Journal of Economic and Social Measurement 26 (2000) 31–62
IOS Press

An approach to longitudinally matching Current Population Survey (CPS) respondents

Brigitte C. Madrian*
University of Chicago and National Bureau of Economic Research, Chicago, IL, USA

Lars John Lefgren
University of Chicago, Chicago, IL, USA

In this paper, we propose an approach for evaluating the trade-offs inherent in different approaches used to match Current Population Survey (CPS) respondents across various CPS surveys. Because there is some measurement error in both the variables used the identify individuals over time and in the characteristics of individuals at any point in time, any procedure used to match CPS respondents has the possibility of both generating incorrect matches and failing to generate potentially valid matches. We propose using the information contained in the variable on whether an individual lived in the same house on March 1 of the previous year as a way to gauge these trade-offs. We find that as measured by reported residence one year ago, increasing the fraction of "invalid" merges that are rejected usually comes at a cost of decreasing the fraction of "valid" merges that are retained. However, there are clearly some approaches that are superior to others in the sense that they result in both a higher fraction of "invalid" merges being rejected and a higher fraction of "valid" merges being retained.

1. Introduction

In the United States, the Current Population Survey (hereafter referred to by its well-known acronym CPS) is the workhorse data set used in empirical analyses of the labor market. There are several reasons for this. First, the CPS is well-suited for analyzing long-term labor market trends, with publicly available microdata going back to 1962 and statistical tabulations going back to the 1930s [1]. Second, the wealth of information collected is diverse, including regularly collected information on things such as labor force participation, unemployment, wages, unionization, hours worked, industry, occupation, education, income, pension and health insurance coverage, and irregularly collected information on things such as job tenure, smoking, computer usage, voting, fertility, and immigration. And third, the size of the CPS samples is generally large enough for reliable statistical inference yet small enough

*Corresponding author: Brigitte C. Madrian, University of Chicago, Graduate School of Business, 1101 E. 58th Street, Chicago, IL 60637, USA. Tel.: +1 773 702 8079; Fax: +1 773 702 0458; E-mail: brigitte.madrian@gsb.uchicago.edu.

to be tractable.[1] As stated in the BLS Handbook of Methods, the primary purpose of the CPS is to "classify the sample population into three basic economic groups: The employed, the unemployed, and those not in the labor force." To this end, a sample of households is contacted each month and household residents are asked a series of questions designed to clarify their labor force status. In some months, households are asked questions on other topics as well, many of which are also related to the labor market, but which may encompass other things as well (e.g. voting behavior).

The vast majority of empirical analyses of the CPS either use only one month of CPS data (a single cross-section), or use a series of CPS surveys, treating them as a series of repeated cross-sections.[2] While most serious users of the CPS are aware that there is, in fact, a longitudinal component to the CPS (described in greater detail in the next section), far fewer empirical analyses exploit this aspect of the CPS. There are two reasons for this. The first, and probably more important reason, is that as a longitudinal panel, the CPS is of very short duration and thus of more limited value than standard longer-term panel datasets such as the National Longitudinal Surveys (NLS), the Panel Study of Income Dynamics (PSID), or even the Survey of Income and Program Participation (SIPP). The second reason is that matching individuals across two (or more) different CPS surveys is not completely straightforward, and the CPS documentation gives very little guidance on how best to do this. The aim of this paper is threefold: first, to describe the longitudinal aspect of the CPS and the potential for matching individuals across survey months; second, to outline the issues involved in accurately matching individuals across survey months for the CPS; and third, to evaluate the trade-offs in different approaches to matching individuals across survey months. Hopefully, this paper will provide some guidance for future users of the CPS interested in exploiting the limited longitudinal nature of the CPS.

Our analysis will focus on matching consecutive March CPS surveys from 1980 to 1998.[3] We confine ourselves to the period after 1980 because we can use a more-or-less consistent criterion for the entire 1980–1998 period. Prior to 1980, several changes, and even omissions, in the variables that identify households and individuals make matching individuals from two consecutive years difficult to impossible for much of the pre-1980 period.[4] For the years prior to 1980 in which it is possible to

[1] There are, of course, exceptions to this generalization. For example, the CPS sample may not be large enough for reliable statistical inference if the sample is sufficiently restricted to smaller geographic units (a particular state or metropolitan area) or particular demographic segments of the population. For these reasons, some researchers have proposed dramatically increasing the size of the CPS [6].

[2] Some recent papers that have merged various CPS surveys include [3,4,7,9,10,12].

[3] Note that neither March 1984 and March 1985, nor March 1994 and March 1995, can be merged. This results from revisions in the household identifiers implemented to protect the confidentiality of survey respondents following revisions in the CPS geographic identifiers. These revisions also affect the ability to match consecutive months during the 1984–1985 and 1994–1995 time periods.

[4] For example, merging the March 1971 and 1972 surveys is precluded by a change in household identifiers, as is merging 1972 to 1973, and 1976 to 1977. Merging households from 1977 to 1978 and 1978 to 1979 is possible, but the omission of individual identifiers from 1977–1979 makes matching individuals tenuous.

match individuals, several algorithms have been proposed and evaluated by others. For more information on these approaches, consult [5,11,14,16].

2. CPS sample design

The CPS is a monthly survey of a probability sample of housing units. It does not, however, survey a completely new set of housing units each month. Rather, the sample is divided into eight representative subsamples called rotation groups, with housing units in each rotation group being interviewed for four consecutive months, followed by an 8-month break, and then by another four months of interviews. Thus, CPS sample housing units are each eligible for 8 different monthly interviews, and rotation groups are referred to in CPS parlance by their "month in sample" or MIS. In any given monthly sample, approximately one-eighth of sample units will be interviewed for the first time (MIS $=$ 1), one-eighth for the second time (MIS $=$ 2), and so on. One-eighth of the sample will be leaving the sample permanently (MIS $=$ 8), and one-eighth will be leaving for the next eight months before being reinterviewed (MIS $=$ 4). These latter two rotation groups, MIS $=$ 8 and MIS $=$ 4, are referred to as the "outgoing rotation groups."

Table 1 illustrates the composition of any monthly CPS sample with respect to these different rotation groups and the cycling of the various rotation groups through the CPS sample. Rotation group A is first interviewed in January of year t (MIS $=$ 1). It is subsequently interviewed in February, March and April of the same year (MIS $=$ 2, 3 and 4 respectively). Following this fourth interview in April of year t, rotation group A then leaves the sample for 8 months and is next interviewed in January of year t+1 (MIS $=$ 5). It continues to be interviewed in February, March and April of year $t+1$ (MIS $=$ 6, 7 and 8), and then leaves the sample permanently after the 8th and final interview in April of year $t+1$. The January $t+1$ sample of the CPS is comprised in part of rotation group A (MIS $=$ 5), along with rotation groups M, L, K and J (MIS $=$ 1, 2, 3 and 4 respectively), and rotation groups z, y and x (MIS $=$ 6, 7 and 8).

Table 1 also illustrates the month-to-month overlap in the CPS sample, along with the potential for matching rotation groups across time in the various CPS interviews. Comparing January of year t with February of year t (or any other two consecutive months) shows that 75% of the CPS sample is common from month to month (in this case, rotation groups A, z, y, o, n, and m); while comparing January of year t with January of year $t+1$ shows that 50% of the CPS sample is common from one year to the next for the same month (rotation groups A, z, y, and x). Fig. 1 shows more generally the fraction of any CPS monthly sample which is common with surrounding months up to 15 months away (beyond 15 months there is no intentional overlap in the CPS samples). Note that because rotation groups drop out of the CPS sample for a period of time, CPS surveys separated by between 4 and 8 months will have no part of their samples which are common, even though surveys separated by 1 to 3

Table 1
Rotation groups in the current population survey

	Month in sample (MIS)							
	1	2	3	4	5	6	7	8
Year t								
January	A	z	y	x	o	n	m	l
February	B	A	z	y	p	o	n	m
March	C	B	A	z	q	p	o	n
April	D	C	B	A	r	q	p	o
May	E	D	C	B	s	r	q	p
June	F	E	D	C	t	s	r	q
July	G	F	E	D	u	t	s	r
August	H	G	F	E	v	u	t	s
September	I	H	G	F	w	v	u	t
October	J	I	H	G	x	w	v	u
November	K	J	I	H	y	x	w	v
December	L	K	J	I	z	y	x	w
Year $t+1$								
January	M	L	K	J	A	z	y	x
February	N	M	L	K	B	A	z	y
March	O	N	M	L	C	B	A	z
April	P	O	N	M	D	C	B	A
May	Q	P	O	N	E	D	C	B
June	R	Q	P	O	F	E	D	C
July	S	R	Q	P	G	F	E	D
August	T	S	R	Q	H	G	F	E
September	U	T	S	R	I	H	G	F
October	V	U	T	S	J	I	H	G
November	W	V	U	T	K	J	I	H
December	X	W	V	U	L	K	J	I
Year $t+2$								
January	Y	X	W	V	M	L	K	J
February	Z	Y	X	W	N	M	L	K
March	*a*	Z	Y	X	O	N	M	L
April	*b*	*a*	Z	Y	P	O	N	M
May	*c*	*b*	*a*	Z	Q	P	O	N
June	*d*	*c*	*b*	*a*	R	Q	P	O
July	*e*	*d*	*c*	*b*	S	R	Q	P
August	*f*	*e*	*d*	*c*	T	S	R	Q
September	*g*	*f*	*e*	*d*	U	T	S	R
October	*h*	*g*	*f*	*e*	V	U	T	S
November	*i*	*h*	*g*	*f*	W	V	U	T
December	*j*	*i*	*h*	*g*	X	W	V	U

Note: Each letter/typeface combination represents a different rotation group.

and 9 to 15 months will. In any given month, 87.5% (7/8) of the CPS sample will be included in at least one future CPS (only sample units with MIS $= 8$ will have not be in any future surveys); similarly, 87.5% of the sample will also have been in at least one previous CPS (only sample units with MIS $= 1$ will have no previous CPS history).

3. Issues in matching individuals across various CPS surveys

While Fig. 1 illustrates the sample overlap across time between various CPS surveys, it substantially overstates the actual fraction of individual respondents that can be matched across surveys. There are at several reasons for this.

