

A Heap of Trouble?

Accounting for Mismatch Bias in Retrospectively Reported Data

(with application to smoking cessation and (non) employment)

by Dean R. Lillard,^{1,2} Haim Bar,¹ and Hua Wang¹
¹Cornell University and ²DIW
May 2008

Corresponding author
Dean R. Lillard
Department of Policy Analysis and Management
432B MVR Hall
Cornell University
Ithaca, NY 14853-4401
Tel. (607) 255-9290
FAX (607) 255-4071
E-mail: DRL3@cornell.edu

This research was supported by Award # R01 HD048828 from the National Institutes of Health. We thank Donald Kenkel, George Jakubson, and Alan Mathios for comments, Robert Strawderman for his statistical advice, and Eamon Molloy for his programming assistance. All errors are our own.

Abstract

Retrospective data are collected for a wide variety of topics. Among others, these topics include marital events (collected through marital histories), births (from fertility histories or constructed from data on age or birth dates), deaths, purchase behavior, and getting, changing, or losing jobs. When people report the timing of events that happened in the distant past, they tend to round up or down the year or time since the event occurred. Consequently, events tend to be “heaped” on multiples of chronological or calendar units (e.g. on units of five or ten for data that naturally occur over years). Such reporting behavior leads to a type of attenuation bias because it causes researchers to match time-varying right-hand side variables to behavior in a year other than the one during which it actually occurred. We investigate heaping in the context of retrospective data on smoking behavior and non employment. We develop and investigate three algorithms researchers can use to mitigate the problems associated with heaping. We show for both outcomes that the methods reduce mismatch bias, sometimes dramatically. The methods offer broad benefits not only because they will produce better estimates but also because there are a growing number of longitudinal studies worldwide that already have and potentially could collect important data through retrospective reports.

I. Introduction

When researchers design and administer surveys they face the challenge of how to collect accurate information about events that happened in the distant past. To do so, researchers can either append information previously collected (e.g. from administrative records) or they can ask respondents to retrospectively report on past events. Retrospective data enhance cross-sectional and longitudinal surveys. With retrospective data one can merge time-varying policy data to time varying event data, even data collected in a single cross-section. Longitudinal surveys are enriched not only when retrospective data are collected in the baseline survey but also when subsequent waves of a panel study collect retrospective data about a social, economic, or physical phenomenon whose importance was unrecognized when the survey began. Because one can link retrospective data to longitudinal survey data collected in previous waves, the addition of retrospective data to longitudinal surveys offers additional benefits over and above retrospective data collected in cross-sectional studies.

Retrospective data are collected for a wide variety of topics. Among others, these topics include marital events (collected through marital histories), births (from fertility histories or constructed from data on age or birth dates), deaths, purchase behavior, getting, changing, or losing jobs.¹

Despite their advantages, retrospectively reported data also have the drawback that people may not accurately report when events occur. It is a stylized fact that, when people report

¹A Google Scholar (<http://scholar.google.com/>) search yielded a large number of “hits” for phrases that likely select studies of events (or that use these events as a covariate). The search yielded: 10,200 hits for “age at marriage,” 3,090 for “time of marriage,” 7,940 for “age at birth,” 1,201 for “age divorce,” “age at divorce,” “time of divorce,” and “time since divorce,” 3,190 for “first sex,” 4,530 for “first job” and duration, and 4,130 for “years worked.”

the timing of events that happened in the distant past, they tend to round up or down the year or time since the event occurred. Consequently, events tend to be “heaped” on multiples of chronological or calendar units (e.g. on units of five or ten for data that naturally occur over years).

Several studies investigate heaping in the context of a more general examination of the quality of retrospective reports of life events in large scale surveys. They identify a number of factors associated with the ability of respondents to recall accurately. The factors include recall the time that has elapsed since the event in question, the salience or social acceptability of an event, respondent characteristics, and interview characteristics.

Recall duration or time since event is a strong predictor of the quality of retrospective reports on marital history in the US Panel Study of Income Dynamics (PSID) (Peters 1988), age at first sex in the National Longitudinal Survey of Youth 1979 (NLSY79) (Wu et al. 2001), and post-partum amenorrhea (the interval after a pregnancy before menstruation returns) in the Malaysian Family Life Surveys (MFLS) (Beckett et al. 2001). Researchers also agree that respondents more accurately report when an event occurred if the event is more salient to the respondent. Kenkel, Lillard, and Mathios (2004) find that smokers are more likely to report the same starting age across different waves of the NLSY79 if they are or were heavier smokers. Although both are salient life events, Peters (1998) shows evidence that dates of divorce are reported less consistently than dates of marriage and conjectures that the difference may arise because divorce is less socially acceptable. Other factors that researchers have linked with recall accuracy include demographic characteristics such as education and race/ethnicity (Kenkel et al, 2004; Peters 1988), question wording (Peters 1998), and even arithmetic facility (Wu et al. 2001).

As we demonstrate below, the presence of heaping can lead to attenuation bias in coefficients of interest. The extent of the attenuation bias varies directly with the nature and severity of heaping. For any given event, when a person “heaps” the true timing of the event is misstated. This heaping causes analysts to match policy data incorrectly. We study heaping in retrospective data of two types - on the age ex-smokers quit their habit and on the month (and year) a person started or stopped working at a job. In both cases one models decisions to start or stop smoking or working as a function of time varying prices a person faces in a particular year (month). The mismatch between the true and reported (heaped) start/stop date is a fundamental source of attenuation bias. In the rest of the paper we label the attenuation bias that results as *mismatch bias*. We distinguish mismatch bias from the attenuation bias that results from classical measurement error because the mismatch error arises even when the underlying explanatory variable is measured perfectly.^{2,3}

Although we investigate heaping in the context of retrospective data on smoking and employment behavior, our goal is to develop a set of general algorithms researchers can use to mitigate the problems associated with heaping in all similar data (which are ubiquitous). While others reject retrospective data on the grounds that there is too much recall error (Tauras and Chaloupka 2001), we believe that it is possible to develop methods to reduce the bias that will yield broad benefits in several ways. They allow researchers to more fully exploit existing

²Hirsch and Schumacher (2004) use the term “match bias” to refer to bias that arises in data with imputed values for some observations. They document the bias that occurs when individuals who are matched by “nearest neighbor” imputation algorithms differ on key characteristics not used to match.

