Journal of Economic and Social Measurement 29 (2004) 445–472
IOS Press

Attrition bias in labor economics research using matched CPS files

David Neumark[a,*] and Daiji Kawaguchi[b]
[a] *Public Policy Institute of California and NBER, 500 Washington St., Suite 800, San Francisco, CA, 94111, USA*
Tel.: +1 415 291 4476; Fax: +1 415 291 4428; E-mail: neumark@ppic.org
[b] *Graduate School of Humanities and Social Sciences, University of Tsukuba, Tsukuba, Ibaraki 305-8573, Japan*
E-mail: kawaguch@sk.tsukuba.ac.jp

Short panel data sets constructed by matching individuals across monthly files of the Current Population Survey (CPS) have been used to study a wide range of questions in labor economics. But because the CPS does not follow movers, these panels exhibit significant attrition, which may lead to bias in longitudinal estimates. The Survey of Income and Program Participation (SIPP) uses essentially the same sampling frame and design as the CPS, but makes substantial efforts to follow movers. We therefore use the SIPP to construct "data-based" rather than "model-based" corrections for bias from selective attrition. The approach is applied to two questions that have been studied with CPS data – union wage differentials and the male marriage wage premium. The evidence suggests that in many applications the advantages of using matched CPS panels to obtain longitudinal estimates are likely to far outweigh the disadvantages from attrition biases, although we should allow for the possibility that attrition bias leads the longitudinal estimates to be understated.

1. Introduction

Short panel data sets constructed by matching individuals across monthly files of the Current Population Survey (CPS) have been used to study a wide range of questions in labor economics. Relative to more conventional panel data sets such as the National Longitudinal Survey (NLS) or Panel Study of Income Dynamics (PSID), panels constructed from the CPS have numerous advantages, including: larger samples; panels that are not restricted to narrow subsets of years or age cohorts; the availability of geographic information; and quick release of data to researchers. There are, however, two offsetting disadvantages. One, which is not the focus of this paper, is the more limited scope of the CPS, providing – in analyses of wages – pairs of observations 12 months apart, with less extensive information than conventional panels. The second disadvantage, which is the focus of this paper, is that matched CPS panels exhibit rather severe attrition. Unlike other conventional panel data sets, in conducting the CPS no effort is made to follow individuals that move. Consequently,

[*] Corresponding author.

typically only 70 to 80 percent of observations can be matched across CPS surveys over a 12-month interval. This constitutes a very high attrition rate, especially over such a short period, making it entirely plausible that estimated relationships based on matched CPS files suffer from substantial biases due to selective attrition.

We use the Survey of Income and Program Participation (SIPP) to assess the extent of attrition bias in estimates of behavioral relationships using matched CPS files. The SIPP is uniquely suited to this purpose because it uses essentially the same sampling frame and design as the CPS but makes substantial efforts to follow individuals that move. This permits the SIPP to be used to construct two alternative data sets with which to study attrition bias such as that which occurs in the CPS. The first is a standard SIPP panel that follows as many individuals as possible if they move. The second is a SIPP panel that mimics the CPS in dropping individuals that move. For any particular empirical application for which matched CPS files have been used, the estimates from these two constructed SIPP data sets can be compared to estimate the bias from attrition related to moving. This provides an approach to studying this particular form of selection bias in matched CPS samples that relies on observed behavior of individuals who "selected out" of the CPS sample, in contrast to the typical approach that would attempt to model the attrition decision – with well-known identification problems. The approach is applied to a couple of standard economic relationships that have been studied with the CPS.[1]

2. Matched CPS files

2.1. Past uses of matched CPS files

Short panel data sets constructed by matching individuals across monthly files of the CPS have been used to study a wide range of questions in labor economics. Recent examples of policy analyses using matched CPS data include: exchange rate changes and employment stability (Goldberg, et al. [13]); the effects of the minimum wage on employment and enrollment of teenagers and on family poverty (Neumark and Wascher [24]; Neumark, et al. [25]); the impact of the Earned Income Tax Credit (EITC) on transitions out of poverty (Neumark and Wascher [23]); and nursing retention (Schumacher [30]). Recent examples of behavioral analyses using matched CPS data include: analysis of wage growth of immigrants versus non-immigrants (Duleep and Regets [7]); the effects of compensating differentials and sex segregation on the sex wage gap in earnings (Macpherson and Hirsch [19]); the effects of unions on wages (Freeman [10]); unmeasured skills and inter-industry wage differentials (Shippen [31]); and earnings mobility (Gittleman and Joyce [12]).

[1] Matched CPS samples have also been used to study outcomes for families, and the approach we take with the SIPP can also be applied to families. While the paper focuses on the analysis of individuals, we have done some preliminary work applying the approach to families. This is potentially more problematic, as changes in family structure are strongly associated with moving.

2.2. Advantages of matched CPS files

Relative to more conventional panel data sets such as the NLS panels or the PSID, panels constructed from the CPS have five principal advantages.[2] First, whereas conventional panels restrict sample sizes because they face large expenses from the collection of longitudinal data, CPS panels provide much larger samples. For example, the monthly CPS covers approximately 50,000 workers. Many studies using the CPS rely on the Outgoing Rotation Group (ORG) – the subset of the sample in each month for which the "Earners Study" is administered, providing information on wages, etc. Since CPS respondents are in the sample for four months, out for eight, and then back in for four, and the Earners Study is administered in the fourth and the last month, it is possible to match up 1/8 of the sample in any month with Earners Study information one year later.[3] Thus, for any one year, it is in principle possible to match up about 75,000 observations with observations one year later (50,000 × 1/8 × 12 months). In contrast, the National Longitudinal Surveys have drawn samples of about 5,000 individuals (in the Original Cohorts) to 12,000 individuals (in the National Longitudinal Survey of Youth, with a military oversample), the SIPP yields panels about 1/2 to 1/5 as large (depending on the year), and the PSID began as a sample of about 5,000 families but has grown to about 8,700 as the families of offspring of the original sample members are integrated into the survey (Hofferth, et al. [16]).

Second, whereas conventional panels (most prominently, the National Longitudinal Surveys) target and follow a specific cohort as it ages, new CPS panels can be constructed for each year over a long period of time, permitting the analysis of an array of policy approaches that have been used over a period of many years, while still affording the statistical advantages of panel data. More generally, coverage of long time spans permits researchers to more adequately control for changes in the economic (or other) environment that may mediate the relationship that is being studied. Third, because the CPS samples the entire population rather than a specific cohort, it permits richer analyses across the age distribution. Fourth, because of the large sample sizes, and because of the ready accessibility of geographic information (which is often suppressed in conventional panel data sets), state-level or even city-level variation in policy can be exploited in empirical analysis (e.g., Adams and Neumark [1]),[4] and subgroups defined by demography, occupation, etc., can be more reliably analyzed. Fifth, CPS data are made available to researchers very quickly, with many files typically available within one or two months after their collection. This makes these data uniquely well-suited for analysis of recent or current issues.

[2] These advantages do not necessarily all arise with respect to each existing panel, but at least some of them always do.

[3] Welch [32] discusses the matching of respondents across CPS surveys.

[4] Since January 1996, the design of the CPS has resulted in the large- and medium-sized metropolitan areas in the sample being self-representing (Bureau of the Census [3]).

In contrast, NLS panels on particular age groups are available for a limited subset of possible sample periods, depending on when a survey of a particular cohort was begun. PSID and SIPP files are produced with a very long lag; as of January 2001, the most recent SIPP files available were for 1996, and the most recent PSID files were for 1997. For topical policy analysis (for example, studying welfare reform) these constraints can be severely limiting, and they may also hamper behavioral research studies to the extent that these could benefit from recent data.

2.3. Disadvantages of matched CPS files

A principal disadvantage of panels constructed from matched CPS files is that these panels exhibit rather severe attrition. Unlike other conventional panel data sets, in conducting the CPS no effort is made to follow individuals or families that move to a new address. Consequently, depending on the year considered, only 70 to 80 percent of observations can be matched across monthly CPS samples 12 months apart, such as would be required with the ORGs. This reflects a very high attrition rate over a one-year period. By way of comparison, in the NLSY, which began in 1979, over 80 percent of eligible participants were still responding in the late 1990s. For the Mature and Young Women cohorts of the NLS, begun in 1968, over 50 percent of participants were responding after almost 30 years of interviewing (Zagorsky and Rhoton [34]). [5] For a more direct comparison, the PSID lost 12 percent of its respondents between 1968 (its year of inception) and its second round interviews in 1969 (Fitzgerald, et al. [9]).

The relatively high attrition in panels created from matched CPS files makes it plausible that estimates based on these files are biased because of attrition that is nonrandom, even conditional on the observable control variables. This could occur quite naturally because the decisions of individuals to move may be, in part, related to the behavior that is being studied. As an example, consider longitudinal estimation of the union wage premium. Suppose first that individuals who experience a wage decline are more likely to move. (Recent evidence reported by Fitzgerald, et al. [9], studying attrition in the PSID, suggests that those who recently experienced unfavorable economic events were most likely to attrit.) Suppose, as nearly all evidence suggests, that the true effect of being represented by a union is to increase wages. Finally, consider the estimation of a first-difference regression of changes in wages on changes in union status (and other controls), and for simplicity suppose that the sample consists only of those who do not change union status, and those who switch into union jobs.