3.1. Non-response

Currently, about 59,000 housing units are designated for data collection each month, of which about 50,000 are occupied and eligible for interview. Of these units, about 6–7% are not interviewed due to "temporary absence (vacation, etc.), other failures to make contact after repeated attempts, inability of persons contacted to respond, unavailability for other reasons, and refusals to cooperate" [2]. This type of non-response will clearly reduce the fraction of CPS respondents that can be matched across time.

3.2. Mortality

Obviously, individuals who die between two different survey months will not be able to be matched from the first to the second survey. The effect of mortality on the matching of CPS respondents across time will be trivial at younger ages, although at older ages it could be quite substantial. For example, to take an extreme case, the annual morality rate in 1996 for individuals aged 85 and older exceeded 15%. This is more than double the average non-response rate in the CPS.

3.3. Migration

While mortality may be important at older ages, residential mobility is likely to be the most important explanation at younger ages for why individuals cannot be matched across time in the CPS. Recall that the CPS is a sample of housing units and not a sample of individuals. Thus, residents *at a particular address* designated for inclusion in the CPS sample will be interviewed following the pattern in Table 1. If the individuals originally interviewed as part of a rotation group move to a new location, they will not be followed to their new location but will be replaced in the survey by the new occupants of the original housing unit, if any. Statistical tabulations from the CPS on residential mobility suggest that between 15% and 20% of the population report living at a different address 12-months previously. Annual migration rates of this magnitude will result in a substantial reduction in the fraction of CPS respondents that can be matched across time, particularly as the time between two different surveys increases.

3.4. Recording errors

In addition to the systematic reductions in the merge rate that result from migration, mortality and non-response, recording errors in the variables that serve to identify individuals over time also influence the match rate in the CPS. Beginning in 1980, individuals *at a point in time* are uniquely identified in the CPS by two variables: a household identifier (HHID), and an individual line number within the household (LINENO).[5,6] In theory, these two identifiers should not change over time: HHID should remain constant for the same housing unit, and LINENO should remain constant for the same individual within a household. But a particular combination of HHID and LINENO does not necessarily uniquely identify individuals *across time* in the CPS because the same combination of HHID and LINEO may be given to *a different* CPS respondent if the original respondent moves away from a sampled housing unit and a new respondent moves in. A third variable, HHNUM (household number), is designed to identify these situations. This variable should equal 1 during the initial interview (MIS $= 1$), and will be incremented by 1 during subsequent interviews any time one household is replaced by another. A change in HHNUM for any HHID would indicate that individuals in the household during the current interview were not the same as those in the household in the previous interview. Thus, in theory, one should be able to match CPS respondents across surveys using the unique and intertemporally consistent combination of HHID, HHNUM and LINEO. In practice, however, any combination of HHID, HHNUM, and LINEO is not necessarily intertemporally consistent because of recording errors. For example, HHNUM may change even when the respondents in a household are the same, or an individual's line number may change even though it shouldn't. We suspect that these inconsistencies arise because the CPS was not designed for utilization as a panel dataset, and consequently the data are not thoroughly checked for the recording errors that lead to inconsistency in these individual and household identifiers over time. Because of these recording errors, matching individuals across various CPS surveys on the basis of HHID, HHNUM and LINENO will give rise to "false positives"– matches that do not represent the same individual (this could arise when the line numbers of two individuals are switched from one survey to the next), and "false negatives"– individuals that do not match even though they are respondents in both surveys (this could arise if HHNUM were incorrectly incremented or if the same individual had a different LINEO in two different surveys).

[5] The variable names used in this paper are taken from CPS Utilities [15], a repackaging of the CPS published by Unicon Research Corporation. Many of the original Census Bureau names have changed over time and are not very descriptive. Appendix Table B1 lists the Unicon names of the variables used in the analysis along with their corresponding Census Bureau names.

[6] Although the combination of HHID and LINENO should in theory uniquely identify individuals at a point in time, there are, nonetheless, infrequent instances of the same combination of HHID and LINENO being given to different individuals.

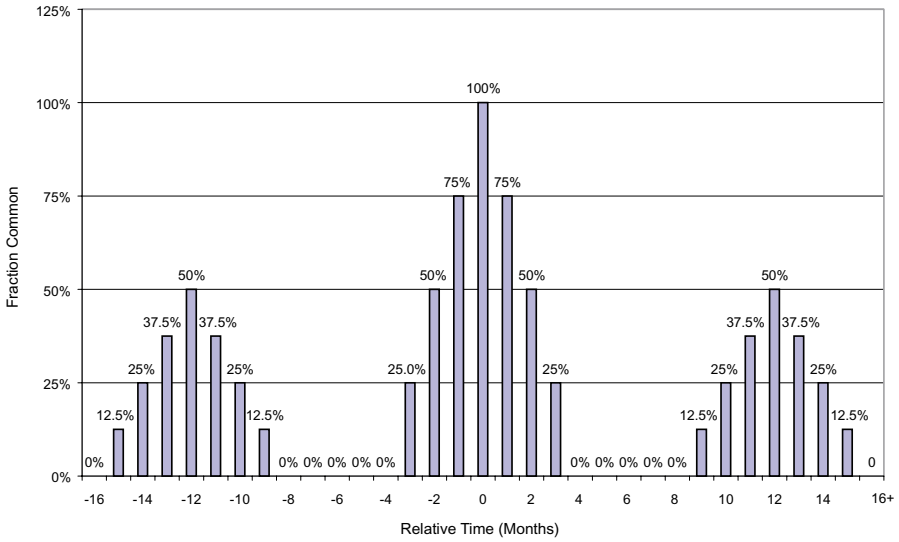

Fig. 1. Fraction of the CPS which is common across months.

In the discussion that follows, we will frequently refer to what we call the "naïve merge rate". This represents all of the merges that occur by matching observations from two different surveys solely on the basis of HHID, HHNUM and LINENO relative to the total number of individuals in time t who could have potentially merged. We use the term "naïve" because this merge rate will include some merges that occur on the basis of HHID, HHNUM and LINENO that do not represent the same individual over time, and it will also exclude some individuals who do not merge on the basis of HHID, HHNUM and LINENO but who are in fact in both of the merged surveys. With no foolproof unique and intertemporally consistent identifier, the researcher is left to his or her own devices in determining what constitutes a valid match, although we will hopefully shed some light on this process in subsequent sections of this paper. Lacking more specific information on non-response and recording errors in the CPS, we cannot examine in much detail how these two factors influence the merging of two CPS surveys. But we can further investigate the effects of migration and mortality. Table 2 shows the relative impact of these two factors on the actual merge rate that is achieved when trying to match individuals from two March CPS surveys on the basis of HHID, HHNUM and LINENO.[7] The statistics in Table 2 are derived from matching the 1980 to 1997 March CPS surveys with the

[7] The March 1994 survey is also matched to the March 1995 survey on the basis of state of residence since HHID, HHNUM and LINEO were only uniquely assigned *within* a particular state for these two years.

Table 2
Factors explaining non-merges in matching the March CPS across years

Respondent age in time t	Naïve merge rate (A)	Naïve non-merge rate (B)	Migration rate (C)		Mortality rate (D)		Residual non-merge rate (E)	
15−19	64.2%	35.8%	16.5%	[46.1%]	0.09%	[0.3%]	19.2%	[53.6%]
20−24	49.3	50.8	35.7	[70.3]	0.11	[0.2]	15.0	[29.5]
25−29	58.3	41.7	31.4	[75.3]	0.12	[0.3]	10.2	[24.4]
30−34	67.2	32.8	21.4	[65.2]	0.15	[0.5]	11.3	[34.3]
35−39	72.8	27.2	15.8	[58.1]	0.20	[0.7]	11.2	[41.2]
40−44	76.0	24.0	12.6	[52.5]	0.25	[1.0]	11.2	[46.5]
45−49	78.2	21.8	10.3	[47.2]	0.38	[1.7]	11.1	[51.0]
50−54	80.1	19.9	8.6	[43.3]	0.59	[3.0]	10.7	[53.7]
55−59	81.7	18.3	7.1	[38.8]	0.93	[5.1]	10.3	[56.1]
60−64	82.7	17.3	6.6	[38.3]	1.46	[8.5]	9.2	[53.3]
65−69	83.9	16.1	5.2	[32.3]	2.16	[13.4]	8.7	[54.3]
70−74	84.1	15.9	4.5	[28.3]	3.27	[20.6]	8.1	[51.1]
75−79	82.8	17.2	4.5	[26.2]	4.93	[28.7]	7.7	[45.1]
80−84	79.4	20.6	4.8	[23.3]	7.68	[37.3]	8.1	[39.4]
85+	72.4	27.6	5.1	[18.5]	15.33	[55.6]	7.2	[25.9]
All	71.0	29.0	16.3	[56.2]	0.86	[3.0]	11.8	[40.8]

Note: Columns A−C and E are derived from calculations made by the authors from the March CPSs from 1980−1998. Column D is from the National Center for Health Statistics and gives the age−specific mortality rate for 1990 [8].

corresponding year-ahead March CPS survey.[8] Column A gives what we call the naïve merge rate in the CPS: the fraction of individuals in time t who have a "match" in time $t+1$ with the same HHID, HHNUM and LINENO conditional on being in a rotation group that is included in time $t+1$. The actual naïve merge rate of about 71% is far from the merge rate of 100% that would obtain in the absence of mortality, migration, non-response in time $t+1$, and changes in household composition between time t and time $t+1$. We call this discrepancy the "non-merge rate": the fraction of individuals in time t who do not appear in time $t+1$ (conditional on being in a rotation group that is included in time $t+1$). Thus, the overall non-merge rate is about 29% (100%–71%).

What accounts for this substantial non-merge rate? In addition to the naïve merge (column A) and non-merge rates (column B), columns C and D of Table 2 give the annual migration and annual mortality rates for each of the age groups listed. Column C shows that as might be expected, there is substantial residential mobility at younger ages, with over one-third of 20–24 year-olds moving from one year to the next. Annual mobility rates also decline quite substantially with age, from a high of over 30% for those in their 20s, to a low of less than 5% for those in their 70s. The percentages in brackets give the fraction of the naïve non-merge rate

[8] Recall that, as noted earlier, the March 1985 and March 1995 CPS surveys cannot be matched to the March 1986 and March 1996 CPS surveys, hence the lack of information on these two years in the tables and figures that follow.