³Bound, Brown, and Mathiowetz (2001) comprehensively review the classical measurement error literature.

retrospective data. They will reduce bias in coefficients of scientific and policy interest. Finally, these methods raise the scientific value of the growing number of longitudinal studies worldwide that already have and potentially could collect important data through retrospective reports.⁴

Below we describe three algorithms one can use to mitigate the bias that heaping introduces to coefficient estimates of interest. Our analysis follows and builds on the work of Little (1992), Torelli and Trivellato (1993), Heitjan and Rubin (1990). Each of these studies recognized the potential problem that heaping might cause. Little (1992) provides a succinct review of event history analysis and missing-data methods. Torelli and Trivellato (1993) propose solutions to heaping in data on unemployment spells of Italian youth that involve the specification of a parametric model of the errors in the reformulated likelihood function, adding dummy variable to flag ex-smokers who heap or do not heap, and smoothing the data as recommended by Heitjan and Rubin (1990). Heitjan and Rubin (1990) propose to solve the problem of heaping by coarsening data over broad intervals centered around the heaping unit. In more recent work, Forster and Jones (2001) model smoking initiation and smoking cessation using UK data in discrete-time hazard models with and without controls for heaping. They implement solutions proposed by Torelli and Trivellato but find little evidence that heaping biases coefficients on cigarette tax in models of smoking duration. Pudney (2007) focuses on heaping in consumption expenditure data, and changes in heaped responses between consecutive waves. Similar to our findings, he observes multiple heaping rules. His analysis focuses on

⁴Worldwide in 2008 there are more than 24 ongoing or planned panel studies (in addition to the ones we use) and at least as many cross-sectional studies that collect or could collect retrospective data of the general form that we analyze.

patterns of transition between heaping points for the same individual. For each of the algorithms we develop, we demonstrate how the algorithm reduces mismatch bias.

We propose both parametric and semi-parametric algorithms. In one of our semi-parametric algorithms, we follow the approach of Heitjan and Rubin (1990) who propose to reduce the mismatch bias through the use of a log-likelihood function defined over a broader interval of the data (five year averages). We show that, in addition to the obvious loss of degrees of freedom this approach entails, it is the preferred method only under a very narrow set of conditions that are unlikely to hold in most data. We offer as an alternative, a parametric approach that preserves degrees of freedom and that accommodates multiple types of heaping. Our parametric solution requires the analyst to identify and include, as controls, observed covariates that predict heaping but which are uncorrelated with the main explanatory variable of interest. These covariates include indicators of the natural heaping units and their interaction with price. We show how bias is reduced in models that include (sets of) these factors.

The rest of the paper proceeds as follows: In section II we introduce and describe the US and international retrospective smoking data we use. In section III we document that, in distributions of when people quit (measured in different ways), heaping occurs in strikingly similar ways in all the countries we study. We show that the wording of retrospective smoking questions shapes the form that heaping takes. A key result in this description is that survey respondents use multiple heaping rules in every distribution we study. This finding has important implications for the method one uses to account for the mismatch bias. In section IV we describe the general problem - a particular type of measurement error - that heaping introduces. We sketch the general form of the bias and show how it enters the likelihood function. In Section V we use Monte Carlo simulation methods under an assumed heaping rule

to show that we can replicate the empirical distribution better using multiple heaping rules rather than a single heaping rule. We also establish a benchmark estimate of the attenuation bias that heaping introduces to a coefficient of interest. We present methods to mitigate the bias in section VI. The presence of multiple heaping rules that we document in Section III informs the methods we recommend and how we specify our empirical models. We present two types of algorithms to account for heaping bias - semi-parametric and parametric. We present and discuss results from our models of smoking cessation in section VII. We then turn our attention to heaping in retrospective reports of starting and ending jobs to highlight the ubiquitous presence of heaping and to illustrate how heaping can hide in ways not immediately obvious to an analyst. In Section VIII we briefly describe the employment data and illustrate how heaping occurs. We present and discuss results from the model of (non) employment in Section IX. Section X summarizes our general findings and concludes the paper.

Our results suggest that heaping biases coefficient estimates, often dramatically, and that our algorithms significantly reduce this bias. While we demonstrate the gains each algorithm affords in the specific context of models of smoking cessation and models of non employment, the algorithms apply to any retrospective data on time-varying events where the covariate of interest also varies across time. Therefore, our algorithms can be applied more generally to a wide range of retrospective data.

II. Data

We use retrospective data on smoking behavior and on employment. Our smoking data are drawn from the US Panel Study of Income Dynamics (PSID), the British Household Panel Study (BHPS), the Tobacco Use Supplements to the Current Population Surveys (CPS-TUS), and the German Socio-Economic Panel (GSOEP). Our employment data are drawn from the US

Panel Study of Income Dynamics (PSID). Here we briefly describe the retrospective smoking data for each data set. We discuss the PSID employment data in more detail in section VIII. We provide more details about each survey and the other data we use in our analysis in the Appendix. There we also provide interested readers with the web address for survey where even more details are available.

Retrospective questions on smoking behavior

Each of these surveys ask former smokers about when they first began and when they last smoked regularly. All of the surveys ask (with minor differences): “How old were you when you started smoking regularly?” The surveys differ in how they ask former smokers to report when they quit.⁵ The PSID and BHPS surveys ask, “At what age did you last smoke regularly?” (The BHPS prompts respondents to estimate the age they quit if they are unsure.) The CPS-TUS asks, “About how long has it been since you last smoked cigarettes every day?” Respondents can answer in days, weeks, months, or years. The GSOEP asks, “When did you give up smoking? (Please provide the year and, if possible, the month).”