Given that unions increase wages and wage gains deter moving, there is more attrition of individuals with negative wage shocks among those who do not switch

[5] This lower attrition is not only because movers are followed, but also because continuing efforts are made to find individuals who were non-respondents in one or even more than one survey year (e.g., Rhoton [29]).

into union jobs than among those who switch into union jobs, as the latter on average experience a wage gain that offsets the negative shock and deters attrition. In a regression context, this implies that the error term in the wage change equation is negatively correlated with a dummy variable indicating a switch to union status, and hence the positive impact of unionization in the wage change equation is understated in the sample of non-attriters.

If we also consider those who move out of union jobs (for whom the variable measuring the change in union status would be minus one), then the same qualitative conclusion follows. There will be more attrition among these individuals than among those who do not switch, as their union status change complements the negative wage shock in inducing attrition. Hence the negative impact of leaving a union job is understated in the sample of non-attriters, so again the gains from unionization are understated.

If instead attrition is associated with wage gains, the same bias results. In this case, for workers switching into union jobs there is more attrition of those with positive shocks compared with those who do not switch – as the union change complements the positive wage shock – leading to understatement of the positive union wage impact. And in the case of those switching out of union jobs there is less attrition of those with positive wage shocks than among those who do not switch, again leading to understatement of the costs of leaving a union job. Thus, overall, in either case attrition leads to those with more extreme wage changes selecting out of the sample, which tends to moderate the estimated effects of a change that actually affects wages. [6]

To see this concretely in terms of the usual selection-bias framework, suppose that – as in Freeman [10] and Jakubson [17] – a usual wage equation is specified for the log wage, denoted by lw_{it}, where the subscripts i and t index individuals and years ($t = 1,\ldots,T$). This is specified as a function of control variables Z_{it}, including human capital and other individual-level controls, individual unobserved heterogeneity C_i, as well as the union status variable U_{it}. Then the regression model is:

$$lw_{it} = \alpha + Z_{it}\beta + \gamma \cdot U_{it} + C_i + \varepsilon_{it}. \tag{1}$$

To eliminate the influence of C_i, which may be correlated with Z_{it} or U_{it}, the first-difference model is estimated:

$$\Delta lw_{it} = \Delta Z_{it}\beta + \gamma \cdot \Delta U_{it} + \Delta\varepsilon_{it}. \tag{2}$$

It is assumed that $E(\varepsilon_{it}|Z_i, UNION_i) = 0$, where $Z_i = [Z_{i1}, \ldots, Z_{iT}]$ and $UNION_i = [U_{i1}, \ldots, U_{iT}]$, so that other than attrition bias, OLS estimation of Eq. (2) would yield unbiased estimates.

[6] Similarly, then, if unions decrease wages, the effect is biased towards zero. To consider one case, suppose those with negative wage shocks attrit. Then among those who switch to union jobs, there is relatively more attrition among those with negative wage shocks, moderating the negative effect of unions.

There is also an indicator for attrition (A_{it}), which is assumed to be a function of the same control variables, and the wage shock. If we assume there is no other error term in the attrition equation (in order to keep the notation to a minimum), and suppose that wage declines are associated with attrition, the equation for the propensity to attrit (A^*) can be written as:

$$A^*_{it} = \Delta Z_{it}\beta' + \gamma' \cdot \Delta U_{it} + \tau' \cdot \Delta\varepsilon_{it},\quad ^7 \tag{3}$$

where $\gamma' < 0$ (assuming that $\gamma > 0$), $\tau' < 0$, and the elements of β' have the opposite signs of the elements of β.

Because we observe only the subsample of non-attriters ($A^*_{it} < 0$), the function we estimate with Eq. (2) is:

$$\begin{aligned}&E(\Delta lw_{it}|\Delta Z_{it}, \Delta U_{it}, \Delta\varepsilon_{it} > -\{(1/\tau') \cdot (\Delta Z_{it}\beta' + \gamma' \cdot \Delta U_{it})\})\\ &\ = \Delta Z_{it}\beta + \gamma \cdot \Delta U_{it} + E(\Delta\varepsilon_{it}|\Delta Z_{it}, \Delta U_{it}, \Delta\varepsilon_{it} > -\{(1/\tau')\\ &\ \cdot(\Delta Z_{it}\beta' + \gamma' \cdot \Delta U_{it})\}).\quad ^8\end{aligned} \tag{4}$$

When $\Delta U_{it} = 1$, because γ' and τ' are both negative, $\Delta\varepsilon_{it}$ can be a larger negative number while still satisfying the inequality $\Delta\varepsilon_{it} > -(1/\tau') \cdot (\Delta Z_{it}\beta' + \gamma' \cdot \Delta U_{it})$ in Eq. (4), compared with the case when $\Delta U_{it} = 0$. Thus, the attrition selection bias induces a negative correlation between ΔU_{it} and $\Delta\varepsilon_{it}$, implying that the least squares estimate of γ in Eq. (2) is biased downward. Note that if γ' is zero, this negative correlation is not induced and there is no attrition bias (attrition is random with respect to changes in union status). On the other hand, when γ' is larger in absolute value (so that a change in union status is more strongly associated with attrition), the bias is likely to be more severe.

If instead attrition is associated with wage gains, then τ' and γ' are greater than zero (and the signs of β' are reversed). In this case the inequality defining inclusion in the sample in Eq. (4) is $\Delta\varepsilon_{it} < -(1/\tau') \cdot (\Delta Z_{it}\beta' + \gamma' \cdot \Delta U_{it})$. When $\Delta U_{it} = 1$, because β' and τ' are both positive, $\Delta\varepsilon_{it}$ can be a larger negative number while still satisfying the inequality, compared to when $\Delta U_{it} = 0$. Thus, the attrition selection bias again induces a negative correlation between ΔU_{it} and $\Delta\varepsilon_{it}$, implying that the least squares estimate of γ in Eq. (2) is biased downward.

The fact that the bias goes the same way regardless of the direction of effect of changes in wages on attrition may seem counter-intuitive. Figure 1 provides a diagrammatic illustration of the attrition bias that provides the underlying intuition.

[7] If we appended an independent error term uncorrelated with the control variables and $\Delta\varepsilon_{it}$ to Eq. (3), none of the conclusions would change, although the equations that follow would be more cumbersome. Similarly, attrition can depend on exogenous variables other than Z that satisfy the same conditions as Z.

[8] Note that the condition for non-attrition (and hence inclusion in the sample) is $\tau' \cdot \Delta\varepsilon_{it} < -(\Delta Z_{it}\beta' + \gamma' \cdot \Delta U_{it})$, but the inequality gets reversed because we divide through by τ', which is negative.

To simplify, suppose $\beta' = 0$. In the top panel, observations are likely to attrit when Δlw is low. As shown by the plot of the density of $\Delta\varepsilon$, and the "cut points" below which attrition occurs, when $\Delta U = 1$ observations attrit only when shocks $(\Delta\varepsilon)$ take on large negative values; thus observations with relatively large negative shocks remain in the sample. In contrast, when $\Delta U = -1$, even moderate negative shocks cause the observations to attrit, so that only observations with positive shocks of small negative shocks remain in the sample. As this discussion shows, ΔU and $\Delta\varepsilon$ are *negatively* correlated in the remaining sample. A similar discussion applies when observations are more likely to attrit when Δlw is high, as illustrated in the lower panel of Fig. 1. Again, we can see that ΔU and $\Delta\varepsilon$ are negatively correlated in the sample of non-attriters. Note that this analysis concerns selection into attrition on unobservables rather than observables. Fitzgerald, et al. [9] demonstrate that correction for biases from selection on observables, even when this selection is endogenous, can be accomplished by weighted least squares.[9]

3. Estimating the bias from selective attrition

3.1. The conventional approach to attrition bias

The usual method to correct regression estimates for selection bias from attrition would be to implement a sample selection correction, following the seminal work on this topic in the context of wages and labor supply by Heckman [14] and [15]. However, if identification of such a model is not to be dependent on functional form and the assumed distributions of the errors, at least one variable that drives attrition but not the behavior of interest (in this particular context, changes in wages) is required (Olsen [26]). In studying behavior at the individual level, such assumptions are often quite problematic. This is exacerbated in the present context because the CPS is not a particularly rich data set, and therefore lacks information on unusual variables that might provide identification. Thus, the paper takes a different approach that provides evidence on attrition bias without requiring the specification of a joint model of attrition and the behavior of interest.

3.2. Using the SIPP to estimate and account for attrition bias

The SIPP is used to assess the extent of attrition bias in estimates of behavioral relationships using matched CPS files. The SIPP is uniquely suited to this purpose because it has many features similar to the CPS files. Indeed, the March CPS and the SIPP are very close substitutes (with each having certain advantages) as the source

[9] In addition, a good deal of research (reviewed in Mack and Petroni [20]) has explored the construction of weights to account for attrition and nonresponse generally in the SIPP.