(column B) that could potentially be accounted for by residential mobility. Across all age groups, residential mobility explains about 56% of the non-merge rate, although the substantive importance of this factor varies significantly with age. For example, residential mobility accounts for over 70% of the non-merge rate for those in their 20s, while explaining less than 25% of the non-merge rate for those over age 80.

The effect of mortality on the non-merge rate also works as expected. At younger ages, mortality rates are trivial and account for very little of the non-merge rate. [9] At older ages, however, mortality rates become large enough to be of substantive importance and could potentially explain a nontrivial portion of the non-merge rate. For example, the 15.3% mortality rate of individuals aged 85 and over accounts for 56% of the non-merge rate for this group. Even at somewhat younger ages, for example 60–64 years, the mortality rate of 1.5% explains almost 9% of the non-merge rate.

The cumulative effect of these two factors–residential mobility and mortality–on the non-merge rate is given in column E as the "residual non-merge rate": the non-merge rate net of migration and mortality (columns B-C-D). Migration and mortality alone reduce the overall naïve non-merge rate of 28.9% to a residual non-merge rate of 11.8%. This is a significant reduction, and the overall residual non-merge rate of 11.8% is not too far from the overall sample non-response rate of 6–7%, the third factor noted above that is likely to impact the merging of individuals in the CPS. Column E also shows that the residual non-merge rate is very high at young ages (for example, 19% at ages 15–19), and declines with age to a low of 7.2% for those aged 85 and older. It is difficult to calculate a non-response rate by age for obvious reasons (if we knew the age of the non-respondent, we would no longer have a non-respondent!), so we cannot infer precisely how much of the variation by age in the residual non-merge rate would be accounted for by age-related variation in the non-response rate.

4. Evaluating the validity of merges across two different CPS surveys

As noted earlier, merging solely on the basis of HHID, HHNUM and LINENO is likely to generate some "false positives" – merges of two observations that do not represent the same individual due to recording errors in individual identifiers. How severe is this problem?

4.1. Differences in gender and race

We first compare the gender and race of individuals who appear to merge. Save for measurement error, these two (more or less) immutable characteristics should be

[9] Mortality rates are customarily expressed as a number per 100,000 population. However, to be consistent with the other percentages in Table 2, we have expressed the mortality rates in column D as a fraction of the total population.

		Gender Time $t+1$		
		Male	Female	Total
Gender	Male	98.62%	1.38%	100.0%
Time t	Female	1.19%	98.81%	100.0%

		Race Time $t+1$			
		White	Black	Other	Total
Race	White	99.73%	0.13%	0.14%	100.0%
Time t	Black	1.11%	98.77%	0.12%	100.0%
	Other	3.30%	0.31%	96.40%	100.0%

the same in both time t and time $t+1$. And, indeed, for most merges they are.

As the two matrices above show, only a small fraction of merges are for individuals who are of a different gender or race. Over 98% of males and females in time t are merged to an individual of like gender in time $t+1$; and over 98% of whites and blacks in time t are merged to an individual of the same race in time $t+1$. That the discrepancies on the basis of these two characteristics is so small is reassuring, but correlation in the demographic composition of housing units over time arising from similarities in residential preferences by race and gender suggests that these two factors alone may not be sufficient to positively identify false matches. Moreover, that there are *any* differences in either gender or race clearly suggests the potential of either "false positives" or measurement error.

4.2. Differences in age

We next turn to age – how similar in age are the individuals who appear to merge? Fig. 2A plots a histogram of the increase in age from t to $t+1$ for individuals who appear to merge. In contrast to race and sex, which we would expect not to change, for most individuals age should change, increasing by one year. In theory there could also be some individuals with March birthdays who are the same age from one March survey to the next if they were interviewed just after their birthday in year t and just before their birthday in year $t+1$. Similarly, there could be some individuals with March birthdays whose age increases by two years as a result of being interviewed just prior to their birthday in year t and just after their birthday in year $t+1$. Fig. 2 shows that for 91.6% of the naïve merges, age increases from t to $t+1$ by one year. For another 1.6%, age increases by two years; and for 3.2% age is reported to be the same. Thus, for about 96% of merges, the difference in age between t and $t+1$ is plausible. Nonetheless, some of these matches may not represent the same individual if, for example, a departed household respondent is replaced by a new respondent of about the same age.

The age difference between t and $t+1$ for the remaining 4% of merges is distributed broadly, from -80 years to $+76$ years. Most of these merges likely represent "false" matches. However, measurement error in the age reported at either t or $t+1$ could easily lead to a change in age outside of the plausible range. The scale in Fig. 2A

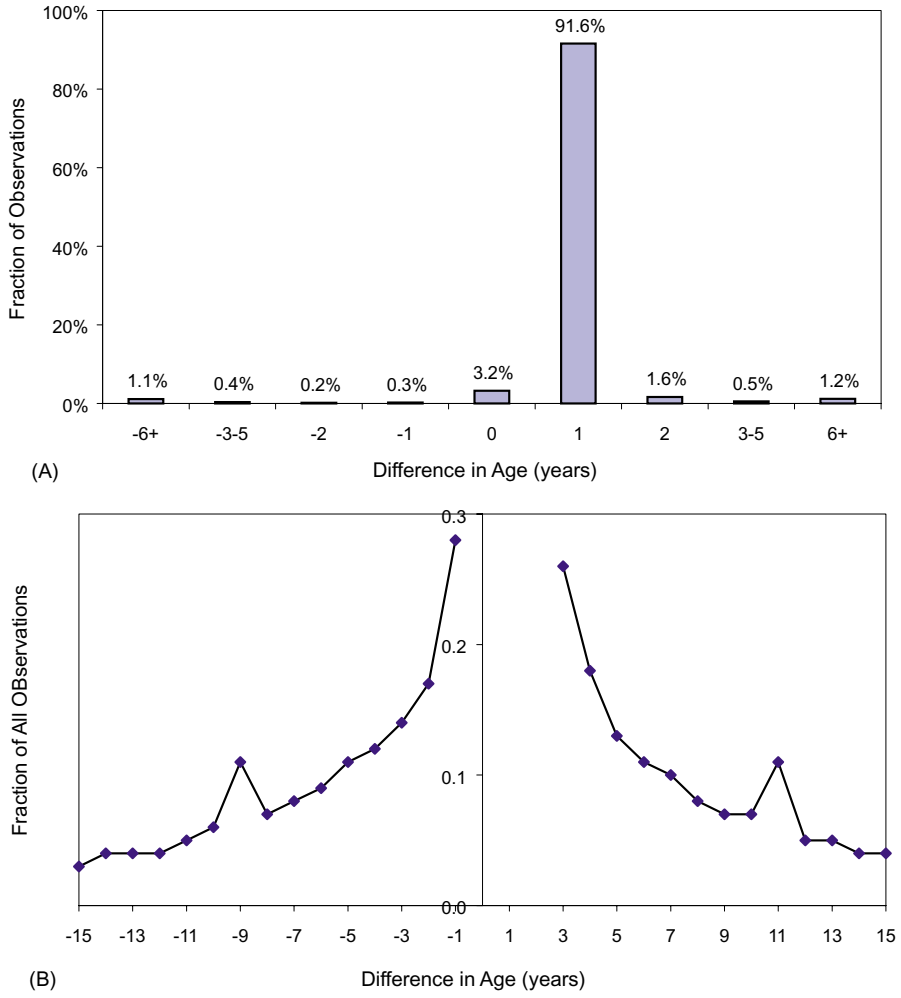

Fig. 2. (A) Distribution of differences in age for matched observations. (B) Distribution of differences in age for selected matched observations.

is so dominated by the mass at 1 year that the variation in the distribution of age differences outside the range of 0–2 years is essentially obscured (and for that reason is not shown – except at highly aggregated levels). But closer inspection reveals that this distribution does suggest the possibility of measurement error. Fig. 2B confines the difference in age to the range from -15 to -1 years and from $+2$ to $+15$ years. Note than none of these age differences individually has any substantial mass, the highest being 0.26% at -1. Nonetheless, there are two things in this Figure that suggest the likelihood of at least some measurement in age for merges that are valid.

First, as the age difference between t and $t+1$ gets farther from 1 (the "expected" age difference), the mass is generally declining.[10] To the extent that measurement error in age in one period tends to be somewhat correlated with the truth rather than random, we would expect the mass to be increasing the closer is the age difference to 1. Second, the exceptions to the general decline in the mass as the age difference gets farther from 1 occur at $+11$ and -9 years: spikes in the probability of these age differences are apparent in Fig. 2B. This is exactly the type of age difference that we would expect to observe if age were miscoded or misreported by exactly 10 years in either t or $t+1$. There is little reason to think that invalid merges would lead to this type of spike, and thus they seem to constitute some evidence of measurement error. Note that these spikes do not imply that measurement error is of an exceptionally large magnitude, at least not at these two points, but they clearly suggest that any rule for determining which merges are "valid" vs. "invalid" on the basis of "excessive" discrepancies in age will lead to some false rejections of validly merged individuals.

4.3. Differences in educational attainment

Like age, educational attainment is another variable that should evolve in a predictable pattern for respondents who are correctly merged, either remaining constant or increasing. Because the questions about and coding of education in the CPS was dramatically changed in 1992, we present two sets of statistics on differences in educational attainment for merged individuals: the first for the period from 1980–1990 when education was consistently coded using the old coding scheme, and the second for the period from 1992–1997 when it was consistently coded using the new coding scheme.[11]

Prior to 1992, two questions were asked about educational attainment in the CPS, the first about the number of years of school attended, and the second about whether the last year of school was completed. The resulting education variables are a linear "years of school" variable (GRDHI) ranging from 1–19 from 1980–1987 and from 0–18 from 1988–1991, and an indicator for whether the final year was completed (GRDCOM). We make the years of school variable consistent from 1980–1991 by

[10] Note that randomly matching individuals from two CPS surveys would also lead to a distribution of age differences that is centered around 1 and that declines as age difference moves farther away from 1. However, the distribution in age difference that results from randomly matching individuals across two CPS surveys is much less convex than the distribution in Figs 2A and 2B. Thus, the evidence of measurement error in age even for valid matches is not just that the mass is declining as age difference moves further from 1, but also that the distribution is so convex.