As we show below, the wording of the question and the response categories that surveys offer respondents shapes the form that heaping takes. These wording differences matter not only because they condition how respondents answer questions but also because an inattentive analyst may inadvertently mask heaping if he fails to account for heaping of one form or another before he uses the data to construct a variable of interest.

After we restrict each sample to respondents who answered the relevant smoking questions, we are left with 4,097 person and 85,056 person-age year observations from the 2003

⁵The PSID and BHPS ask questions of anyone who ever smoked. The CPS-TUS and GSOEP ask questions of those who have smoked at least 100 cigarettes.

PSID, 8,238 person and 230,461 person-age year observations from the 2002 BHPS, 11,820 person and 256,864 person-age year observations from the 2002 GSOEP, and 236,077 person and 5,194,230 person-age year observations from the combined cross-sections of the CPS-TUS.⁶ Just under half of each sample quit smoking as of the year they were surveyed (2,035 in the PSID, 2,554 in the BHPS, 4,527 in the GSOEP, and 96,783 in the CPS-TUS).

In addition to a fairly standard set of demographic variables, we match the price of 20 cigarettes to each respondent in each year of his life. In the UK cigarette prices are measured as the real price of the British brand Capstan (in 2008 British pounds) from 1904 to 2002. In Germany prices are measured as the average price of cigarettes (in constant 2002 Euros) as reported in the German Statistical Office publication VI D 46/4-39. This series runs from 1951 to 2002. We draw data on the price of cigarettes in each US State from 1954 to 2003 from Orzechowski and Walker (2005). These data measure the average price per pack across all cigarette brands in a given state in November of each calendar year. In Britain and Germany our measure of cigarette prices varies over time but not across geography. In our US data, cigarette prices vary over time and also across states.

To match US cigarette prices by year and state to PSID or CPS respondents we need to assign a person to a state of residence in each year. This means we must assign PSID and CPS-TUS respondents to a state of residence in years before they were surveyed. For PSID respondents we impute a state of residence in each year of life using an algorithm developed in Lillard and Molloy (2008). This algorithm allows us to track PSID respondents over time as

⁶Although the PSID is a panel study and has asked retrospective smoking questions multiple time (see the Appendix), we only analyze data from 2003 because the other surveys' retrospective smoking data are cross-sectional. Lillard, Bar, and Wang (2008) analyze separately the role heaping plays when multiple observations on the same smoker are available.

they live in different states - a feature that allows us to estimate price effects on smoking cessation decisions with more precision. In the CPS-TUS we impose the assumption that a person always lived in his current state of residence and assign to him the corresponding cigarette prices (taxes) for all years.⁷

Table 1 presents summary statistics for the main variables of interest and some of our control variables. Smoking cessation rates were similar across the four data sets - around 2 percent of smokers quit each year in the PSID, CPS-TUS, and GSOEP samples and around 1 percent of smokers quit each year in the BHPS sample. US smokers paid between \$2.08 and \$2.25 in the average year for a pack of cigarettes while the average German and average British smoker respectively paid € 7.22 and £3.62 for a pack of 20 cigarettes.

Our control variables include life events and health shocks. Around 2 percent of smokers in the CPS-TUS and BHPS samples experience the birth of a child each year. A similar percentage of PSID smokers had a child but we lump together the birth of a child with retirement, marriage, and divorce in an index of life events. About 12 percent of the PSID sample experienced one of these life events in each year. We include this index in the models we run on the PSID data. About 2 percent of smokers got married in the BHPS each year.⁸ The PSID

⁷This immobility assumption is used in many studies that examine the effects of factors that vary across states (and over time). It is used by Douglas and Hariharan (1994) and Douglas (1998) in the context of smoking decisions. Acemoglu and Angrist (2000) use it in the context of schooling decisions. Lillard and Molloy (2008) use continuous residence histories available for a subsample of the PSID to document the degree of geographic residence mismatch bias the assumption introduces at various stages of the life-course. On average, an immobility assumption assigns people to their correct state of residence for about 75 percent of their person years. Because people move across state lines less frequently after age 30 (when quits are more likely to occur) the immobility assumption is less problematic for analyzing smoking cessation than it is for the analysis of smoking initiation.

⁸The Appendix describes in greater detail how time-varying data were constructed.

collects retrospective data on several types of health events that we use to map back physical and mental health shocks to the year an individual reported to have experienced them. These shocks include having a heart attack, a stroke, being diagnosed with lung disease, heart disease, asthma, cancer, hypertension, developing emotional problems, learning disabilities, and mental problems.

The samples are quite similar in terms of both time-varying and time-invariant characteristics. The average age of smokers while they smoked ranged from 32.5 (GSOEP) to 35.3 (BHPS). The slightly older age of smokers in the British sample is partly explained by the broader coverage of calendar years for smokers in that sample. The BHPS data cover the years from 1918 to 2002. The US data cover 1954 to 2003 and the GSOEP data cover the years 1951 to 2001. Half of the PSID and BHPS sample of smokers are women while 41 percent of the sample of German smokers are female and 47 percent of the CPS-TUS sample of smokers are female. All of the smokers were roughly 46 years old when the retrospective data were collected. Smokers in the US and Germany smoked about 21 years the average UK respondent smoked about 27 years.⁹

III. Heaping in the distribution of smoking cessation event measures

Figure 1 presents three distributions of smoking cessation events, each of which measures when an ex-smoker quit in units of time asked in the PSID, BHPS, CPS-TUS, and GSOEP surveys. The PSID and BHPS ask about the age a person last smoked regularly. The CPS-TUS asks how much time has elapsed since a person last smoked. The GSOEP asks the

⁹We also include controls for own education for the US (measured as time-varying in the PSID and current education in the CPS-TUS), parental education dummies for Germany, current household income for the CPS-TUS, and race and ethnicity for all countries. To preserve space we do not show the means for these variables however they are available on request.

calendar year an ex-smoker quit. Panel A plots PSID and BHPS data on the distribution of the age at which an ex-smoker quit. Panel B plots CPS-TUS data on the years since a person last smoked. Panel C plots GSOEP data on the calendar year in which an ex-smoker quit.