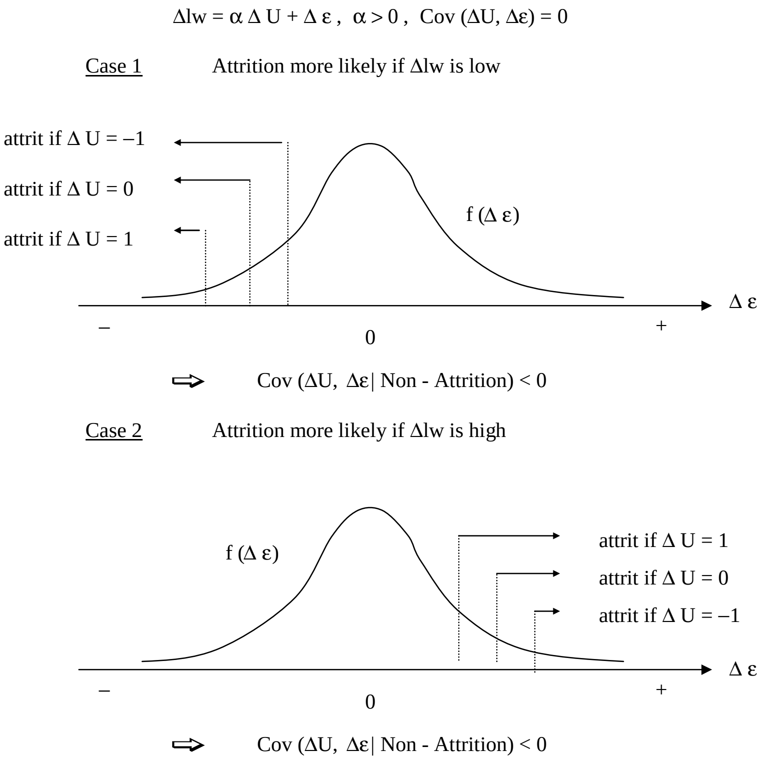
Δlw = α Δ U + Δ ε , α > 0 , Cov (ΔU, Δε) = 0
Case 1
Attrition more likely if Δlw is low
attrit if Δ U = –1
attrit if Δ U = 0
attrit if Δ U = 1
f (Δ ε)
Δ ε
–
0
+
⇨ Cov (ΔU, Δε| Non - Attrition) < 0
Case 2
Attrition more likely if Δlw is high
f (Δ ε)
attrit if Δ U = 1
attrit if Δ U = 0
attrit if Δ U = –1
Δ ε
–
0
+
⇨ Cov (ΔU, Δε| Non - Attrition) < 0

Fig. 1. Diagrammatic Illustration of Attrition Bias.

of annual income and poverty estimates in the United States (Citro and Kalton [4], pp. 85–86).[10] However, whereas the matched CPS files provide a "quasi-panel," in the sense that the surveyors simply return to the same address to interview the people who currently reside there, the SIPP follows individuals who change their address. More specifically, SIPP interviewers attempt to interview in person everyone who remains within 100 miles of a SIPP primary sampling unit, and interviewers are instructed to conduct telephone interviews with movers who do not satisfy this criterion (Citro and Kalton [4], p. 92). Because of this difference in sampling strategies, many individuals that would be lost in matched CPS files can be followed in the SIPP.

This permits the SIPP to be used to construct two alternative data sets with which to study attrition bias. The first is the standard SIPP panel that follows as many individuals as possible as they move. The second is a SIPP panel that mimics the CPS in dropping individuals that move. For any particular empirical application for which matched CPS files have been used, these two constructed SIPP data sets can

[10] Coder and Scoon-Rogers [6] discuss many of the differences and similarities.

be used to estimate the same relationship studied with the CPS. The estimates with and without the movers included can be compared to estimate the attrition bias in the estimated relationship, and the results for the sample with the movers included can provide estimates in which attrition bias is mitigated. Thus, using the SIPP data in this manner provides an approach to studying attrition bias in matched CPS samples, based on a method that relies on observed behavior of individuals that "selected out" of the CPS sample, in contrast to the model-based approach to attrition.

To see more explicitly how the SIPP data can be used, denote by M the set of individuals who move over a one-year period in the SIPP, beginning from some specified wave and month of a SIPP panel, and denote by N the non-movers. Assume for now, for purposes of discussion, that all movers are captured and included in M. Then continuing with the earlier example, estimating Eq. (2) for the set N mimics what is done with the CPS data. If the SIPP data are chosen judiciously, the results should match results from the CPS quite well, although there may be some differences of measurement.[11] Such estimates provide the baseline for the analysis, although they will be compared with corresponding CPS estimates for the same sample period to establish the comparability of the SIPP data.

The simplest way to gauge the extent of attrition bias is then to estimate Eq. (2) using the larger set $M + N$ – i.e., adding in the movers – as these estimates should be free of attrition bias. The differences in the estimates of γ and β measure the bias from attrition (differences in intercepts are generally of little interest).

A stringent test for attrition bias can be constructed. Now letting A_{it} denote those who move – and therefore "attrit" from the CPS-type sample – Eq. (2) can be augmented to include A_{it} and interactions of this indicator with all of the included variables:

$$\Delta lw_{it} = \alpha'' A_{it} + \Delta Z_{it}\beta + \Delta Z_{it} \cdot A_{it}\beta'' + \gamma \cdot \Delta U_{it} + \gamma'' \cdot \Delta U_{it} \cdot A_{it} + \nu_{it}. \quad (5)$$

In this specification, if the estimates of β'' and γ'' are not significantly different from zero, then the hypothesis that the parameters β and γ are equal in the two subsamples cannot be rejected. Any subset of these parameters can also be tested (most notably the union effect in this application).[12] One potential problem is that such tests are potentially over-restrictive, in that there could in principle be a significant difference between the two subsamples, but because of the smaller size of the set of movers, the estimates based on the full sample $(M + N)$ and the sample of non-movers only (N) might still be statistically indistinguishable. As a matter of practical consequence, then, this would indicate that estimates from matched CPS samples do not exhibit significant attrition bias. Of course, this problem does not

[11] For example, income is measured on a monthly rather than an annual basis (in the March CPS), which introduces some differences in the classification of families as poor. (See Coder and Scoon-Rogers [6].)

[12] This closely parallels what Falaris and Peters [8] call the "comparison method," for obvious reasons. See also Becketti, et al. [2].

arise if the data fail to reject the restrictions that β'' and γ'' equal zero. But in cases where the restrictions are rejected, care must be taken to compare the estimates using $M + N$ and N to assess whether the attrition bias is a concern.

Constructing a test is non-trivial because the samples are non-independent. One approach is to estimate the regression for $M + N$, and to then estimate the regression for N, first unconstrained and then restricting the coefficients (or some pertinent subset of them) to equal the estimates for $M + N$. This is intuitively appealing but will understate the p-values by failing to account for sampling variation in the estimates from $M + N$. Another approach is to apply the Hausman test framework. For example, denote by g_{M+N} the estimate of γ from the pooled sample and by g_N the estimate from the sample of non-movers. Then, under the null of no attrition bias, g_N and g_{M+N} are both consistent estimates of γ, but the latter is efficient, while under the alternative of attrition bias, g_{M+N} is consistent but g_N is inconsistent. Thus, as desired, the test of the null that $(g_{M+N} - g_N)$ equals zero is a test of the null of no attrition bias. This is not quite the usual Hausman set-up, in which the estimate that is efficient under the null is the one that is inconsistent under the alternative. But the essential idea underlying such tests still applies – i.e., that the covariance between an efficient estimator and its difference with respect to an inefficient estimator, both of which are consistent under the null, is zero. However, this requires that the estimate using $M + N$ is efficient, so specification tests for heteroscedasticity must be implemented and weighted least squares (WLS) used if necessary.

Of course, one could argue that if it is possible to recover from the SIPP data estimates that do not exhibit attrition bias, research should just use the SIPP instead of matched CPS files. In some cases, this is a reasonable position. But recall all of the advantages of the CPS relative to the SIPP that were discussed earlier.

The discussion to this point has been in the context of one particular application, namely the use of matched CPS files to study the effects of union status on wages. But the goal of this research is to try to provide some general information, rather than simply improving on existing estimates in one particular study. Therefore, these same methods will be applied to another context, in particular the effects of marriage on wages of men (see, e.g., Korenman and Neumark [18]). Because address changes associated with marriage are likely and the effects of marriage may be different for those who change address at the time of marriage, longitudinal estimates of the marriage premium obtained only from a set of individuals without address changes may be particularly prone to attrition bias.

These two applications were chosen because the changes on which they focus – changes in jobs (which we assume often accompany changes in union status), and marital status transitions – seem relatively likely to be associated with moving. Hence, they may provide estimates toward the upper range of the effects of attrition bias in matched CPS samples, although that is only speculation. As a consequence, if these applications as a whole yield little evidence of attrition bias in matched CPS files, the combined evidence should prove relatively reassuring with respect to the use of these files. On the other hand, evidence of attrition bias would suggest caution in drawing conclusions from matched CPS files, without attempting to verify, in the particular context being studied, that attrition bias was unlikely to be a concern. If instead applications were considered in which attrition bias was unlikely – on a priori grounds – to be a substantive concern, the results would be relatively uninformative. Aside from this, the paper focuses on applications for which matched CPS files have been used to address topics that have been of broad interest to labor economists.