[11] For 1991, education in t and $t+1$ were coded differently, the former according to the old coding scheme and the latter according to the new. We have evaluated the differences in education for this year after attempting to make the two different coding schemes for education equivalent. Since, as shown in Fig. 3, the old and new coding schemes paint a similar picture with respect to differences in educational attainment between t and $t+1$, we simply note that for 1991 the picture is similar as well. The details of this are presented in the Appendix.

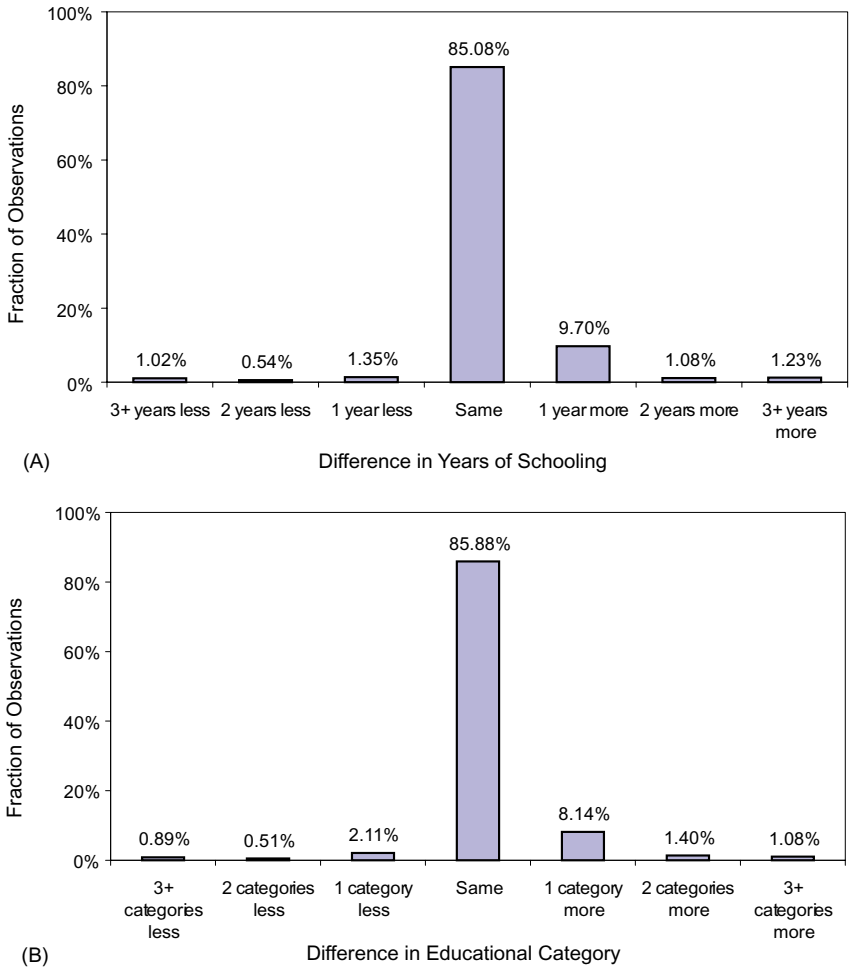

Fig. 3. (A) Distribution of differences in educational attainment for matched observations, 1980–1990. (B) Distribution of differences in educational attainment for matched observations, 1992–1997.

subtracting 1 for the period from 1980–1987. We also subtract 1 if the final year of school attended was not completed (we do this for all years of schooling except 0, which corresponds to kindergarten). The resulting variable measures years of school completed, and differences in this variable from t to $t+1$ measure differences in years of school completed.

Beginning in 1992, education was no longer classified according to years of school, but according to the highest level of educational attainment classified into categories. Some of the categories correspond to degrees (e.g. doctorate degree, professional degree, associates degree, high school degree), others to more arbitrary measures

(e.g. 1–4 years of school, 5–8 years of school, grade 9, grade 10, grade 11, etc.). Movement from one category to another could represent an incremental change in schooling of less than a year (for example, the movement from 12th grade no diploma to high school graduate), exactly a year (for example, 9th grade to 10th grade), or more than a year (for example, grades 1–4 to grades 5–6, or bachelor's degree to doctorate degree). We have classified educational differences according to changes in educational category which we have ordered based on whether or not a movement from one to another would be plausible. So, for example, since a bachelor's degree could plausibly be followed by either a master's degree, a professional degree, or a doctorate degree, we classify a movement from the bachelor's degree category in time t to any of these other three categories in time $t+1$ as having achieved "1 category more" in education. The appendix provides more detail on how we categorize differences in the level of educational attainment between t and $t+1$.

Figs 3A (1980–1990) and B (1992–1997) show the distribution of one-year differences in educational attainment for individuals who appear to merge. As anticipated, there are substantial masses at 0, approximately 85% of merges in both 3A (1980–1990) and 3B (1992–1997), and at 1, about 10% of merges in 3A (1980–1990) and 8% in 3B (1992–1997). In addition, 1–2% of merges show an educational discrepancy of either 1 year (category) less, or 2 years (categories) more, differences that might be plausible with measurement error in education in either t or $t+1$. As with the distribution of age differences, the distribution of educational differences is generally declining as we move away from the "expected" difference of 0 or 1, and less than 3% of merges have educational differences that fall outside the range of -1 year (category) to $+2$ years (categories).

5. Tradeoffs in establishing "valid" merge criterion

The discussion in the last section clearly points to some infrequent but large discrepancies in the age and educational attainment of observations that merge across two different CPS surveys when merging is done solely on the basis of HHID, HHNUM and LINENO. This suggests that taking other factors, such as age and education, into consideration will surely increase the likelihood that merged observations truly represent the same individual rather than two different individuals. But the discussion in the last section also suggests that if there is any sort of measurement error in these other factors, a tradeoff is involved in making use of this additional information. On the one hand, taking other factors into account is likely to lead to the rejection of merges that do not represent the same individual. On the other hand, doing so is also likely to lead to the rejection of merges which do in fact represent the same individual but for whom these other factors were measured with error in either time period t or $t+1$. How significant are these tradeoffs?

In order to answer this question, we must first establish a way to gauge the tradeoff between invalidating incorrect merges vs. invalidating correct ones. One variable

that may illuminate this issue is household residence one year ago. The migration statistics listed in Table 2 come from the CPS, and specifically, from a question that asks whether respondents lived at the same location on March 1 of the previous year. During the period under consideration in this paper, 1980–1998, this question was asked in every year except 1980 and 1985 when the question referred to March 1 five years previous (that is, 1975 for 1980 and 1980 for 1985). Using the answers to this question, we can ascertain how effective various merge criterion are at invalidating the merged observations which report in time $t+1$ that they did not live at the same address one year ago (and who are thus more likely to represent a false merge). We can also ascertain how effective the merge criterion are at validating the merged observations which report in time $t+1$ that they did live at the same address one year ago (and who are thus more likely to represent true merges).

We establish several criteria for establishing the "validity" of a merge based on discrepancies in some or all of the characteristics discussed above: sex, race, age and education. Clearly, one could consider more factors than these, although these four factors alone give rise to literally thousands of different potential merge criteria. Fig. 4 and Table 3 illustrate the tradeoffs between invalidating likely false merges vs. validating likely true merges for a small subsample of these potential criteria. The criteria included in Fig. 4 (and listed in Table 3) are chosen from the potential universe either because they seemed to us to be sensible or because they represented an interesting alternative. It is important to note that there is no single "best" criterion. Each of these criteria involves a tradeoff between invalidating false positives and retaining true positives, although some criteria do appear to be dominated by others. The "best" criterion is ultimately a judgment to be made by the researcher and may vary depending on the research question being asked and/or the methodology being used to answer it.

The letters representing the merge criteria in Fig. 4 and Table 3 stand for the characteristics on the basis of which discrepancies between t and $t+1$ are used to invalidate a merge: sex (S or s), race (R and r), age (A and a), and education (E or e). The symbol "|" stands for "or", so that the point R|A represents merges that are invalidated on the basis of discrepancies in race or age. The absence of an "|" between letters stands for "and", so that the point RA represents merges that are invalidated on the basis of discrepancies in *both* race and education.

Discrepancies in sex and race are easy to define: either these characteristics are the same in t and $t+1$, or they are not. As noted in the previous section, what constitutes a discrepancy in age and education is less straightforward. We consider two different sets of criteria on the basis of age and education, one more restrictive than the other. The first, and less restrictive set of criteria, considers a merge to be valid on the basis of age if the difference in age between t and $t+1$ is no less than -1 and no greater than 3, and on the basis on education if educational attainment is either the same from t to $t+1$, decreases by no more than 1 year (category), or increases by no more than 2 years (categories). The second, and more restrictive set of criteria, considers a merge to be valid on the basis of age if the difference in age between t and $t+1$ is no less

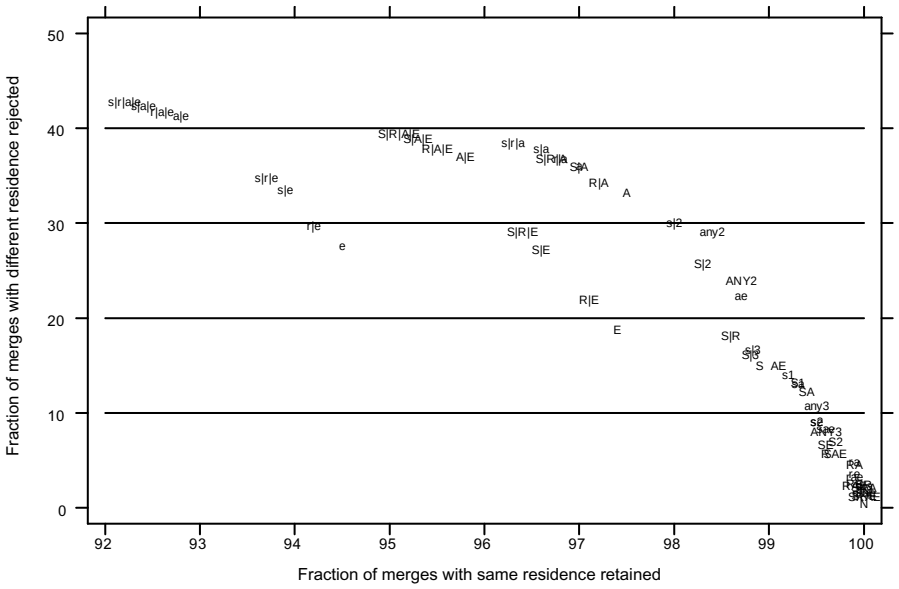

Fig. 4. The trade-offs of various merge criterion.

than 0 and no greater than 2, and on the basis on education if educational attainment is either the same from t to $t+1$ or increases by no more than 1 year (category).