The distributions plotted in Figure 1 highlight two striking patterns. First, retrospectively reported data are heaped on the units of time to which it is natural to round - those that are evenly divisible by 5. Second, respondents to each survey appear to differentially round to units divisible by 5 and units divisible by 10. This pattern suggests that respondents are using at least two heaping rules - round to the nearest multiple of 5 or round to the nearest multiple of 10 - and that they do so with different probabilities. We respectively indicate these two types of heaping by the labels A5/A10 for age heaping, T5/T10 for time heaping, and C5/C10 for calendar year heaping. The relatively greater proportion of the sample that heaps on units of 10 is preliminary evidence that there are at least two types of heaping rules used in response to retrospective survey questions. The presence of multiple heaping rules is going to be important when we discuss methods by which to reduce the mismatch bias that heaping causes.

While it seems apparent from Figure 1 that respondents are using multiple heaping rules in the units mentioned in each survey's question, only closer inspection reveals that respondents are also heaping in units other than those mentioned in the survey question. Figure 2 examines more closely the distribution of responses to the CPS-TUS survey. In Panel A we again plot the distribution of the amount of time since an ex-smoker quit but we modify the distribution in two ways. First, we plot the distribution using a calendar scale by subtracting elapsed time from 2002 (the survey year). Second, we remove from the distribution two types of respondents; those who reported an elapsed time that was a multiple of 5 (T5 and T10 responses) and those whose implied quit age (age in 2002-elapsed time) was a multiple of 5 (A5 and A10 responses).

The resulting marginal distribution suggests that the remaining respondents are more likely to report an elapsed time that corresponds to a calendar year that is divisible by 10 (i.e. a C10 calendar year such as 1990, 1980, 1970, 1960) than they are to report a time that is the next adjacent calendar year. There is no obvious tendency to heap on C5 calendar years.

Panel B similarly plots the distribution of elapsed time but on an age scale. Here a person's quit age equals their age in 2002 minus the time since they smoked regularly (in years). We similarly plot the marginal distribution of implied quit ages after removing from the distribution the respondents whose time since quit responses corresponded to T5/T10 elapsed time or C5/C10 years. The marginal distribution of the implied quit age also shows some evidence that may indicate heaping on ages that are multiples of 5 or 10 (especially for age 30, 35, and 40). The lack of heaping at the upper tail of the implied age distribution is not surprising because older ex-smokers are also more recent quitters. Presumably respondents recall and report more recent events with more accuracy.

Together Figure 1 and Figure 2 establish two important points. First, heaping is pervasive.¹⁰ Second, respondents use multiple heaping rules within and across multiple units of time. This latter point is important not only because it constrains the methods one can use to solve the mismatch bias but also because analysts may inadvertently mask heaping if they fail to recognize that heaping is present before they manipulate raw data to construct a variable of

¹⁰Almost identical patterns of heaping are observed for many other outcome that include the years a person worked full-time, the number of times a person volunteered at a religious organization last year, the elapsed time since a person developed asthma, emotional problems, heart disease, arthritis, diabetes, high blood pressure, lung disease, mental problems, had a stroke, was diagnosed with cancer and in the time that must pass until a person will be fully vested in a pension plan. However, heaping is less likely for salient life events. For example we see no heaping in the distribution of age at first marriage.

interest (e.g. if they construct a smoker's quit age from data on elapsed time but fail to recognize that some respondents heap answers on calendar time units).

Before we present our methods to correct for heaping, we briefly describe the statistical problems it introduces and produce baseline estimates of the bias that it might generate.

IV. Heaping as measurement error

When respondents use (multiple) heaping rules to round the timing of an event to something other than the true timing, the likelihood function changes. In the parametric approach we can write the probability of the response (the actual quitting age, x) as $\Pr(X=x) = f(x|\boldsymbol{\theta})$ where $\boldsymbol{\theta}$ is a vector of parameters, and f is a probability distribution function. We also assume a parametric form of multiple heaping patterns, so that the i -th heaping rule has the distribution function h_i : $\Pr(Y=y|X=x) = h_i(y|x, \lambda_i)$. For example, we may have one heaping rule for every multiple of 5 which takes the form of a Bernoulli random variable with parameter (λ_5) p_5

$$Y = y|X = x = \begin{cases} y = x & \text{with probability } 1 - p_5 \\ T5 & \text{with probability } p_5 \end{cases} \quad (1)$$

where $T5$ is the multiple of 5 that is closest to x . Similarly, we can define a heaping rule for multiples of 10 with parameter p_{10} , or, using the previous notation in the paper, define similar rules for $A5$, $A10$, $C5$, $C10$, and so forth.

Of course, we can modify the parametric form so that the probability of heaping to a certain value, y , depends on covariates (e.g., age at time of survey) or define a different support for each heaping rule. For example, we can consider the following (symmetric) heaping rule

$$\Pr(Y = T10 | X = x) = \begin{cases} 0 & \text{if } |x - T10| > 5 \\ \frac{p_{10}}{1 + |x - T10|} & \text{if } |x - T10| \leq 5 \end{cases} \quad (2)$$

According to this rule, the probability of observing a *T10* age, given that the true age is x , increases the closer *T10* is to x .

In certain cases it may be reasonable to use non-symmetric heaping distributions, and in many cases, including the one in this paper, the support sets of different rules can overlap. Also, each heaping point can be associated with a different heaping rule. For example, T5 rules may apply to the range 25-55, but for 15, 20, and 60 and above we may use different rules.