3.3. Prior work on attrition in matched CPS files and other panels

Previous research has looked at some evidence on the effects of non-matches in CPS panels, although the emphasis has been different from that in the present study, and the analysis more limited. Specifically, Peracchi and Welch [28] consider the question of representativeness of the cross-section and panel data sets resulting from matching CPS files, using the March files from 1979 to 1991. For the analysis of cross-sectional files, they compare matched and unmatched families or individuals. They do find some differences in terms of both demographic characteristics and labor market outcomes. Their ability to study bias in the panel files is more limited, of course, since they do not have longitudinal data on unmatched workers. They study this question in two ways. First, they use bounds analysis treating this as a selection problem; and, second, they use grouped data to obtain what the authors characterize as "indirect" information about whether match failure creates any bias in these panels. Their findings suggest that for the only transitions they study – those among labor force states – biases are small. While this is a valuable contribution, in the absence of data on unmatched individuals or families the evidence is suggestive but not decisive. The present study, by recovering data on unmatched individuals or families, should be more informative and provide more direct evidence. Furthermore, the evidence will focus specifically on bias in the estimates of behavioral effects, which differs from issues regarding representativeness of the samples.

Other recent research has studied attrition in conventional panels. MaCurdy et al. [21] fully explore the role of attrition in the NLSY, focusing on its effects on representativeness of the data. They also look at differences between those who attrit "permanently," and those who attrit temporarily (missing some waves of the survey but returning later). Similarly, Fitzgerald et al. [9] study the changing representativeness of the PSID. The approach in the present study differs from the approach taken in these papers by asking about the impact of attrition bias on estimates of behavioral relationships.

Research by Falaris and Peters [8] is closer to what we do. They use data from three NLS cohorts and the PSID to examine how survey attrition affects estimates of models of schooling choices. Falaris and Peters estimate regressions using data on (1) people who always respond to the surveys ("stayers"); (2) people who miss some surveys but re-appear in later waves ("intermittent" attriters); and (3) permanent attriters. They further subdivide their subsamples of attriters into people who attrit after the observed behavior is measured (in this case, after age 25), and people who attrit earlier, precluding measurement of this behavior unless they are intermittent attriters. In their particular context, late attrition is much more common, and more of it is permanent. Paralleling the estimation and testing of Eq. (5), they test whether the same statistical model describes the behavior of stayers and different types of attriters. Their findings focus on family background effects on schooling choices (highest grade completed and age of school completion), for which they generally find little effect of attrition on the model estimates. However, while similar in spirit, the work in this paper differs in two important ways. First, the behaviors Falaris and Peters study, and the key explanatory variables, do not seem like obvious candidates for variables that are likely to be intimately related to moving or other sources of attrition, although this is of course speculative. Second, in this study we are most interested in the behavior that would otherwise emerge *after* the attrition. Falaris and Peters can draw such inferences for the intermittent attriters, but it seems likely that this group is most like the non-attriters, so it may not be surprising that they find relatively few differences. Regardless of the substantive importance of these differences, the present study differs fundamentally in focusing on matched CPS files – a potentially valuable alternative source of panel data about which serious questions regarding attrition arise.

4. Data set construction

The central question that determines whether our approach is informative is how well the SIPP does at tracking movers. It would be ideal to track *all* movers, since in that case it would be possible, in principle, to divide the SIPP sample into the two "ideal" subsamples – one representing individuals or families that the CPS tracks, and one representing those individuals and families that the CPS fails to track. This ideal goal cannot be achieved, but it is obviously important to assess how close to it one can get with the SIPP, since the research will only be informative if a reasonably high fraction of movers is tracked.

A second requirement for the SIPP analysis to be informative regarding non-matching in the CPS is that this non-matching must be strongly related to moving, since this is the main dimension on which the SIPP does better than the CPS. An early study of this question using 1979–1983 CPS data (for households) indicated that 42 percent of non-matched households are movers; the next largest share is non-interviews (Pitts [27]). The research by Peracchi and Welch also suggests that moving is an important component of non-matching, concluding that "the main source of attrition in the CPS is failure to follow people of college age in matched households and young households who move" (p. 160). The remainder of this section describes the construction of the individual-level data sets – in particular the methods of following individuals in the SIPP – and assesses the extent to which these conditions hold.

Table 1
Codes for SIPP Item 36B

01	Interviewed	16	Entire household institutionalized
02	No one home	17	Demolished
03	Temporarily absent	18	House or trailer moved
04	Refused	19	Converted to permanent business
05	Unable to locate	20	Merged
06	Other Type A	21	Condemned
09	Vacant	22	Deleted (sample adjustment, error)
10	Occupied by persons with usual residence elsewhere	23	Entire household deceased, moved
		24	Moved, address unknown
11	Unfit or to be demolished	25	Moved within a country beyond limit (cannot be reached by telephone)
12	Under construction, not ready		
13	Converted to temporary business	26	All sample persons re-listed on new control card
14	Unoccupied site for mobile home		
15	Permit granted, construction not started		

The original SIPP data sets were obtained from the Data Extraction System, which is maintained by the US Census Bureau on the Internet. The core files of the 1990, 1991, 1992, and 1993 SIPP panels are used in this analysis. In each SIPP panel, respondents are surveyed three times in a year (with each interview referred to as a wave), and the interview collects information from respondents covering the previous four months, with most data available on a monthly basis. The most common CPS files used in constructing panels from matched monthly files are the ORG files, corresponding to the 4th and 16th months in the CPS sample, because these months include information on labor market earnings; certainly these files would be used in a CPS-based study of union- or marriage-related wage differentials. Thus, to use SIPP data corresponding to the CPS data as closely as possible, attention was similarly restricted to data from the 4th and 16th months of each of these four SIPP panels, using the first and fourth interviews.

The SIPP data are used to try to identify individuals who would have attrited in the CPS because of a change of address. Of course, there is also attrition in the SIPP, and it is of interest to know the reasons for this attrition, because if a good deal of attrition in the SIPP is attributable to changes of address, the SIPP data are less useful for the research. Thus, the first task is to determine the reason for attrition in the SIPP panels. To discuss this, it is necessary first to define some of the terms regarding data collection in the SIPP. Two data collection instruments are used in the SIPP: a control card and a questionnaire. An interviewer uses a control card that is issued for each address unit throughout the longitudinal survey. A questionnaire is used to collect the data, with a new questionnaire used for each interview.

The first step is to determine if a specific observation is interviewed in the specific months used (months 4 and 16). Item 36B in the control card, which is coded for each interview month, is used to determine interview status and, as noted above, corresponds to the address unit. This same item also provides the reason for attrition. The codes for this item are given in Table 1.

In addition to these codes, the code 00 is assigned for an observation that is out of the sample at the time of interview. Empirically, over the course of the panel, the code 00 is observed following codes 02–06, 16, 22–26, and 01. It is not possible to assign the reason for attrition for observations for which the code 01 is followed by 00, as these are cases in which observations are dropped after an interview without a specific reason being provided. However, in the other cases the reason for attrition can be identified. Specifically, the reason for attrition between month 4 and month 16 can be ascertained by tabulating item 36B for each month conditional on the code 00 for the next month, using the following algorithm:

1. Tabulate item 36B at month 16.
2. Tabulate item 36B at month 15 if item 36B at month 16 $= 00$.
3. Tabulate item 36B at month 14 if item 36B at month 16 $= 00$ and item 36B at month 15 $= 00$.

..., until month 4 is reached.

This algorithm is used to identify the reason for attrition for the last month prior to the code switching to 00. These reasons for attrition are grouped into the categories relevant for the research, as follows: category A–household did not move but cannot be contacted, corresponding to item 36B codes 02–06; category B–deceased, out of country, corresponding to item 36B codes 16 and 23; and category C–household moved and cannot be contacted, corresponding to item 36B codes 22, 24, 25, and 26.

With these categories, the data can be broken down as follows. First, those who move and are tracked by the SIPP can be classified straightforwardly as movers tracked by the data, using the "address ID" that simply keeps track of the number of address changes of individuals followed in the survey. On the other hand, observations in category C are movers who are not successfully tracked by the SIPP, while categories A and B include observations that attrit for reasons unrelated to moving. The full classification of observations is reported in Table 2.

For each year, the table begins by reporting the number of observations in the sample by the 4th month. In 1990, for example, there are 58,249 such persons, of whom 58,149 are interviewed in the 4th month, which is our starting sample. Of these individuals, 50,504 are interviewed in the 16th month, while 7,109 attrit, for an attrition rate of 12.2 percent. In addition, for an additional 536 observations there is an individual identified in the SIPP as an interviewee in month 16, but the demographic information on this individual does not match that in month 4, leading us to discard these observations.