In Fig. 4 and Table 3, capital letters (S, R, A, E) are associated with the merge criteria that are less restrictive with respect to discrepancies in age and education, while lowercase letters (s, r, a, e) are associated with the merge criteria that are more restrictive with respect to discrepancies in age and education. The criteria which depend only on sex and/or race are denoted by capital letters. ANY2 stands for discrepancies in any two of the four factors (any2, ANY3 and any3 are similarly defined). S|2 stands for a discrepancy in sex or two of the remaining 3 factors (s|2, S|3 and s|3 are similarly defined).

Table 3 lists each of the criteria along with the fraction of merged observations which report living in the same residence one year ago that are retained as valid merges, and the fraction of merged observations who report living in a different residence one year ago that are invalidated according to the merge criterion. This table gives the numbers that correspond to each of the points in Fig. 4.

The x-axis of Fig. 4 gives the fraction of merged observations who report living in the same residence one year ago that are retained as valid merges under each of the criteria. In this graph, the fraction of merges living in the same place last year that are retained ranges from a low of about 92% under criterion s|r|a|e to a high of a high on 99.98% under criterion SRAE. The y-axis gives the fraction of merged observation who report living in a different residence one year ago that are rejected as invalid merges under each of the criteria. This fraction ranges from a low of 0.5%

Table 3
The tradeoffs of various merge criteria

Merge Criteria – merges invalidated on the basis of discrepancies in:	Less restrictive age and education criteria		More restrictive age and education criteria	
	% of merges w/ diff. residence rejected	% of merges w/ same residence retained	% of merges w/ diff. residence rejected	% of merges w/ same residence retained
Naïve merge (N)	0.0%	100.0%	0.0%	100.0%
Sex (S)	9.2	99.04	9.2	99.04
Race (R)	3.5	99.65	3.5	99.65
Age (A, a)	20.8	97.65	22.6	97.19
Education (E, e)	12.0	97.43	18.5	94.61
Sex and Race (SR)	1.3	99.96	1.3	99.96
Sex and Age (SA, sa)	7.4	99.50	7.9	99.46
Sex and Education (SE, se)	3.9	99.66	5.5	99.55
Race and Age (RA, ra)	2.6	99.90	2.8	99.90
Race and Education (RE, re)	1.4	99.93	2.1	99.91
Age and Education (AE)	9.2	99.17	13.8	98.85
Sex, Race and Age (SRA, sra)	1.1	99.96	1.2	99.96
Sex, Race and Education (SRE, sre)	0.6	99.97	0.8	99.97
Sex, Age and Education (SAE, sae)	3.3	99.72	4.9	99.62
Race, Age and Education (RAE, rae)	1.2	99.95	1.8	99.93
Sex, Race, Age and Educ. (SRAE, srae)	0.5	99.98	0.8	99.97
Sex or Race (S|R)	11.4	98.74	11.4	98.74
Sex or Age (S|A, s|a)	22.6	97.19	23.8	96.78
Sex or Education (S|E, s|e)	17.4	96.82	22.2	94.10
Race or Age (R|A, r|a)	21.6	97.40	23.2	96.95
Race or Education (R|E, r|e)	14.1	97.16	19.9	94.35
Age or Education (A|E, a|e)	23.6	95.91	27.2	92.94
Sex or Race or Age (S|R|A, s|r|a)	23.2	96.95	24.4	96.54
Sex or Race or Education (S|R|E, s|r|e)	18.7	96.56	23.1	93.86
Sex or Age or Education (S|A|E, s|a|e)	24.9	95.52	28.0	92.60
Race or Age or Education (R|A|E, r|a|e)	24.2	95.68	27.7	92.72
Sex or Race or Age or Educ. (S|R|A|E, s|r|a|e)	25.4	95.28	28.3	92.38
Any three of four factors (ANY3, any3)	4.7	99.67	6.5	99.57
Any two of four factors (ANY2, any2)	14.9	98.85	18.1	98.58
Sex and one of three other factors (S1, s1)	8.0	99.44	8.5	99.39
Sex and two of three other factors (S2, s2)	4.0	99.70	5.4	99.61
Sex or two of three other factors (S|2, s|2)	16.0	98.46	18.8	98.24
Sex or all three other factors (S|3, s|3)	9.9	99.01	10.3	99.00

Note: Authors' calculations based on merging the March 1980–1998 CPSs as described in the text.

under criterion SRAE to a high of 28.3% under criterion s|r|a|e. What more general observations can we make about these various merge criteria from Fig. 4 and Table 3?

First, the two extremes of the various merge criteria noted above, SRAE and s|r|a|e, illustrate the tradeoff between rejecting falsely merged observation and retaining validly merged observations quite starkly: the criterion that does the best at retaining apparently "valid" merges does poorly in rejecting apparent "invalid" merges (SRAE), while the criterion that does the best at rejecting apparently "invalid" merges does so

at the expense of also invalidating a substantial fraction of apparently "valid" merges (s|r|a|e).[12] Ideally, the best place to be in Fig. 4 would be the upper right-hand corner with a high retention rate for "valid" merges and a high rejection rate for "invalid" merges. But, as just noted, those criteria that tend to be higher up on the graph (higher rejection rate for "invalid" merges) also tend to be further to the left rather than to the right (lower retention rate for "valid" merges).

Second, there is a set of criteria which appears to trace out a concave frontier for the tradeoff between rejecting "invalid" and retaining "valid" merges: there are some criteria that invalidate a substantial number of "false" merges without invalidating many "true" merges, but the marginal cost of increasing the fraction of "false" merges that are invalidated in terms of "true" merges that are also invalidated is increasing.

Third, there are some criteria which are clearly dominated by others (and which are inside the concave frontier). For example, S|R|E is clearly dominated by the criteria s|a, S|R|A, r|a, S|A, a, R|A, A, s|2, and any2. Relative to S|R|E, all of these criteria result in a higher fraction of merges with a different residence last year being rejected and a higher (or equivalent, in the case of any2) fraction of merges with the same residence last year being retained.

Fourth, as would be expected, the criteria that are more restrictive with respect to age and education (those designated by lowercase letters) lead to both a higher fraction of merges with a different residence last year being rejected, and a much lower fraction of merges with the same residence being retained.

Note that Fig. 4 only shows the tradeoff between retaining valid merges vs. rejecting invalid merges for a small fraction of the universe of merge criterion that could be adopted: considering additional factors (for example, consistency in marital status between t and $t+1$, or consistency in the relationship to household head between t and $t+1$), and increasing or decreasing the stringency of merge standards with respect to age and education would further "flesh out" the apparent frontier in Fig. 4. We view Fig. 4 as a point of departure for evaluating different merge criteria, not as an illustration of the only merge criteria worth considering.

6. Evaluating potential merge criterion

To further evaluate the validity of the potential criteria on which to merge successive CPS surveys, we will restrict ourselves to eight of the criteria in Table 3 and Fig. 4. They are: S|2 and s|2, ANY2 and any2, S|R|A and s|r|a, and S|R|A|E, and s|r|a|e. We look at the "sex or any two" (S|2 and s|2) and "any two" criterion (ANY2 and any2) because relative to the substantial mass of merge criteria in the lower-right

[12] We use the terms "valid" and "invalid" in reference to the merges quite loosely here. We have no way of judging exactly which merges are valid and which are not. While residence last year is certainly likely to be highly correlated with whether or not a merge is valid, we recognize the potential for measurement error in this variable.

hand corner of Fig. 4, these criterion seem to offer a relatively large improvement in the fraction of merges with a different residence that are rejected (invalidation of "false positives"), without a very large corresponding decrease in the fraction of merges with the same residence that are retained (retention of "true positives"). We look at the "any of four" criterion s|r|a|e because of the criterion in Fig. 4, this one gives the highest rejection rate of merges with a different residence. We also look S|R|A|E and the "any of three" criteria s|r|a and S|R|A because relative to s|r|a|e, these criteria reject almost as many merges with a different residence but retain a much higher fraction of merges with the same residence.

Table 4 presents some additional statistics on which to assess the performance of these various merge criteria. The first row of Table 4 gives the fraction of respondents in time t that are deemed to have a valid match in time $t+1$ under each of these criteria. This merge rate varies from a high of 71.0% when using the naïve criterion (HHID, HHNUM and LINENO alone) in establishing a valid merge, to a low of 65.1% when valid merges must also be of the same gender and race and satisfy the more restrictive age and education criteria. The second row of Table 4 gives the merge rate of row 1 relative to the naïve merge rate.

The third row of Table 4 shows the fraction of valid merges according to each of these criteria that report the same residence in both t and $t+1$. This fraction ranges from a low of 97.3% using the naïve criterion to a high of 98.3%. Note that these fractions are very high, and as such, indicate that the validated merges under each of these criteria are indeed likely to represent the same individuals. Numbers of this level need not even indicate any false merges because of slight discrepancies in the time frame of the mobility question and the March CPS survey. The mobility question refers to residence on March 1 of the previous year, while the CPS surveys themselves are conducted during the week containing the 19th of March, so any individual who moved during the first two or three weeks of March could be a sample respondent in time t and still correctly answer in time $t+1$ that he or she had not lived in the same residence on March 1 of the previous year. Indeed, if mobility were uniformly distributed across all 52 weeks of the year, 4–5% of respondents would be in this position. In fact, mobility is not uniformly distributed across the year, nor across weeks within the month, and the first two weeks of March are likely to have lower than average changes in residence. Nonetheless, the real possibility that some small fraction of individuals could be in the survey in time t and legitimately report in time $t+1$ that they had a different residence on March 1 of the previous year suggests that the small fraction of "valid" merges reporting a different place of residence could still be legitimate.