In the more general case, suppose there are K heaping rules in our model, denoted by h_1, \dots, h_K , and denote the identity function by I (so that $I(x)=x$). When a true age is drawn according to $f(x|\theta)$, there is a probability mass function $g(h|x, \mu)$ that determines probability that the respondent will select a function h from $H = \{h_1, \dots, h_K, I\}$ and report $h(x)$. Note that the selection of the heaping pattern by the respondent can depend on x .

Hence, conditional on the true age x , the probability that the respondent will select a certain heaping rule, h , is $g(h|x, \mu)$, and given x and h , the probability of observing y is $h(y|x, \lambda_h)$.

So for a given x , the probability of observing y is

$$\sum_{h \in H} g(h|x, \mu) h(y|x, \lambda_h) \quad (3)$$

The likelihood of a single observation is:

$$L(\mu, \lambda, \theta | y_j) = \int f(x | \theta) \sum_{h \in H} g(h | x, \mu) h(y_j | x, \lambda_h) dx \quad (4)$$

and the likelihood function for n observations is:

$$\prod_{j=1}^n L(\mu, \lambda, \theta | y_j) \quad (5)$$

Our model encompasses and generalizes previously suggested models. For example, the widely used method in Heitjan and Rubin (1990) deals with a data set in which the time-unit is months and two groups of heaping points are observed (multiples of 6 or 12 months). The Heitjan and Rubin paper discusses a few rounding procedures that depend on the culture (rounding up, down, or to a mid-point of an interval). They propose a likelihood function similar to ours, and as in our paper, their model allows for overlapping heaping intervals. In their discussion they consider only the case where the probability of heaping inside the relevant interval is uniform. Wright and Bray (2003) propose a mixture model for rounding where the observed values take one of two forms (rounded to the nearest integer or to one decimal point). Their heaping rules do not admit overlapping intervals and the probability of heaping in their paper does not depend on the true value.

V. Monte Carlo simulation and analysis

To further emphasize that respondents use multiple heaping rules, we use Monte Carlo methods to replicate the empirical distribution for one data set. The raw data on the age smokers last smoked regularly as reported in the BHPS and PSID appear to be distributed according to an Inverse Gaussian (inverse Normal) distribution. The pdf and cdf for the Inverse Gaussian (IG) distribution are given by:

$$f(y) = \sqrt{\frac{B}{2\pi y^3}} \exp\left(-\frac{B}{2y}\left(\frac{y-A}{A}\right)^2\right) \quad (6)$$

and

$$F(y) = \Phi\left(\sqrt{\frac{B}{2\pi y^3}} \frac{y-A}{A}\right) + \exp\left(\frac{2B}{A}\right) \varphi\left(\sqrt{\frac{B}{y}} \left(\frac{-y-A}{A}\right)\right) \quad (7)$$

where A and B denote mean and scale parameters.

We use the Inverse Gaussian distribution to replicate the main features of the heaped data by using a mixture of two heaping rules - one on A5 ages that are evenly divisible by 5 and one on A10 ages that are evenly divisible by 10. We assume ex-smokers heap on primary and secondary heaping unit using the following formula:

$$\text{Heaping probability} = P1*(1/(1+|Ax-A5|)) + P2*(1/(1+|Ax-A10|)) \quad (8)$$

where P1 is the probability that a person heaps on A5 ages, P2 is the probability that a person heaps on A10 ages, Ax is the true quit age, A5 is the closest age that is a multiple of 5, and A10 is the closest age that is a multiple of 10. In our simulation we set P1=P2=.10. Of course P1 and P2 need not be constants. As noted by Hausman et al. (1998) and Torelli and Trivellato (1993) heaping probabilities may be a function of other covariates (e.g. the time that has elapsed between the age an ex-smoker actually quit and the age at which retrospective information is collected). Below we show evidence that P1 and P2 vary with this elapsed time.

In figure 3 we plot one of the 100 Monte Carlo draws from the non heaped distribution of age last smoked regularly (using an assumed IG distribution for those ages), one of the 100 draws from the distribution that includes the heaping rule (8), and the empirical distribution of

the reported quit ages of the 2003 PSID ex-smokers. These three distributions are respectively labeled “IG MC,” “IG-MC Heaped,” and “US ex-smokers (PSID).” As can be seen, the heaping rule described above replicates the main features of the observed (heaped) PSID data. Mass is shifted to A5 and A10 ages with relatively more to A10 ages than to adjacent A5 ages. Thus, it is apparent that at least two heaping rules are present in these retrospectively reported data.

We also use the above heaping rule to get a first glimpse of the potential extent of mismatch bias associated with heaping. To do so, we first assume a distribution for the raw data on smoking prevalence by age. Using that distribution we then artificially cause 10 percent of the sample to quit in a particular year. For this exercise we require people to quit so that the age of smoking cessation is distributed IG (Inverse Gaussian). That is, we require 20 percent of smokers to quit (all of whom began at the same age) and then distribute who quits at each age according to fraction of smokers that the IG quit age distribution implies. Our final step is to age everyone forward by fifteen years and then apply a known heaping rule. The known heaping rule yields a “reported quit age” that differs in a known way from the true quit age. Our estimate of the bias is then the simple difference in the fraction of people who quit in the true year with the fraction of people whose reported quit age (i.e. the heaped age) corresponded to that year.

Not surprisingly, bias increases with heaping. In this very artificial setting, the estimated coefficient on price can be biased downwards by between 1.3 to 13 percent of the true coefficient when we allow P1 and P2 to vary from .05 to .50. As we show below, these probabilities are not likely to be equal. Because the probability of heaping varies systematically with observable characteristics, we can use the extra information to mitigate systematic differences in the bias.

VI. Correcting for heaping - three approaches

We next describe three methods by which analysts might reduce the bias we described above. To provide a benchmark we first estimate a model we label as “**Naive**” that takes the reported quit data as given.