The fourth and fifth rows of the table present the critical information on how many of the observations *retained* from month 4 to month 16 move. As the table shows, again focusing on 1990, 43,291 are non-movers and 7,213 are movers. The first sample is the one that "mimics" the CPS, in that it loses all attriters, whether because of moves or other reasons. The overall attrition rate from this sample – defining attriters as those who really attrit in the SIPP, plus the movers who are tracked – is 24.6 percent ($\{7,213 + 7,109\}/58,149$), very similar to CPS attrition rates over the course of a 12-month period. The observations that are not followed in the SIPP are then broken into the three categories described above, revealing that 3,800 attrit

Table 2
Classification of Individual Data by Moving and Attrition Status

Year	1990	1991	1992	1993	All years
Entered sample by 4th month	58249	37478	51380	52092	199199
Interviewed at 4th month	58149	37424	51235	51995	198803
Interviewed at 16th month, given interviewed at 4th month (*SIPP match*):	50504	32281	45042	45511	173338
Non-movers[1] ("CPS sample")	43291	27863	38682	39172	149008
Movers	7213	4418	6360	6339	24330
Month 4 to month 16 attriters:	7109	4836	5753	6068	23766
Attrition not related to moving	3800	2613	2918	3281	12612
No interview, Type A (HH not moved but can't be contacted):					
No one home	149	122	102	168	541
Temporarily absent	164	91	148	206	609
Refusal	2839	2028	2194	2427	9488
Unable to locate	9	3	5	0	17
Other	131	87	95	81	394
No interview, Type B (deceased, out of country), entire household out-of-scope	508	282	374	399	1563
Attrition related to moving	2474	1644	2123	2176	8417
No interview, Type C (HH moved and can't be contacted):					
Moved, address unknown	1582	1068	1473	1574	5697
Moved within country beyond limit (phone interview failed)	68	68	47	54	237
Other	824	508	603	548	2483
The reason cannot be assigned[2]	835	579	712	611	2737
Inconsistent match based on demographic information:	536	307	440	416	1699

[1] When an observation's address code is not changed but the state code is changed, the observation is treated as a non-mover (45 cases in pooled sample).
[2] Coded as not matched or not in sample by 16th month; reason cannot be determined.

for reasons unrelated to moving (categories A and B), while 2,474 attrit for reasons related to moving (category C); 835 cannot be classified. The table also gives an indication of the reasons for attrition related to moving. In just under two-thirds of the cases, the respondent has moved and the address is unknown. In a handful of cases the address is known, but the address is within the country yet outside the limit to which a SIPP interviewer will travel and the phone interview failed.[13] The "other" category includes the remaining cases (including, for example, moves outside the country and outside the limit of a SIPP interviewer, for which a phone interview failed).

These numbers indicate that of the 9,687 movers (7,213 of whom are followed, plus the 2,474 category C individuals), 74.5 percent are successfully followed in the SIPP. Comparable numbers for 1991, 1992, and 1993 are, respectively, 72.9 percent,

[13] According to the 1990 SIPP User's Guide, only 4 percent of the population is outside this limit.

75.0 percent, and 74.4 percent – figures that are quite robust. These percentages are central to this research. It is in principle conceivable that attrition bias is more severe with respect to the movers that are not tracked by the SIPP; while the research will not implement standard selection type corrections to explore this question, it will compare observables for the tracked and non-tracked movers. Nonetheless, it seems quite clear that recovering nearly three-quarters of the movers should give a relatively firm idea of the extent of attrition bias from moving.

At the same time, in the estimation of Eqs (2) and (5) using the SIPP, the observations for movers that are tracked by the SIPP are weighted up to represent the subsample of attriters. For example, corresponding to the figures just cited, each observation is multiplied by a weight of approximately 1.33 (1/.75). Otherwise the resulting estimates (of Eq. (2), especially) would not represent the appropriately-weighted average of coefficients for non-movers and movers. This assumes, of course, that whether a mover is tracked in the SIPP is random.

The evidence in Table 2 demonstrates the feasibility of using the SIPP data at the individual level to construct (1) data sets that mimic CPS panels constructed from matched monthly CPS files, and (2) data sets that recover a substantial percentage of movers who are lost in CPS-type matches. This offers considerable potential with regard to testing for attrition bias in behavioral relationships estimated using matched CPS panels.

5. CPS-SIPP matching comparisons

In this section we assess whether the SIPP data are informative about biases from attrition in economic relationships estimated with the CPS data. First, suppose (as assumed earlier) that the SIPP captures all "would-be" attriters in the CPS. Then the artificial match using the SIPP that mimics the CPS match – i.e., the match that throws out all movers – should have similar features to the actual CPS match. In particular, the characteristics of those who are matched should be similar across the two data sets, as should the characteristics of those who are not matched.

Information useful in assessing how well these conditions hold in the individual data is reported in Table 3. Columns (1)–(3) report results for the matched CPS data, for the same years covered in the previous table.[14] The first column reports descriptive statistics for the full set of CPS observations (as of the first observation in the potential match), and the following two columns break these out by those matched 12 months

[14] Details on the matching in the CPS, and the procedures chosen to maximize comparability between the CPS and SIPP data, are given in the notes to the table. Madrian and Lefgren [22] compare the performance of several methods of matching in the CPS and recommend the use of household number, household ID, line number, race, sex, and age, based on tradeoffs between false matches and keeping the match rate sufficiently high. The matching method employed for the CPS data in this paper is almost identical to the method they recommend, except for the use of household number.

Table 3
CPS and SIPP Match Comparisons, Individual Data, All Years Pooled

Data source	CPS			SIPP			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Interviewed and information for matching available	293332			198803			
Observations	Total	Non-match	Match	Total	Non-match + movers (CPS non-match)	Non-movers (CPS match)	Matched movers
N	293332	76087	217245	198803	49795	149008	24330
Descriptive statistics							
Black	0.106	0.132	0.097	0.123	0.146	0.116	0.122
Native/Asian	0.042	0.052	0.038	0.039	0.047	0.037	0.047
Female	0.520	0.506	0.524	0.519	0.510	0.523	0.523
Married	0.557	0.415	0.605	0.559	0.452	0.594	0.448
	[226227]	[56947]	[169280]	[151759]	[37277]	[114482]	[17572]
Union	0.162	0.124	0.175	0.157	0.122	0.171	0.110
	[118108]	[31174]	[86929]	[85510]	[22586]	[62924]	[11944]
HS drop out	0.223	0.231	0.220	0.254	0.263	0.258	0.231
HS graduate	0.362	0.359	0.363	0.359	0.357	0.360	0.340
Some college	0.221	0.228	0.218	0.192	0.201	0.189	0.218
College graduate	0.195	0.182	0.199	0.190	0.179	0.193	0.211
	[225663]	[56773]	[168890]	[151974]	[37326]	[149008]	[17601]
Age 0–14	0.229	0.251	0.221	0.236	0.251	0.232	0.278
Age 15–23	0.121	0.180	0.100	0.123	0.171	0.107	0.185
Age 24–30	0.109	0.162	0.091	0.112	0.170	0.092	0.201
Age 31–37	0.118	0.122	0.116	0.119	0.124	0.117	0.133
Age 38–50	0.176	0.136	0.190	0.173	0.131	0.187	0.119
Age 51–63	0.113	0.068	0.129	0.110	0.070	0.123	0.049
Age 64–75	0.089	0.047	0.104	0.085	0.049	0.097	0.024
Age 76–90	0.045	0.034	0.049	0.042	0.032	0.045	0.011
Hourly rate	11.023	9.846	11.446	10.992	9.752	11.435	9.736
of pay (calculated)	[118103]	[31174]	[86929]	[83928]	[22078]	[61850]	[11688]

CPS data are from 1990–1991, 1991–1992, 1992–1993, and 1993–1994 matches, based on Outgoing Rotation Group files. Data from the 4th and 16th months in the SIPP surveys from the 1990, 1991, 1992, and 1993 panels are used, to correspond to the CPS data. There are four rotation groups (sub-panels of a particular panel) in the SIPP, and the 4th months for these fall in February, March, April, and January for rotation groups 1, 2, 3, and 4, respectively. Thus, CPS data from January through April were used to be consistent with interview months of the SIPP sample. CPS matching was based on household identification number, line number (individual identifier within household transcribed by interviewer), race, sex, and age. If an observation had a missing value for any of these variables, it was dropped. Individuals are allowed to age 0 to 2 years between interviews. These procedures follow Peracchi and Welch [28]. Non-unique matches (10 in 1993 and 20 in 1994) were, however, classified as non-matches, with no further effort made to match these observations (in contrast to Peracchi and Welch [28]). Characteristics are those reported in the 4th month of survey. Variable definitions are as follows: married equals 1 if an individual is married and spouse is present, 0 otherwise; and union equals 1 if an individual is a union member, 0 otherwise. In some cases the number of observations used to calculate statistics is in square brackets under the statistics. The numbers of observations are reduced for marital status and education because the marriage and education questions are only asked for those over age 15, and for the education question there is additional missing information. The sample universe for the union membership question is those employed in private or government sectors. In neither data set are matches ever based on data known to be allocated.

later, and those unmatched. Overall, the match rate is 74.1 percent. Non-matched observations are more likely to be non-white (with the proportion black higher by 0.035), and slightly less likely to be female (by 0.018). Non-matched observations are considerably less likely to be married (by 0.19) and unionized (by 0.051). In terms of education, the differences are less pronounced, with the proportions generally within 0.01 except for college graduates. Non-matched observations tend to be younger, especially within the 15–30 age range. (These features of the data were very similar in each year of CPS data used.) Finally, non-matched observations have hourly wages that are lower by about 16 percent, consistent with lower rates of marriage and unionization, a higher proportion black, and younger ages.

The next three columns ((4)–(6)) report comparable figures for the SIPP, first reporting overall descriptive statistics, then those for the match that mimics the CPS, with descriptive statistics for those who could not be followed plus the movers (the "CPS non-match"), and then those who could be followed and did not move (the "CPS match"). The estimates indicate that the CPS and SIPP display very similar characteristics. The descriptive statistics in column (1) and column (4) match up quite closely. Furthermore, the CPS-type match based on the SIPP data corresponds closely to the actual CPS match. The estimates in column (2) are similar to those in column (5), while those in column (3) are similar to those in column (6). Finally, the differences between those matched and not matched are similar, as revealed by a comparison of the differences between columns (5) and (6) with the differences between columns (2) and (3). In particular, columns (5) and (6) reveal that non-whites are less likely and women are more likely to be matched, by similar amounts to the differences between columns (2) and (3) (e.g., 0.030 vs. 0.035 for blacks, and 0.013 vs. 0.018 for women). Similarly, as with the actual CPS match, married individuals are considerably more likely to be matched (0.142, vs. 0.190 in the actual CPS match), as are union members (0.049 vs. 0.051). There are small differences by education, and again, as for the actual CPS match, younger individuals are less likely to be matched, with the magnitudes of the differentials similar across the artificial and actual CPS matches. As the last row shows, there is also a similar wage gap (17 percent in this case).