The sixth row of Table 4 shows the fraction of invalidated merges according to each of these criteria that report the same residence one year ago. These fractions range from a low of 64.1% under criterion ANY2 to a high of 86.8% under criterion s|r|a|e. To rule out the likelihood of incorrectly invalidated merges, these numbers would ideally be small. While certainly much lower than the fraction of valid merges who report the same residence one year ago (row 3), they are much greater than

0. That they are much lower than the fraction of valid merges who report the same residence one year ago certainly suggests that the restrictions on what constitutes a valid merge do in fact distinguish between valid and invalid merges. But that they are still quite high, and that they get higher as the merge rate falls, suggests that the cost of weeding out the invalid merges is simultaneously weeding out a substantial number of valid merges as well.[13]

To further gauge the tradeoff between invalidating incorrect merges vs. invalidating correct ones, we also look at changes in marital status between t and $t+1$ for "valid" and "invalid" merges, and changes in the relationship to household head. The first thing to note about marital status is that it can legitimately change – marriage, divorce and death are not infrequent events.[14] So, even among valid merges, there should be some fraction of individuals who report a change in marital status. Similarly, relationship to household head could also change. The fourth row of Table 4 shows that the fraction of valid merges reporting the same marital status in both t and $t+1$ ranges from a low of 97.0% using the naïve criterion to a high of 97.9%; the fraction with the same relationship to household head ranges from 97.9% to 98.4%. And, although not shown, the fraction of valid merges reporting the same marital status follows the expected pattern with respect to age, with a higher fraction of "valid" merges showing a change in marital status at young ages when people are most likely to marry, and at very old ages when spouses are most likely to die. As with the fraction of valid merges reporting the same residence last year, the fraction reporting the same marital status or relationship to household head in both surveys increases as the merge rate (row 1) declines. These statistics, in conjunction with those on mobility, lead us to believe that all of the selected criteria do a good job of rejecting invalid merges: on the basis of mobility, marital status, and relationship to household head, it does not appear that any of these criteria falsely retain a substantial number of invalid merges.

The seventh row of Table 4 shows that the fraction of invalidated merges reporting the same marital status (row 7) in both years ranges from a low of 63.7% under criterion ANY2 to a high of 87.6% under criterion s|r|a|e. In general, these fractions are similar to those for the same residence last year (row 6), although one might expect a fair amount of similarity in marital status even for merges that do not represent the same individuals. The fraction of invalidated merges reporting the same relationship to household head (row 8) is somewhat higher, ranging from a low of 82.7% under criterion ANY2 to a high of 93.1% under criterion s|r|a|e.

The ninth row of Table 4 gives the fraction of respondents in the $t+1$ survey that are not merged according to each of the criteria. As expected, the criteria which result in the lowest fraction of respondents in time t who are merged also result in

[13] Interestingly, there is a substantial decline between 1993 and 1996 in the fraction of "invalid" merges who report the same residence one year ago, perhaps due to the 1994 redesign of the CPS.

[14] For the sake of simplicity, we have classified individuals as being either married (spouse present or spouse absent) or single (widowed, divorced or never married).

Table 5
Factors contributing to the invalidation of merges under different merge criterion

Merge criteria	Sex	Race	Age	Education
Less restrictive age and education criteria				
ANY2	50.6%	13.6%	93.8%	74.2%
S|2	60.6	10.8	74.8	59.2
S|R|A	33.3	12.3	79.8	–
S|R|A|E	23.2	8.6	55.6	52.9
More restrictive age and education criteria				
any2	44.4%	11.9%	94.3%	84.7%
s|2	52.6	10.2	80.4	72.2
s|r|a	29.9	11.1	83.5	–
s|r|a|e	15.3	5.7	42.8	69.8

Note: Authors' calculations based on merging the March 1980–1998 CPSs as described in the text.

the highest fraction of respondents in time $t+1$ who are not merged. Moreover, the various merge criteria are fairly symmetric: the merge rates for time t respondents (row 1) are almost identical to the merge rates for time $t+1$ respondents for each of the merge criteria.

The last row of Table 4 gives the fraction of non-merged $t+1$ respondents who report living in the same residence one year ago. This ranges from a low of 49.5% under the naïve merge criterion to a high of 56.0% under criterion s|r|a|e. The fraction of non-merged $t+1$ respondents reporting the same residence is below the fraction of $t+1$ respondents in invalidated merges reporting the same residence one year ago (row 6), but is still fairly high: about half of the $t+1$ respondents who don't merge at all on the basis of the naïve merge criterion report living in the same residence one year previously. This reflects the combined effect of two likely factors. First, CPS-eligible non-response households in time t who are surveyed in $t+1$ and who have not moved in the past year will report living in the same residence one year ago despite not being matched (or matchable) to individuals in the time t survey. Second, coding error in HHID, HHNUM or LINENO for individuals who are in fact in both surveys and who presumably could be merged will result in non-merged $t+1$ individuals who report living in the same residence one year ago. This type of coding error will give rise to "false negatives"– individuals who do not merge but who are, in fact, included in both surveys.

Table 5 shows the general role of gender, race, age and education in invalidating merges under each of the selected merge criterion. Note that these factors can sum to more than 100% in each row because more than one factor can contribute to the invalidation of a merge. Discrepancies in race are the least common reason for the invalidation of merges. This is not surprising – the strong patterns of residential segregation by race are likely to lead to a substantial amount of correlation in the race of housing unit occupants, even when residential mobility does occur. Differences in gender are somewhat more important than differences in race, but are still not a factor in almost two-thirds (or more) of the invalidated merges under the S|R|A,

Table 6
Demographic characteristics and merge criteria

Characteristic	Merge Status			
	Naïvely merged individuals	Non-merged individuals	"Valid" s|r|a|e merged individuals	"Invalid" s|r|a|e merged individuals
Age (years)	44.4	36.3	44.5	42.2
Gender				
Male (%)	46.7	48.7	46.7	46.4
Female (%)	53.3	51.3	53.3	53.6
Race				
White (%)	87.5	86.0	87.8	84.0
Black (%)	9.1	9.8	9.0	10.9
Other (%)	3.4	4.2	3.3	5.1
Married (%)	61.9	47.5	62.6	54.5
Education				
$<$ High school (%)	28.0	30.6	27.0	38.6
High school (%)	36.0	34.5	36.9	26.2
Some college (%)	18.5	19.7	18.4	19.0
College (%)	10.7	10.0	10.9	9.4
$>$ College (%)	6.8	5.3	6.8	6.8
Moved in past year (%)	11.4	28.1	11.1	14.4
Real family income	$25,064	$19,043	$25,341	$21,787
Real own income	$5,084	$5,207	$5,185	$3,750

Note: Authors' calculations based on merging the March 1980–1998 CPSs as described in the text.

s|r|a, S|R|A|E, and s|r|a|e criterion. Rather, differences in age and education are the biggest factors in the rejection of merged observations for these criteria.

As noted earlier, while we have focused on gender, race, age and education as criteria to use in validating merged observations, other criteria could be used in addition to or instead of some of these factors. However, whatever the criteria used, without the ability to identify when discrepancies in these criterion are correctly applied to invalidate incorrect merges vs. incorrectly applied to invalidate correct merges, there is a tradeoff between ensuring the integrity of the matched sample versus generating potential sample selection biases by excluding a greater number of observations from the sample of matches. Table 6 gives the demographic characteristics for time t respondents who: are naively merged (column 1); are not naively merged (column 2); are validly merged according to the criterion s|r|a|e (column 3); and are invalidated according to criteria s|r|a|e (column 4). This table clearly illustrates that there are some discrepancies in the characteristics of individuals on the basis of their merge status, although the differences between naively merged individuals whose merges are invalidated or not is smaller than the differences between individuals that do and do not merge at all. The impact that this kind of selection bias might have on important parameter estimates within the context of statistical models is not known, and quite likely will vary depending on the circumstances at hand. But being aware of the potential severity and direction of the selection bias is likely to be important.

7. Conclusions

This paper has, hopefully, illustrated several things. The first is that despite the lack of individual identifiers in the CPS that are guaranteed to be intertemporally consistent, merging individuals across two different CPS surveys is far from impossible. The second thing to be learned from this paper is that measurement error in both demographic and household identification variables means that it will be very difficult to find an algorithm to correctly merge all individuals who should merge without also merging individuals who shouldn't. This results in a fundamental tradeoff: ensuring the integrity of the matched sample versus generating sample selection bias by excluding potentially valid matches from the sample of "valid" merges.

Based on the analysis presented here, it appears that a naïve merge on the basis of HHID, HHNUM and LINEO alone is likely to result in "merged" observations that do not represent a match between the same individuals. Imposing additional merge criteria on gender, race, age and education (and potentially other factors as well) will invalidate many of these incorrect merges. In general, the criteria used to invalidate merges appear to have a tradeoff: as measured by reported residence one year ago, increasing the fraction of invalid merges that are rejected comes at a cost of decreasing the fraction of valid merges that are retained. However, there are clearly some merge criteria which are superior to others in the sense that they result in both a higher fraction of invalid merges being rejected and a higher fraction of valid merges being retained. Each researcher must ultimately make his or her own decisions regarding which merge criterion to adopt, and the appropriate criterion my very well depend on the application at hand.

While we have focused on gender, race, age and education as criteria to use in validating merged observations, other criteria could be used in addition to or instead of some of these factors (e.g. marital status, relationship to household head, etc.). [15] Given the factors that we have looked at, it appears to us that criterion S|R|A does a good job of balancing the need to invalidate incorrect merges without needlessly invalidating too many correct merges. The general algorithm that we would propose for merging the CPS after 1980 is to merge first by HHID, HHNUM and LINENO, and then impose merge the criterion S|R|A by excluding merges for which any of the following is true: i) gender differs; ii) race differs; or iii) the difference in age between t and $t+1$ is less than -1 or greater than 3.

[15] The results in this paper also suggest that residence last year could be used as a merge criterion. However, residence last year is only asked in March, and thus could only be used as a merge criterion for March-to-March merges, and even then not for all March-to-March merges because in some years (1985, for example), the question refers to residence five years ago rather than residence one year ago. Thus, the usefulness of this variable in merging CPS surveys is somewhat more limited that the traditional demographic variables used to merge.

Appendix A: comparing educational attainment in two different years

This appendix outlines briefly how we define differences in educational attainment across two different years. We break this down by CPS "regime" according to the types of educational attainment questions that were asked.