Recent quitters

Our first method to correct for heaping restricts the sample to current smokers and ex-smokers who quit within the past five years. This sample restriction assumes that these smokers do not heap because they quit recently. We label these models as “**Recent**.” This method has the advantage that it is less prone to mismatch bias but has the disadvantage that it only uses variation in cigarette prices over five years to explain quitting behavior. When prices vary seldom or if there is no geographic variation in prices a single year, this method may be difficult to implement.

Coarsening data

Our second method follows the suggestion of Heitjan and Rubin (1990) and “coarsens” the data by averaging of all right-hand side variables over five years centered around T5 ages/time/calendar years. The dependent variable in these models equals “1” if a smoker quit in any year inside the interval and “0” otherwise. We label these models as “**Coarse**.” This method obviously involves efficiency losses because it throws away variation in explanatory variables that occurs inside the interval over which data are averaged. It also has other drawbacks we discuss below.

Parametric controls for heaping and associated bias - the “T5” method

Our third and preferred method retains all observations but introduces parametric control variables to flag ages and years that correspond to A5, C5, or T5 heaping units. We account for potential mismatch bias by interacting each of these indicator variables with cigarette price. We

also include each respondent's age when surveyed as a measure of more general recall bias (of a type not associated with a T5 age/time/calendar unit). We label this method as the “**T5**” method.

As we noted above, any sample may include up to seven types of survey respondents - those who answer truthfully and those who round answers to units of 5 or to units of 10 in age, time, or calendar years. While our **T5** method will capture most of the bias present when some respondent round to T10 units, we also estimate a more flexible specification we label as the “**T5+**” method that adds indicators for age/time/calendar years that are evenly divisible by 10 and their interaction with cigarette price. Together with a variable that measures age when surveyed and the interaction of survey age with price, this more flexible specification allows for the six types of heaping. We also estimate variants of the other two methods that we label “**Recent+**” and “**Coarse+**” in which we include measures of heaping that are not resolved by the treatment of the data those methods involve. For example, the **Coarse** method requires one to pick a particular type of heaping (e.g. A5 heaping) and coarsen the data appropriately. As we explain below, this treatment of the data fails to fully account for mismatch bias from even broader heaping rules (e.g. A10) and it fails to fully account for mismatch bias from other types of heaping (T5/T10 and C5/C10) except under the very unusual conditions we describe above. We therefore include in **Coarse+** indicators of heaping in A10/T10/C10 and their price interactions. Similarly, we include other types of heaping indicators (and their interaction with price) in **Recent+** specification.

Our basic estimating equation parameterizes heaping of these types in the general form:

$$y_t = \beta_0 P_t + \beta_1 g(A5, A10, C5, C10, T5, T10, Age_{t+k}) + \beta_2 g(\bullet) P_t + \beta_3 X_t + \varepsilon_t \quad (9)$$

where $g(\bullet)$ is, in our specifications, a vector of indicator variables that flag observations in years that correspond to natural heaping units in age, calendar, or chronological time. The coefficient β_2 on the interaction between these heaping indicators and cigarette price captures the bias associated with heaping. X_t is a vector of other control variables.¹¹

To graphically illustrate the advantages of the **T5** or the **T5+** specifications, Figure 4 presents the true and reported quit ages for four ex-smokers who heap. All four smokers report that they quit at age 40 but two of the ex-smokers round using a A10 rule. One A10 heaper quit at age 36 and rounds up to report he quit at age 40. The other A10 heaper quit at age 44 and rounds down to report he quit at age 40. Similarly, one A5 heaper quit at age 38 but rounds up to report he quit at age 40. The other A5 heaper quit at age 42 but rounds down to report he quit at age 40.

If one coarsens the data over a five year interval centered on A5 ages, the coarsening method will resolve bias for A5 and A10 heapers who actually quit at ages inside the interval. However, the coarsening method will not resolve the bias for A10 heapers whose actual quit age lies outside the coarsening interval. From Figure 4 it should also be fairly obvious that, even when there were no A10 heapers, coarsening the data on five-year intervals centered on A5 ages will similarly leave unresolved the bias for ex-smokers who heap on calendar time (C5 or C10) or chronological (elapsed) time (T5 or T10) units when their true quits lie outside of the

¹¹This specification differs from the solution proposed by Torelli and Trivellato (1993) because the indicator variables are exogenous to each individual. They flag person-years that correspond to A5/A10, C5/C10, or T5/T10 observations (not to ex-smokers who heap their answers). The specification is similar to one used by Crockett and Crockett (2006) though they do not present a likelihood function or discuss mixture models. Nicolas (2002) uses Spanish data to estimate a variant of our **T5** specification but he only includes T5 heaping. He finds evidence of some downward bias.

coarsened unit interval. If there were no heapers on 10 year rules, the coarsening method would resolve heaping bias from A5, C5, and T5 heapers only in the unlikely case that all people were born every five years (so they all turned an A5 age in the same calendar year) and all were surveyed in a calendar year that was a C5 year (so all T5 heapers reported elapsed time that corresponded to an A5 age). In the more general case, the coarsening method will leave heaping mismatch bias unresolved. It is also apparent that restricting the sample to those who quit recently will fail to fully resolve mismatch bias. However, the **T5** and **T5+** specifications are flexible enough to accommodate all types of heaping.

VII. Results - Smoking cessation

Table 2 reports results from discrete-time hazard models of the probability a smoker quits in a given year. We estimate the above models (labeled at the top of each column) for the 2003 PSID, multiple waves of the CPS-TUS, the 2002 GSOEP and the 2002 BHPS. In all cases we report probit coefficients on cigarette price, the heaping indicators, and their interaction with cigarette price. At the bottom of each column we also report the implied marginal change in the probability a smoker quits when the price of cigarettes increases by one standard deviation. This marginal probability is computed by predicting a baseline probability for each sample member, adding one standard deviation to the cigarette price, predicting the new probability for each sample member. We then average the two predicted probabilities over the whole sample. The difference is the estimated marginal effect we report.