While this appears to establish that the CPS-type match using the SIPP mimics the actual CPS match well, the methods used in this study to estimate the effects of attrition bias rely on the movers who can be followed in the SIPP being representative of the non-matches in the CPS. Column (7) reports figures useful in assessing whether this holds, reporting descriptive statistics for the movers who are followed in the SIPP. The ideal here would be for the numbers in columns (7) and (2) to be similar. Even if this does not hold, it would be desirable for the numbers in column (7) to be more like those in column (2) than in column (3), indicating that the followed movers are more like the CPS attriters than the CPS non-attriters. These conditions hold for the most part. The proportions black, native/Asian, married, and union members in column (7) are relatively close to those in column (2), and certainly closer than they are to those in column (3). The age distributions are also quite similar for the matched movers in the SIPP and the non-matches in the CPS. With respect to wages, this condition holds quite closely, as the average wage for matched movers is quite close to that for non-matches in the CPS. The only variables for which the numbers in column (7) are closer to those in column (3) are the proportion female and education, although for these variables the differences between the matched and unmatched observations were relatively small in the first place. Of course ultimately these conditions need to hold for the unobservables conditional on the observables, rather than necessarily holding for the observables themselves. Since that is impossible to test, all one can do is look at the observables. And even here, observables that change over time cannot be measured for the attriters. In what sense, then, do comparisons of distributions of some of the observables tell us something about the distributions of the unobservables? Suppose we are estimating a generic regression using the CPS data

$$y_i = x_i\beta + \eta_i, \tag{6}$$

for the sample of non-attriters $(A_i = 0)$. (This can be interpreted as our wage change regression.) Because of attrition related to wage changes, it is possible that $E(\eta_i|x_i, A_i = 0) \neq 0$, which can bias the estimates of β. The only way the sample of non-attriters gives us unbiased estimates is if $E(\eta_i|x_i, A_i = 0) = E(\eta_i|x_i) = 0$, which we do not want to assume to be true.

However, our approach is informative if the CPS data and the CPS-match using the SIPP data are representative of the same sample, and hence are expected to yield similar estimates, and if the matched movers in the SIPP are representative of CPS attriters. For the first condition to hold, we require that $E(\eta_i|x_i, A_i = 0)$ is the same in the two data sets (and that β is also the same). For the matched movers in the SIPP to be representative of CPS attriters, we require that $E(\eta_i|x_i, A_i = 1)$ is the same for CPS attriters and the matched movers.

All we can compare across the two data sets, though, are the conditional distributions of $x_i^f|A_i = 1$ and $x_i^f|A_i = 0$, where the 'f' superscript indicates the subset of variables in x that are fixed over time and hence fully observable for both attriters and non-attriters. If we are willing to assume that the joint distributions of the unobservables and the observables are identical, then if the distributions of the observables are identical, the distributions of the unobservables conditional on the observables are also identical. Note that this discussion focuses on necessary but not sufficient conditions. In particular, we have not examined whether the distributions referred to above are identical, but only the means. Also, we have not examined the distributions of time-varying variables, which are observable only for non-attriters. But we believe this brief discussion outlines what is implicitly in researchers' minds when they look at distributions of the available observables to try to learn something about representativeness of data sets selected on some characteristic or behavior.

Table 4
Union Wage Premium Analysis, Sample Details and Descriptive Statistics

	CPS sample		SIPP sample	
Sample construction				
Total matched observations	217245		173338	
Age>=16	166239		129528	
Employed at both month 4 and 16	90937		73309	
Employed by private firms or government	76579		63139	
Wage is available for both month 4 and 16	76165		61882	
Union membership status is available for both month 4 and 16	76165		61882	
All explanatory variables for wage regression are available	76005		61882	
CPS match	–		52694	
Matched movers	–		9188	
Descriptive statistics				
Union	0.208		0.193	
Hourly rate of pay (calculated)	11.68		11.55	
Hours worked per week	38.71		39.17	
Weekly earnings	475		–	
Monthly earnings	–		1993	
Weeks worked in a month	–		4.285	
Union transitions (%)				
Not covered – not covered	74.53		76.88	
Not covered – covered	4.68		3.86	
Covered – not covered	4.22		4.25	
Covered – covered	16.57		15.01	
		SIPP non-movers		SIPP movers
No change in union status		91.89		91.88
Change in union status		8.11		8.12

Weeks worked in a month in the SIPP is reported in integer values. Thus, reported values are either rounded up or down if workers worked a fraction of a week. Over-frequent rounding downward by interviewees may have made the calculated hourly rate of pay higher for the SIPP sample.

6. Assessing the attrition bias

6.1. Union wage premium

Our first analysis focuses on the estimation of union wage premia using matched CPS files. As a preliminary, Table 4 reports some information relevant to this analysis. The first panel describes the sample construction, beginning with the matched samples (column (3) in Table 3 for the CPS data, and columns (6) and (7) for the SIPP data). Focusing on those aged 16 and over reduces the sample by about one-quarter, and requiring employment in months 4 and 16, by private firms or government, reduces it by a bit more than half. The second panel reports basic descriptive statistics. These are quite close in the two data sets, although some differences are expected in part because the CPS data refer to a week while the SIPP data refer to a month, and in part because of other slight differences in the surveys. The third panel first compares transitions across union status in the two data sets, showing that changes are slightly more common in the CPS. Finally, the last two rows ask whether, in terms of these transitions, movers and non-movers in the SIPP look very different. The probabilities of changes in union status are virtually identical among non-movers and movers, despite our expectations that union status changes would be associated with moving.

Table 5 reports the regression results.[15] We begin in the first three columns by reporting cross-section (OLS) and first-difference estimates using the matched CPS files. The cross-section estimate of the union wage premium is 0.160. The first-difference estimate is a good deal lower, at 0.044. By way of comparison, Freeman [10] reports a cross-sectional estimate of 0.19 and a longitudinal estimate of 0.09, using matched May CPS data for 1974 and 1975 (see his Table 6), and Jakubson [17] reports a cross-sectional estimate of 0.179 and a longitudinal estimate of 0.080, using PSID data from 1976 to 1980. The specifications are not identical and the years are different, but qualitatively the results are similar, although our longitudinal estimate is below the estimates in these studies.

Next, to establish whether the SIPP data using the CPS match provide a good baseline, the same specifications are reported in columns (4)–(6) using the SIPP data but excluding matched movers. The estimates are uniformly higher by about 0.02 to 0.03, but otherwise the pattern in going from OLS to first-difference estimates is the same.

The crux of the analysis comes in the remaining eight columns. Columns (7)–(10) report the same set of estimates, but now using all of the matched SIPP data (i.e., adding in the matched movers). The OLS estimates are very close to those using the SIPP match that mimics the CPS, with the estimate of the union premium lower by a trivial 0.004. The first-difference estimates differ by only 0.001 or 0.002. Similarly, the Hausman tests reported in columns (8)–(10) do not indicate that there is attrition bias in the estimates using the CPS match with the SIPP data, in turn suggesting that there is no serious bias from attrition in the longitudinal estimate of the union premium using matched CPS files.

Nonetheless, the slightly higher fixed effects estimate with the matched movers included suggests that those individuals who move may experience greater gains to becoming union members than the average individual. To explore this more fully, columns (11)–(14) report estimates of the specification augmented to include a dummy variable for attriters (matched movers), and more importantly an interaction between this dummy variable and the union variable, paralleling Eq. (5) in the text. In this specification the union-attrition interaction identifies the differential effect of changes in union status for those who move, as the attrition variable is time-invariant. The first-difference estimates reveal that the estimated impact of unionization is only slightly larger for attriters than for non-attriters, by 0.01, and the differential is not

[15] All estimation in this paper was done using Stata 7.