1980–1991: As noted in the text, prior to 1992, two questions were asked about educational attainment in the CPS, the first about the number of years of school attended, and the second about whether the last year of school was completed. The resulting education variables are a linear "years of school" variable (GRDHI) ranging from 1–19 from 1980–1987 and from 0–18 from 1988–1991, and an indicator for whether the final year was completed (GRDCOM). We make the years of school variable consistent from 1980–1991 by subtracting 1 for the period from 1980–1987. We also subtract 1 if the final year of school attended was not completed (we do this for all years of schooling except 0 which corresponds to kindergarten). The resulting variable measures years of school completed, and differences in this variable from t to $t+1$ measure differences in years of school completed.

1992–1998: Beginning in 1992, education was no longer classified according to years of school, but according to the highest level of educational attainment classified into categories. Tables A1 and A2 list these various categories and their values as assigned by the CPS. We have classified differences in educational category between t and $t+1$ according to how plausible the movement

Table A2
Merge criterion and differences in educational category

Educational	Educational category, time $t+1$															
category, time t	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46
31: $<$ Grade 1	MR	MR	x	x	x	x	x	x	x	x	x	x	x	x	x	x
32: Grades 1−4	LR	MR	MR	x	x	x	x	x	x	x	x	x	x	x	x	x
33: Grades 5−6	x	LR	MR	MR	x	x	x	x	x	x	x	x	x	x	x	x
34: Grades 7−8	x	x	LR	MR	MR	LR	x	x	x	x	x	x	x	x	x	x
35: Grade 9	x	x	x	LR	MR	MR	LR	x	x	x	x	x	x	x	x	x
36: Grade 10	x	x	x	x	LR	MR	MR	LR	x	x	x	x	x	x	x	x
37: Grade 11	x	x	x	x	x	LR	MR	MR	MR	LR	x	x	x	x	x	x
38: Grade 12, no diploma	x	x	x	x	x	x	LR	MR	MR	MR	MR	MR	X	x	x	x
39: HS diploma/GED	x	x	x	x	x	x	LR	LR	MR	MR	MR	MR	X	x	x	x
40: Some college, no degree	x	x	x	x	x	x	x	LR	LR	MR	MR	MR	MR	x	x	x
41: Assoc. degree (vocational)	x	x	x	x	x	x	x	LR	LR	LR	MR	MR	MR	x	x	x
42: Assoc. degree (academic)	x	x	x	x	x	x	x	LR	LR	LR	MR	MR	MR	x	x	x
43: Bachelor's degree	x	x	x	x	x	x	x	x	x	LR	LR	LR	MR	MR	MR	MR
44: Master's degree	x	x	x	x	x	x	x	x	x	x	x	x	LR	MR	MR	MR
45: Professional degree	x	x	x	x	x	x	x	x	x	x	x	x	LR	MR	MR	MR
46: Doctorate degree	x	x	x	x	x	x	x	x	x	x	x	x	LR	MR	MR	MR

Note: MR denotes cells which satisfy the more restrictive educational merge criterion that are described in the text; LR denotes cells which also satisfy the less restrictive merge criterion that are described in the text.

Table A3
Merge criterion and education in 1991 and 1992

Educational category in 1992	Years of schooling in 1991	
	More restrictive merge criteria	Less restrictive merge criteria
31: $<$ Grade 1	0	0−1
32: Grades 1−4	0−3	0−4
33: Grades 5−6	4−6	3−7
34: Grades 7−8	6−8	5−9
35: Grade 9	8−9	7−10
36: Grade 10	9−10	8−11
37: Grade 11	10−11	9−12
38: Grade 12, no diploma	11−12	10−13
39: HS diploma/GED	11−12	10−13
40: Some college, no degree	11−15	11−16
41: Assoc. degree (vocational)	11−15	11−16
42: Assoc. degree (academic)	11−15	11−16
43: Bachelor's degree	15−16	14−18
44: Master's degree	16−18	16−18
45: Professional degree	16−18	16−18
46: Doctorate degree	17−18	17−18

from one category to another would be and also according to how movements across categories correspond to changes in the years of school completed for the 1980–1991 period. Tables A1 and A2 show a matrix of educational attainment in time t by educational attainment in time $t+1$. In Table A1, the contents of each cell gives our appraisal of the relative difference in the two categories, with 0 meaning that the categories are either the same or roughly equivalent, 1 meaning that it would not be implausible to see an individual move from one category to the next over the course of one year, and so on. Table A2 shows how we categorize educational differences in defining the various merge criteria.

1991–1992: Table A3 shows the combinations of years of schooling in 1991 and educational category in 1992 which satisfy the less restrictive educational merge criteria and the more restrictive educational merge criteria.

Appendix B: Details on the procedure used to longitudinally match CPS respondents

This appendix outlines in greater detail the specific steps used in longitudinally matching the CPS in this paper. All of the statistical programs used to match the CPS in this paper are available from the authors upon request.

1) Make two data extracts, one for time t and one for time $t+1$, both of which contain the variables necessary to merge and any additional variables to be used in a statistical analysis. For the analysis in this paper, this was done on a PC using the CPS Utilities data extraction program. The variables that we extracted

Table B1
CPS variables used in the analysis

Description	Unicon Name	Census Bureau Name	CPS surveys[b]
Month-in-sample	MIS	MIS	76−98
Household identifier	HHID	HH-IDENT-NUM	77−88
		I-IDNUM	88B−95
		H-IDNUM	96−98
Household number	HHNUM	ITEM9	77−88
		H-HHNUM	88B−98
Line number	LINENO	LINENO	79−88
		A-LINENO	88B−98
State of residence	STATE	MST-STATE	77−88
		HG-ST60	88B−98
Sex	SEX	SEX	63−88
		A-SEX	88B−98
Race	RACE	RACE	63−88
		A-RACE	88B−98
Age	AGE	AGE	76−88
		A-AGE	88B−98
Highest grade attended	GRDHI	HI-GRADE	76−88
		A-HGA	88B−91
Whether highest grade was completed	GRDCOM	GRADE-COMPL	76−88
		A-HGC	88B−91
Highest level of education	GRDATN	A-HGA	92−98
Moved in the past year	MIGSAM[a]	MIGSAME	80−98
	MIGSAM1[a]		
Marital status	MARSTAT	MAR-STAT	76−88
		A-MARITL	88B−98
Relationship to household head/reference person	RELHD	REL-HEAD	76−88
		A-EXPRRP	88B−98
Total family income	FAMINC	F-INC-TOT	76−88
		FTOTVAL	88B−98
Total individual earnings	INCERN	PINCERN	80−88
		PEARNVAL	88B−98

Source: CPS Utilities Electronic Documentation [15].

a. The Unicon variable MIGSAM refers to residence one year previous for all surveys except 1985 and 1995 when it refers to residence five years previous. The Unicon variable MIGSAM1 refers to residence one year previous in the 1995 survey. There is no corresponding variable for residence one year previous in the 1985 survey for either the Census CPS or the Unicon repackaging.

b. A major redesign of the CPS in 1988 resulted in the release of two different versions of the 1988 CPS, the regular release and the B release.

for this paper are listed in Table B1. For a March-to-March merge, respondents with a MIS of 1–4 should be included in the time t extract, while respondents with a MIS of 5–8 should be included in the time $t+1$ extract. Respondents with a MIS of 5–8 in time t or with a MIS of 1–4 in time $t+1$ can be excluded since these respondents are not included in the sampling frame of both surveys. For a month-to-month merge (e.g. March-to-April), respondents with a MIS of 1–3 or 5–7 should be included in the time t extract, while respondents with a MIS of 2–4 or 6–8 should be included in the time $t+1$ extract.

2) Recode MIS in the time $t+1$ data to correspond to the appropriate value that respondents in time t would have if they were in both surveys. For March-to-March merges, subtract 4 from the $t+1$ MIS. For a month-to-month merge, subtract 1 from the $t+1$ MIS. Other variables that will be used to determine the validity of matches (e.g. sex, race, age, etc.) will need to be given different names in the two extracts so that both the time t and time $t+1$ values are preserved.
3) Sort the time t and $t+1$ data by MIS, HHID, HHNUM and LINENO. For a March 1994-to-March 1995 merge, the data must be sorted by MIS, STATE, HHID, HHNUM, and LINENO. This would be true for some of the month-to-month merges in the 1994–1995 time period as well and results from the fact that the CPS only assigns unique household identifiers (HHID) within state over part of this time period.
4) Match merge the sorted t and $t+1$ data extracts on the basis of the variables used to sort the data above.

One problem that may arise (depending on which CPSs are being merged), is the presence of multiple post-merge observations with the same identifying variables (HHID, HHNUM, LINENO). This occurs because even though HHID, HHNUM and LINENO are meant to uniquely identify individuals, in some CPS surveys there are multiple respondents who have the same HHID, HHNUM and LINENO. If, for example, there are two individuals with the same HHID, HHNUM and LINEO in both of the CPS surveys being matched, we will end up with four merged observations. Two of the merged observations will be (potentially) correct, and two of them will be incorrect. We deal with this issue in a way designed to preserve as many potentially correct matches as possible.

First, we create a unique identifier for all respondents in both t and $t+1$ (these identifiers are not unique across t and $t+1$, only within t and $t+1$ – we do not merge on the basis of these identifiers). After merging the t and $t+1$ data extracts as described above, we flag the post-merge observations that do not have a unique value of the t and/or $t+1$ identifiers that we create. Among these flagged observations, we then deleted those that do not have the same sex in t and $t+1$. We then flag the remaining post-merge observations that still did not have a unique value of the t and/or $t+1$ identifiers that we created. Among these flagged observations, we then delete those that do not have the same race in t and $t+1$. We repeat this process, deleting those observations with different values of age according to the less-restrictive age criteria and then according to the more-restrictive age criteria, different values of education according to the less-restrictive education criteria and then according to the more-restrictive education criteria, and finally, we delete those flagged observations with differences in their relationship to household head in time t and $t+1$. We then go through and flag any remaining observations with non-unique identifiers and delete all of these observations.