For models that include heaping indicators and the interaction of cigarette price with the heaping indicator or the survey age, we estimate a predicted marginal effect by constraining at zero the coefficients on the heaping indicators and the interaction between the heaping indicator and price and the interaction between the survey age and price. That is, we impose the

assumption that the association between price and heaping indicators/survey age reflect only mismatch bias and that the differential probability to quit at A5/T5/C5 (10) time units reflects only reporting errors. We then compute a baseline quit probability for all respondents, increase the price by a standard deviation and compute a second quit probability. As before, we average these two probabilities over all sample members and take the difference as our estimate of the implied effect of a change in cigarette price.

Note that we additionally divide the implied marginal effect for the **Coarse** and **Coarse+** models by five to convert the implied effect from the probability a person quits in a five year window to the probability a person quits in any year in that five-year window.

The results in Table 2 generally show that the **Naive** model yield implied marginal effects of price change on the probability a smoker quits that are small and sometimes of the opposite sign than would be predicted by demand theory. Estimates range from -.002 in the CPS-TUS sample to .0032 in the GSOEP. Restricting the sample to current smokers and smokers who quit recently results in generally large predicted effects of a standard deviation in price. The implied effects shown in the columns headed **Recent** range suggest that a standard deviation in cigarette price raises the probability a smoker quits by 1 percentage point in the CPS-TUS, by 2 percentage points in the BHPS, and by 4.3 percentage points in the GSOEP. Only in the 2003 PSID does the implied effects remain small (.005) and roughly the same as the effect implied by the coefficients in the **Naive** model.

The data coarsened over a five year interval (centered on A5 ages in the PSID and BHPS, on T5 years in the CPS-TUS, and on T5 calendar years in the GSOEP) not only fail to resolve the mismatch bias but appears to exacerbate it. The implied marginal effects shown at the bottom of the columns labeled **Coarse** are all smaller (and sometimes of the opposite sign

predicted by theory) than the implied effects from the **Naive** model. The implied marginal effect of a standard deviation increase in cigarette price range across the samples from -.005 in the CPS-TUS to .0021 in the GSOEP. By contrast, the coefficients estimated on the heaping indicators and their interaction with cigarette price in our preferred specification, **T5**, suggest the presence of substantial mismatch bias of the expected form. When that bias is controlled, the implied marginal effect of a standard deviation in cigarette price is .105 in the PSID, .027 in the CPS-TUS, .059 in the GSOEP, and .008 in the BHPS. Except for the BHPS, relative to the implied effects for all other specifications, the implied marginal effects are largest for the **T5** models.¹²

Table 3 estimates more flexible specifications of the basic **Recent**, **Coarse**, and **T5** models. In particular we add measures of heaping that is likely still present in the data for each sample. In general, the addition of heaping indicators for 10 year units and their interaction with price reduces mismatch bias in all model specifications. The biggest reduction in bias is seen in the **Coarse+** specification. Using the coefficient estimates on price from the basic specification in Table 2 an analyst would have concluded that increases in the cigarette price has either small or no statistically significant effect on quit probabilities (or even worse that increases in price reduce the probability a smoker quits). However, once one accounts for mismatch bias from people who round on A10/T10/ and/or C10 units, the results in Table 3 in the columns labeled **Coarse+** now imply that a standard deviation increase in the cigarette price raises the probability a smoker quits by between .017 and .019 (from a baseline quit probability of .02). Only in the BHPS sample does the **Coarse+** specification result in an implied effect of a price change that is

¹²Note that the effect of cigarette price will be biased downwards in all models estimated on the CPS-TUS because we assume a person always lived in one state (his current state).

the opposite sign predicted by theory. Adding A10/T10/C10 indicators and their price interactions in the T5+ model shows that there is some mismatch bias associated with rounding on units of 10. However, the implied effects of a marginal change in the cigarette price are essentially unchanged.

Robustness checks

To account for the mismatch bias that heaping introduces, it must be the case that the smokers who heap their reported quit ages on A5/A10, C5/C10, and T5/T10 years are no different in unobservable ways than smokers who report a quit age in some other year. That is, the heaping indicators must be orthogonal to the determinants of smoking. While that assumption seems plausible, it is of course possible that unobserved factors determine both the probability an ex-smoker quits and the probability that he reports having quit on a heaped age/year. One might suppose, for example that smokers who pay little attention to price may also pay not report accurately. It is possible that these smokers respond less to price increases.

We examine these questions in two ways. First we examine observable differences between heapers and non heapers. Heapers do differ from non heapers on observable characteristics we include in our models. For example, older ex-smokers are more likely to heap. Heapers are older than non heapers on average. The difference in the mean ages of the two groups of ex-smokers in the PSID, CPS-TUS, GOEP, and BHPS respectively was 7, 3, 3, and 8 years. In all samples heapers were more likely to be male and were better educated but (in the CPS-TUS) did not differ in terms of family income. While better educated quitters may be less responsive to cigarette price changes it is not obvious why that should be the case.¹³

¹³In unreported regression we estimate the probability an ex-smoker heaps on A5 ages. The probability of heaping is generally uncorrelated with observable characteristics except the

Second, in the PSID we are able to observe multiple reports of the age a smoker quit across up to six waves of the data (1986, 1990, 1999, 2001, 2003, and 2005). To examine whether unobserved characteristics might be driving our results, we create an indicator variable that flags ex-smokers who did not report a heaped age in 2003 but did report an A5 age in any of the other years for which retrospective smoking data were available. We then interact that variable with cigarette price. We rerun our **T5+** model with indicators for heapers in 2003, heapers in a year other than 2003, the interaction between these two variables (to identify people who heaped both in 2003 and also in some other year) plus the interaction between cigarette price and the flag for quitters who reported an A5 age in a year other than 2003. In principle, the coefficient on the interaction term tests whether heapers respond differently to price than do non heapers. The results suggest that heapers are marginally *more* responsive to price than non heapers (though the coefficient is not statistically different from zero at conventional significance levels).