Table 5
Union Wage Premium Regression Estimates

Data source	CPS			SIPP, CPS match			SIPP							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	OLS	FD	FD WLS1	OLS	FD	FD WLS1	OLS	FD	FD WLS1	FD WLS2	OLS	FD	FD WLS1	FD WLS2
Union	0.160	0.044	0.045	0.185	0.071	0.074	0.181	0.072	0.075	0.076	0.188	0.071	0.075	0.075
	(0.003)	(0.005)	(0.005)	(0.004)	(0.008)	(0.006)	(0.004)	(0.007)	(0.006)	(0.006)	(0.004)	(0.006)	(0.006)	(0.006)
Education	0.098	–	–	0.090	–	–	0.091	–	–	–	0.091	–	–	–
	(0.001)			(0.001)			(0.001)				(0.001)			
Experience	0.036	0.001	0.001	0.035	0.011	0.010	0.036	0.013	0.010	0.011	0.035	0.036	0.009	0.008
	(0.000)	(0.003)	(0.003)	(0.001)	(0.008)	(0.008)	(0.000)	(0.008)	(0.007)	(0.007)	(0.001)	(0.002)	(0.007)	(0.007)
Experience^2/100	−0.059	−0.032	−0.030	−0.057	−0.027	−0.024	−0.057	−0.031	−0.029	−0.030	−0.057	−0.032	−0.029	−0.029
	(0.001)	(0.006)	(0.004)	(0.001)	(0.004)	(0.004)	(0.001)	(0.036)	(0.003)	(0.003)	(0.001)	(0.004)	(0.004)	(0.004)
Black	−0.104	–	–	−0.116	–	–	−0.118	–	–	–	−0.118	–	–	–
	(0.005)			(0.006)			(0.006)				(0.006)			
Female	−0.261	–	–	−0.262	–	–	−0.252	–	–	–	−0.252	–	–	–
	(0.003)			(0.004)			(0.004)				(0.004)			
Move	–	–	–	–	–	–	–	–	–	–	−0.045	–	–	–
											(0.018)			
Union × Move	–	–	–	–	–	–	–	–	–	–	−0.011	0.010	0.013	0.013
											(0.012)	(0.017)	(0.016)	(0.008)
Constant	5.352	–	–	5.435	–	–	5.405	–	–	–	−5.412	–	–	–
	(0.011)			(0.015)			(0.014)				(0.014)			
Year and month dummy variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Test of hetero-scedasticity (p-value)	–	4.46 (<0.01)	–	–	4.09 (<0.01)	–	–	3.52 (<0.01)	–	–	–	3.81 (<0.01)	–	–
Hausman test	–	–	–	–	–	–	–	0.113	0.303	0.583	–	–	–	–
Observations	76005	76005	76005	52694	52694	52694	61882	61882	61882	61882	61882	61882	61882	61882
R^2	0.350	–	–	0.321	–	–	0.325	–	–	–	0.321	–	–	–

FD and FD WLS denote first-difference and first-difference weighted least squares. For each first-difference estimation, a test for heteroscedasticity was implemented through a regression of the squared residual on the predicted dependent variable and its square. The results from this auxiliary regression were used to obtain the weight used in standard weighted least squares (WLS1), in particular, the inverse of the square root of the predicted variance (see Wooldridge [33], pp. 259–260). In WLS2, the underrepresentation of movers because of attrition related to moving is corrected through up-weighting observations on movers by the factor 1.346; since the number of matched movers is 24330 and there are 8417 attriters due to moving, each matched mover represents 1.346 such attriters ($\{24330 + 8417\}$/24330). Standard errors are reported in parentheses. Standard errors of OLS estimates are robust against individual clustering. Union is the dummy for union membership. Experience is potential experience. The number of observations reported is twice the number of individuals in the data set. In columns (11)–(14) all variables shown in the table are interacted with the dummy variable for move, to ensure that the coefficient on Union × Move does not reflect differences in other coefficients; without these additional interactions, though, the results for Union × Move were nearly identical. The Hausman test is calculated for the single variable "Union," using the formula $H = (g_{cps} - g_{sipp})^2/(Var(g_{cps}) - Var(g_{sipp}))$, where g_{cps} and g_{sipp} are the estimated coefficients of union using the SIPP (CPS match) and the full SIPP match. The reported test statistic has a χ^2 distribution with 1 degree of freedom; the 5% critical value is 3.84.

Table 6
Male Marriage Premium Analysis, Sample Details and Descriptive Statistics

	CPS sample		SIPP sample	
Sample construction				
Total matched observations	103325		82735	
Age>=16	77231		60260	
Employed at both month 4 and 16	48498		39367	
Employed by private firms or government	38707		32409	
Wage is available for both month 4 and 16	38451		31752	
Marital status is available for both month 4 and 16	38451		31752	
All explanatory variables for wage regression are available	38355		31752	
CPS match	–		27022	
Matched movers	–		4730	
Descriptive statistics				
Married	0.699		0.664	
Hourly rate of pay (calculated)	1340		1317	
Hours worked per week	41.44		42.08	
Weekly earnings	573		–	
Monthly earnings	–		2410	
Weeks worked in a month	–		4.286	
Marriage transitions (%)				
Not married – not married	28.48		31.11	
Not married – married	1.60		2.44	
Married – not married	1.07		1.41	
Married – married	68.85		65.04	
		SIPP non-movers		SIPP movers
No change in marital status		97.76		86.96
Change in marital status		2.24		13.04

See notes to Table 4. Married is the dummy variable for married and spouse present.

significant. A downward bias in the non-mover sample – or equivalently larger effects for attriters – is a reasonable expectation, for the same reasons discussed in relation to Eqs (1)–(4) and Fig. 1.

6.2. Marriage wage premium for men

Our second analysis focuses on the estimation of marriage premia for men using matched CPS files. Table 6 parallels Table 4, providing descriptive statistics for the analysis sample. The descriptive statistics are very similar across the CPS and SIPP samples. The bottom panel compares marital status transitions across the two data sets, showing that changes are slightly more common in the SIPP sample. Although the differences in levels are small, the relative importance of the differences are not negligible; for example, only 2.67% of the CPS sample experienced a transition, compared with 3.85% of the SIPP sample, and it is the subsamples of changers that identify the marriage premium in the first-difference estimation. Finally, the last rows of the table show the relationship between changes in marital status and moving. Here, the differences are pronounced. In particular, 13.04% of movers changed marital status, compared with only 2.24% of non-movers. This difference clearly indicates that marriage or divorce is likely to be associated with change of residence.

Table 7 – which has the same structure as Table 5 – reports the regression results. Looking at the CPS data, in the first three columns, the cross-sectional estimate of the marriage premium is 0.142, while the fixed effects estimate is near zero and statistically insignificant.[16] By way of comparison, also looking at men Chun and Lee [5] report a cross-sectional estimate of 0.117 using March CPS data from 1999 restricted to those aged 18 to 40, and Gray [11] reports a cross-sectional estimate of 0.058 and a fixed-effects estimate of 0.014 using NLSY79 data from 1989 to 1993, for those aged 24–34.

The estimation results using the SIPP data with the CPS match are reported in columns (4)–(6). The results are almost identical to the results using the CPS sample. This assures that the SIPP data with the CPS match serve well as a baseline to evaluate attrition bias. The key results using the SIPP sample that includes the matched movers are reported in columns (7)–(14). The cross-sectional estimate in column (7) is close to the estimate based on the SIPP data with the CPS match (0.146 vs. 0.156). However, the first-difference estimates, ranging from 0.014 to 0.018, are a bit larger (and positive), although not statistically significant. The significance of the difference in the estimates based on the SIPP data using the CPS match vs. the SIPP data with the full match are formally tested using the Hausman test, and the equivalence of the estimates is rejected at the 5% level in the FD estimates without weighting, and nearly rejected based on the weighted estimates. The higher longitudinal first-difference estimates using the full match in the SIPP imply that the estimates based only on non-movers are subject to attrition bias that understates the effects of marriage, which is, again, as expected. The finding of more evidence of attrition bias in estimating marriage premia is not surprising if we remember (based on Table 6) that the probability of moving is much higher for those who change marital status.

Nonetheless, while the formal statistical test tends to point to attrition bias, the longitudinal estimates using the CPS match vs. the full match are not substantively different, suggesting that in this particular context, at least, the attrition bias we do find is not much of a concern. For a closer examination of the attrition bias, the results reported in columns (11)–(14) allow us to test the equivalence of the marriage premium between movers and non-movers. The first-difference estimates indicate

[16] These results suggest that the marriage premium found in the cross-sectional estimate is due to selection rather than a productivity effect or discrimination. However, Korenman and Neumark [18] note that if the marriage premium grows over time (perhaps because marriage is associated with greater human capital investment), we may find no effect of marriage in a short first difference, because the effective change in years married is very small. They present evidence is consistent with this latter explanation, although that is not the focus of this paper.

Table 7
Male Marriage Wage Premium Regression Estimates

Data source	CPS			SIPP, CPS match			SIPP							
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)
	OLS	FD	FD WLS	OLS	FD	FD WLS1	OLS	FD	FD WLS1	FD WLS2	OLS	FD	FD WLS1	FD WLS2
Married	0.142	−0.003	−0.002	0.156	−0.006	−0.007	0.146	0.018	0.014	0.016	0.149	0.005	−0.000	−0.000
	(0.005)	(0.013)	(0.013)	(0.007)	(0.018)	(0.018)	(0.006)	(0.012)	(0.013)	(0.013)	(0.006)	(0.014)	(0.015)	(0.015)
Education	0.090	–	–	0.083	–	–	0.084	–	–	–	0.084	–	–	–
	(0.001)			(0.001)			(0.001)				(0.001)			
Experience	0.043	0.002	−0.000	0.041	0.006	0.006	0.042	0.013	0.011	0.013	0.041	0.011	0.009	0.010
	(0.001)	(0.004)	(0.004)	(0.001)	(0.011)	(0.018)	(0.001)	(0.010)	(0.011)	(0.011)	(0.001)	(0.010)	(0.011)	(0.011)
Experience^2/100	−0.069	−0.041	−0.035	−0.065	−0.025	−0.020	−0.066	−0.033	−0.029	−0.030	−0.065	−0.031	−0.028	−0.028
	(0.001)	(0.008)	(0.006)	(0.002)	(0.006)	(0.005)	(0.002)	(0.005)	(0.005)	(0.005)	(0.002)	(0.005)	(0.005)	(0.005)
Black	−0.140	–	–	−0.156	–	–	−0.159	–	–	–	−0.159	–	–	–
	(0.008)			(0.009)			(0.009)				(0.009)			
Move	–	–	–	–	–	–	–	–	–	–	−0.010	–	–	–
											(0.028)			
Married × Move	–	–	–	–	–	–	–	–	–	–	−0.056	0.049	0.040	0.037
											(0.014)	(0.024)	(0.026)	(0.025)
Constant	5.327	–	–	5.385	–	–	5.368	–	–	–	5.376	–	–	–
	(0.015)			(0.019)			(0.018)				(0.018)			
Year and month dummy variables	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Test of hetero-scedasiticity (p-value)	–	3.07 (<0.01)	–	–	3.60 (<0.01)	–	–	2.93 (<0.01)	–	–	–	1.89 (0.016)	–	–
Hausman test	–	–	–	–	–	–	–	4.877	3.165	3.473	–	–	–	–
Observations	76710	76710	76710	54044	54044	54044	63504	63504	63504	63504	63504	63504	63504	63504
R^2	0.341	–	–	0.310	–	–	0.313	–	–	–	0.313	–	–	–