Table B2 shows the number of time t and time $t+1$ respondents that are merged to more than one individual because they have a non-unique individual identifier in either

time t, time $t+1$, or both periods, for each of the 1980–1998 March-to-March merges that are possible. It also notes how many observations with non-unique identifiers remain after we apply each of the deletion criteria just described. Note that for many of the March-to-March merges, non-unique identifiers are not a problem. For the March-to-March merges in which there are individuals with non-unique identifiers, these individuals constitute only a small fraction of the total sample.

At this point, we then calculate the fraction of t and $t+1$ respondents that are successfully merged. We then apply the criteria discussed in the paper to flag those merged observations that do not appear to represent the same individuals. For the purposes of statistical analysis on matched CPS survey respondents, the next and final step would be to delete from the sample the observations from time t and $t+1$ that do not merge along with the merged observations that are rejected according to whatever criterion is adopted to make this assessment.

Acknowledgements

We thank Eanswythe Grabowski and Erzo Luttmer for comments on a previous draft of this paper. Research support from the National Institutes on Aging is gratefully acknowledged.

References

[1] Bureau of Labor Statistics (1996), Short History of the CPS, http://www.bls.census.gov/cps/bhistory.htm., Accessed March 1, 1999.
[2] Bureau of Labor Statistics (1997), Handbook of Methods, http://stats.bls.gov/opub/hom/pdf/homch1.pdf., Accessed March 1, 1999.
[3] D.M. Cutler and B.C. Madrian, Labor Market Responses to Rising Health Insurance Costs: Evidence on Hours Worked, *RAND Journal of Economics* **29** (1998), 509–530.
[4] D.S. Evans and L.S. Leighton, Some Empirical Aspects of Entrepreneurship, *American Economic Review* **79** (1989), 519–535.
[5] A. Katz, K. Teuter and P. Sidel, Comparison of Alternative Ways of Deriving Panel Data from the Annual Demographic Files of the Current Population Survey, *Review of Public Data Use* **12** (1984), 35–44.
[6] S.A. Levitan and F. Gallo, Workforce Statistics: Do We Know What We Think We Know – And What Should We Know? *Journal of Economic and Social Measurement* **16** (1990), 87–124.
[7] E.F.P. Luttmer, Does the Minimum Wage Cause Inefficient Rationing? Unpublished paper, University of Chicago, 1998.
[8] National Center for Health Statistics, GMWK291 – Death Rates for 72 Selected Causes by 5-Year Age Groups, Race, and Sex: United States, 1979–1996, http://www.cdc.gov/nchswww/data/gm291_1.pdf (ages 0–39) and http://www.cdc.gov/nchswww/data/gm291_ 2.pdf (ages 44+). Accessed February 27, 1999.
[9] D. Neumark and W. Wascher, The Effects of Minimum Wages on Teenage Employment and Enrollment: Evidence from Matched CPS Surveys, in: *Research in Labor Economics*, Vol. 15, S.W. Polachek, ed., JAI Press, Connecticut, 1996, pp. 25–63.
[10] F. Peracchi and F. Welch, Trends in Labor Force Transitions of Older Men and Women, *Journal of Labor Economics* **12** (1994), 210–242.

[11] A. Pitts, Matching Adjacent Years of the Current Population Survey, Unpublished paper, Unicon Research Corporation, California, 1988.
[12] L.M. Segal and D.G. Sullivan, The Growth of Temporary Services Work, *Journal of Economic Perspectives* **11** (1997), 117–136.
[13] StataCorp, *Stata Statistical Software: Release 5.0*, Stata Corporation, College Station, TX, 1997.
[14] Unicon Research Corporation, Appendix S – Discussion Regarding: Matching of CPS Files, from the CPS Utilities Annual Demographic and Income Supplement: March 1976–1998 (Disc B) CD-ROM (Manual/section5), Unicon Research Corporation, California, 1999.
[15] Unicon Research Corporation, CPS Utilities Annual Demographic and Income Supplement: March 1964–1998 CD-ROM, Unicon Research Corporation, California, 1999.
[16] F. Welch, Matching the Current Population Surveys, *Stata Technical Bulletin* **12** (1993), 7–11.

Table B2
Details on observations with non-Unique Individual Identifiers When Longitudinally Merging the CPS

March-to-March merge	Number of post-merge observations with non-unique Ids	Observations with non-unique identifiers remaining after deletion on the basis of differences in:				
		Sex	Race	Age	Education	Household relationship
1980–1981						
1980 respondents	210	40	40	2	2	2
1981 respondents	261	64	64	8	6	4
1981–1982						
1981 respondents	208	62	60	6	6	2
1982 respondents	228	34	34	12	10	6
1982–1983						
1982 respondents	298	86	86	10	10	6
1983 respondents	234	62	60	6	6	2
1983–1984						
1983 respondents	266	76	76	16	14	6
1984 respondents	343	104	100	22	116	8
1984–1985						
1984 respondents	280	62	60	10	8	2
1985 respondents	285	66	64	18	18	6
1986–1987						
1986 respondents	264	86	86	16	14	10
1987 respondents	357	102	92	8	8	4
1987–1988						
1987 respondents	0	0	0	0	0	0
1988 respondents	318	92	82	14	8	0
1988–1989						
1988 respondents	0	0	0	0	0	0
1989 respondents	0	0	0	0	0	0
1989–1990						
1989 respondents	0	0	0	0	0	0
1990 respondents	0	0	0	0	0	0
1990–1991						
1990 respondents	0	0	0	0	0	0
1991 respondents	0	0	0	0	0	0
1991–1992						
1991 respondents	0	0	0	0	0	0
1992 respondents	0	0	0	0	0	0
1992–1993						
1992 respondents	0	0	0	0	0	0
1993 respondents	0	0	0	0	0	0
1993–1994						
1993 respondents	24	10	10	0	0	0
1994 respondents	0	0	0	0	0	0
1994–1995						
1994 respondents	0	0	0	0	0	0
1995 respondents	243	149	132	6	6	6
1996–1997						
1996 respondents	0	0	0	0	0	0
1997 respondents	2	0	0	0	0	0
1997–1998						
1997 respondents	0	0	0	0	0	0
1998 respondents	0	0	0	0	0	0

Note: Authors' calculations based on merging the March 1980–1998 CPSs as described in the text.

Table A1
Merge criterion and differences in educational category (1992+)

Educational category, time t	Educational category, time $t+1$															
	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46
31: < Grade 1	0	1	2	3+	3+	3+	3+	3+	3+	3+	3+	3+	3+	3+	3+	3+
32: Grades 1−4	−1	0	1	2	3+	3+	3+	3+	3+	3+	3+	3+	3+	3+	3+	3+
33: Grades 5−6	−2	−1	0	1	2	3+	3+	3+	3+	3+	3+	3+	3+	3+	3+	3+
34: Grades 7−8	−3+	−2	−1	0	1	2	3+	3+	3+	3+	3+	3+	3+	3+	3+	3+
35: Grade 9	−3+	−3+	−2	−1	0	1	2	3+	3+	3+	3+	3+	3+	3+	3+	3+
36: Grade 10	−3+	−3+	−3+	−2	−1	0	1	2	2	3+	3+	3+	3+	3+	3+	3+
37: Grade 11	−3+	−3+	−3+	−3+	−2	−1	0	1	1	2	3+	3+	3+	3+	3+	3+
38: Grade 12, no diploma	−3+	−3+	−3+	−3+	−3+	−2	−1	0	1	1	2	2	3+	3+	3+	3+
39: HS diploma/GED	−3+	−3+	−3+	−3+	−3+	−2	−1	−1	0	1	2	2	3+	3+	3+	3+
40: Some college, no degree	−3+	−3+	−3+	−3+	−3+	−3+	−2	−1	−1	0	1	1	1	3+	3+	3+
41: Assoc. degree (vocational)	−3+	−3+	−3+	−3+	−3+	−3+	−3+	−2	−2	−1	0	0	1	2	2	3+
42: Assoc. degree (academic)	−3+	−3+	−3+	−3+	−3+	−3+	−3+	−2	−2	−1	0	0	1	2	2	3+
43: Bachelor's degree	−3+	−3+	−3+	−3+	−3+	−3+	−3+	−3+	−3+	−1	−1	−1	0	1	1	2
44: Master's degree	−3+	−3+	−3+	−3+	−3+	−3+	−3+	−3+	−3+	−3+	−2	−2	−1	0	1	1
45: Professional degree	−3+	−3+	−3+	−3+	−3+	−3+	−3+	−3+	−3+	−3+	−2	−2	−1	−1	0	1
46: Doctorate degree	−3+	−3+	−3+	−3+	−3+	−3+	−3+	−3+	−3+	−3+	−3+	−3+	−2	−1	−1	0

Note: This table reflects the authors' assessment of the differences in the years of schooling associated with the educational categories in the CPS in 1992 and later years.

Table 4
Evaluating the different criterion for establishing a "valid" merge

	Merge criteria								
	Naïve	Less restrictive age and education criteria				More restrictive age and education criteria			
		ANY2	S|2	S|R|A	S|R|A|E	Any2	s|2	s|r|a	s|r|a|e
Merge rate									
As a fraction of time t respondents	71.0%	69.9%	69.6%	68.3%	67.1%	69.6%	69.3%	68.0%	65.1%
As a fraction of the naïve merge rate	100.0	98.3	97.9	96.2	94.5	97.9	97.6	95.7	91.6
Fraction of "valid" merges									
With the same residence last year	97.3	97.9	98.0	98.2	98.2	98.1	98.1	98.3	98.3
With the same marital status last year	97.0	97.7	97.7	97.9	97.9	97.7	97.8	97.9	97.9
With the same HH relationship last year	97.9	98.2	98.2	98.3	98.3	98.2	98.3	98.3	98.4
Fraction of "invalid" merges									
With the same residence last year	–	64.1	69.4	75.5	81.5	64.5	68.6	76.9	86.8
With the same marital status last year	–	62.1	66.4	77.2	82.9	65.2	68.3	78.8	87.8
With the same HH relationship last year	–	82.7	84.0	88.3	90.9	83.7	84.8	89.0	93.1
Fraction of non-merged $t+1$ respondents									
$t+1$ respondents not merged	28.8	30.0	30.3	31.6	32.7	30.3	30.6	31.9	34.8
With the same residence last year	49.5	50.1	50.5	51.8	53.4	50.2	50.6	52.2	56.0

Note: Authors' calculations based on merging the March 1980–1998 CPSs as described in the text.