See notes to Tables 5 and 6. In columns (11)–(14) all variables shown in the table are interacted with the dummy variable for move, to ensure that the coefficient on Married × Move does not reflect differences in other coefficients; without these additional interactions, though, the results for Married × Move were nearly identical. The Hausman test is calculated for the single variable "Married."

positive and marginally significant increases in wages for workers who marry and move, compared with no effect (a relatively precise estimate of zero) for non-movers.

The analysis of the marriage premium for males suggests that using matched CPS files to obtain longitudinal estimates of this premium entails some risk of underestimating the premium, due to attrition bias. However it is important to note that the large drops in the estimated marriage premium from around 0.15 in the cross-sectional estimates to 0 to 0.02 in the longitudinal estimates were found regardless of the sample used for estimation, and that the size of the attrition bias, which is at most around 0.02, is small in comparison.

7. Conclusion

While matching CPS panels is a popular means of forming panel data sets permitting longitudinal estimation, because the CPS does not follow residential movers these panel data sets suffer high attrition. This paper analyzes whether this attrition leads to bias in the longitudinal estimation of some standard behavioral relationships in labor economics: the effects of unions on wages; and the marriage wage premium for males. This question is explored using SIPP data to first mimic the CPS by excluding all residential movers, and then exploiting the fact that the SIPP successfully follows many movers to construct a data set with far less attrition. To the extent that the SIPP panels including movers constitute a random sample of the population, comparison of results using the "CPS-style" match with the SIPP, and using the full match, should be informative regarding the extent of attrition bias in these estimated relationships. Because the CPS and SIPP are very similar surveys aside from this difference in attrition due to moving, this analysis of attrition bias in the SIPP should be informative about attrition bias in the CPS.

The results for the longitudinal analysis of union wage effects reveal negligible and statistically insignificant evidence of attrition bias despite the high attrition rate in the matched CPS files. In contrast, the longitudinal analysis of the marriage premium for males finds statistically significant attrition bias. The amount of bias, however, does not seem to be serious in an economic sense. There is no way to draw definitive conclusions about whether these results generalize to other applications using matched CPS files. However, one point that is reassuring is that we found little evidence of attrition bias even in a case where attrition was very strongly related to the identifying information (changes in marital status). Thus, we regard the evidence as suggesting that in many applications the advantages of using matched CPS panels to obtain longitudinal estimates are likely to far outweigh the disadvantages from attrition biases, although we should allow for the possibility that attrition bias leads the longitudinal estimates to be understated.

In contrast, we suspect that attrition bias is likely to be more severe when the change in the dependent variable caused by the change in the independent variable of interest largely determines whether there is attrition in matched CPS panels. In such a case, a healthy suspicion of attrition bias is probably warranted, and it may be possible to use the data and methods we have proposed in this paper to assess the extent of attrition bias.

Acknowledgements

This paper reflects the views of the authors, and not those of the Public Policy Institute of California.

References

[1] S. Adams and D. Neumark, Do living wage ordinances reduce urban poverty? *Journal of Human Resources* **38** (2003), 490–521.

[2] S. Becketti, W. Gould, L. Lillard and F. Welch, The Panel Study of Income Dynamics after fourteen years: an evaluation, *Journal of Labor Economics* **6** (1988), 472–492.

[3] Bureau of the Census, CPS Technical Paper 63: Design & Methodology, US Bureau of the Census, Washington, DC, 1997.

[4] C.F. Citro and G. Kalton, *The future of the Survey of Income and Program Participation*, National Academy Press, Washington, DC, 1993.

[5] H. Chun and I. Lee, Why do married men earn more: productivity or marriage selection? *Economic Inquiry* **39** (2001), 307–319.

[6] J. Coder and L. Scoon-Rogers, *Evaluating the quality of income data collected in the Annual Supplement to the March Current Population Survey and the Survey of Income and Program Participation*, US Bureau of the Census, Washington, DC, 1996.

[7] H.O. Duleep and M.C. Regets, Measuring immigrant wage growth using matched CPS files, *Demography* **34** (1997), 239–249.

[8] E.M. Falaris and H.E. Peters, Survey attrition and schooling choices, *Journal of Human Resources* **33** (1998), 531–554.

[9] J. Fitzgerald, P. Gottschalk and R. Moffitt, An analysis of sample attrition in panel data, *Journal of Human Resources* **33** (1998), 251–299.

[10] R.B. Freeman, Longitudinal analyses of the effects of trade unions, *Journal of Labor Economics* **2** (1984), 1–26.

[11] J.S. Gray, The fall in men's return to marriage: declining productivity effects or changing selection? *Journal of Human Resources* **32** (1997), 481–504.

[12] M. Gittleman and M. Joyce, Earnings mobility and long-run inequality: an analysis using matched CPS data, *Industrial Relations* **35** (1996), 180–196.

[13] L. Goldberg, J. Tracy and S. Aaronson, Exchange rates and employment instability: evidence from matched CPS data, *American Economic Review* **89** (1999), 204–210.

[14] J.J. Heckman, Sample selection bias as a specification error, *Econometrica* **47** (1979), 153–161.

[15] J.J. Heckman, Shadow prices, market wages, and labor supply, *Econometrica* **42** (1974), 679–694.

[16] S.L. Hofferth, W-J.J. Yeung and F.P. Stafford, *Panel Study of Income Dynamics*, ICPSR, 1996.

[17] G. Jakubson, Estimation and testing of the union wage effect using panel data, *Review of Economic Studies* **58** (1991), 971–991.

[18] S. Korenman and D. Neumark, Does marriage really make men more productive? *Journal of Human Resources* **26** (1991), 282–307.

[19] D.A. Macpherson and B.T. Hirsch, Wages and gender composition: why do women's jobs pay less? *Journal of Labor Economics* **13** (1995), 426–471.

[20] S. Mack and R. Petroni, *Overview of SIPP nonresponse research*, US Bureau of the Census, 1994.

[21] T. MaCurdy, T. Mroz and R.M. Gritz, An evaluation of the National Longitudinal Survey of Youth, *Journal of Human Resources* **33** (1998), 346–436.

[22] B.C. Madrian and L.J. Lefgren, An approach to longitudinally matching Current Population Survey (CPS) respondents, *Journal of Economic and Social Measurement* **26** (2000), 31–62.

[23] D. Neumark and W. Wascher, Using the EITC to increase family earnings: new evidence and a comparison with the minimum wage, *National Tax Journal* **54** (2001), 281–317.

[24] D. Neumark and W. Wascher, The effects of minimum wages on teenage employment and enrollment: estimates from matched CPS data, *Research in Labor Economics* **15** (1996), 25–63.

[25] D. Neumark, M. Schweitzer and W. Wascher, The effects of minimum wages on the distribution of family incomes: a non-parametric analysis, *NBER Working Paper No.* **6536** (1998).

[26] R.J. Olsen, A least squares correction for selectivity bias, *Econometrica* **48** (1980), 1815–1820.

[27] A. Pitts, *Matching adjacent years of the Current Population Survey*, mimeograph, Unicon Research Corporation, Santa Monica, CA, 1988.

[28] F. Peracchi and F. Welch, How representative are matched cross-sections? Evidence from the Current Population Survey, *Journal of Econometrics* **68** (1995), 153–179.

[29] P. Rhoton, *Attrition and the National Longitudinal Surveys of Labor Market Experience: avoidance, control and correction*, Center for Human Resource Research, The Ohio State University, 1984.

[30] E.J. Schumacher, Relative wages and exit behavior among registered nurses, *Journal of Labor Research* **18** (1997), 581–592.

[31] B.S. Shippen, Unmeasured skills in inter-industry wage differentials: evidence from the apparel industry, *Journal of Labor Research* **20** (1999), 161–169.

[32] F. Welch, Matching the Current Population Surveys, in: *Stata Technical Bulletin Reprints*, (Vol. 2), J. Hilbe, ed., Computing Resource Center, Santa Monica, CA, 1993, pp. 34–40.

[33] J.M. Wooldridge, *Introductory Econometrics*, South Western College Publishing, Cincinnati, OH, 2000.

[34] J. Zagorsky and P. Rhoton, *Attrition and the National Longitudinal Surveys' women cohorts*, Center for Human Resource Research, The Ohio State University, 1999.