VIII. Another application - minimum wage and employment

We next turn to a second outcome - not being employed - to illustrate two additional and important points. First, we show that heaping is present in other types of retrospective data. Second, and perhaps more importantly, we show that the natural heaping unit in retrospective data on employment is not the same as it was for the smoking data and that it takes a form that is not immediately obvious. Finally we demonstrate again that our parametric specification reduces bias associated with the heaping.

difference between the age when interviewed and the age a smoker said he quit.

Economists have long investigated whether the minimum wage empirically impacts employment as theoretically predicted. That question remains open in part because of a debate through a set of studies that went back and forth about the empirical evidence by David Card and Alan Krueger (1994a, 1994b, 1995a, 1995b) and David Neumark and William Wascher (1992, 2000, 2006). We examine the question here not to try to settle this debate but because retrospective reports about when a person begins or ends employment is an example of a topic of interest that is likely affected by heaping.

For this exercise we draw data from the 2003 PSID on the month and year a person started and stopped working at up to four jobs for a given employer. The question we use takes the form, “When did you (HEAD/your WIFE) start and when did you stop working for this employer? Please give me all of the start and stop dates if you have gone to work for (this employer/yourself) more than once.” Respondents report both the month and year they started or stopped working in each job. This questions are answered by PSID “heads” and “wives” who are currently employed and by those currently not working.

We use the information on the month and year each person began or ended a job to construct a time series of monthly indicator of employment. The employment indicator equals “1” if a month falls between the start or end date of any job and “0” otherwise. We then create a “non employed” indicator that equals 1 minus the employment indicator. This indicator thus captures people who are unemployed and those who are out of the labor force for any other reason. We drop all months prior to the first month of reported work. We keep observations from January 1970 through December 2002. Finally, we merge individual data on age, race, sex, years of completed schooling, and state of residence in each calendar year.

We use the data on state of residence to merge data on the higher of the state or federal minimum wage each person faced. These data were drawn from the compilation of state labor laws published in the article titled “State Labor Legislation Enacted in [year]” that has appeared in the January edition of the Monthly Labor Review for all the years included in the sample. These data are supplemented by data on wage orders in various industries in each state as reported in state legislation (see Lillard.2001).

We estimate models of non employment separately by sex for two samples of women and men. We estimate the models on the full sample and on the sample of women and men with 12 or fewer years of education. This sample restriction tries to capture the population most likely to be affected by changes in the minimum wage.

After we restrict the sample to respondents who answered the relevant employment questions, the sample sizes for the models of non employment are 4,457 person and 354,331 person-month observations for the sample of all women, 2,913 person and 208,819 person-month observations for the sample of low-educated women, 4,223 person and 419,749 person-month observations for the sample of all men, and 2,704 person and 237,494 person-month observations for the sample of low-educated men. In the average month, roughly 19 percent of each sample was not employed and faced a minimum wage of \$6.50 (in constant 2008 dollars). Table 4 presents summary statistics.

In Figure 5 and Figure 6 we present distributions of months people started or stopped a job to show that heaping is present. Figure 5 presents, for men and women who were unemployed at the 2003 survey, the distribution of the month they reported having last worked. Panel A shows the distribution for unemployed men. Panel B shows the distribution for unemployed women. For both sexes, there appears to be heaping in January, June, and

December (primarily for women). Of course some of this heaping may reflect true behavior and not recall bias. It is plausible, for example, that men and women stop working in June (by quitting or because they have seasonal jobs) if they have young children or if they move from one job to another. However, from the perspective of estimating the effect of minimum wage on non employment, the source of the heaping is relatively unimportant. One must still account for the possibility of a differential effect of the minimum wage in these months.

Figure 6 shows one intuitively obvious and one much less obvious way in which heaping presents itself in the month respondents report having started their first job. The obvious heaping pattern is in the higher probability that a person reports having started a job in January. Between 30 and 47 percent of men and women reported they began their first job in January. However, even here there is evidence that people heap their answers. In particular, we divide the sample into those who began their first job within the past five years, those who began between 5 and 10 years ago and those who began more than 10 years ago. Figure 6 shows that respondents are as much as 15 percentage points more likely to report having started in January if the start was temporally further back in time. This difference indicates likely heaping.

Figure 6 also presents evidence that suggests that people heap in another way that is much less obvious. In particular, people report a start month that coincides with the month they are being surveyed by the PSID. In 2003, the PSID interviewed respondents between March and November. Here, we present the distribution of start months for people who were surveyed in June. It is apparent that people are more likely to report having started to work in the month of the survey and that the propensity of doing so is greater for people who started their first job

longer ago.¹⁴ Because nothing obviously connects the month the PSID surveyed its respondents to the month that person began to work, the heaping of reported start months on the month a person was surveyed strongly suggests they are heaping their answers. It also suggests that heaping is a potential source of mismatch bias when estimating the effect of time-varying determinants of employment.

IX. Results - non employment

Table 5 presents coefficient estimates for the probability a person is not employed. We present models for four samples - all women, women with 12 or fewer years of completed schooling, all men, and men with 12 or fewer years of completed schooling. For each sample we present coefficient estimates from a naive model that ignores heaping and the model that corresponds to equation (9) above (with heaping indicators defined appropriately). We use five heaping indicators - one that indicates months that correspond to the month a person was surveyed, and indicators for January, June, September, and December. At the bottom of each table we also present the implied marginal change in the probability of not being employed for a standard deviation change in the minimum wage. We show the percent of each sample that is not employed and the implied marginal change relative to the baseline rate of non employment.

All models show that the probability of non employment is higher in months when the minimum wage is higher. In the naive models, a standard deviation increase in the minimum wage raises the probability a person is not employed by between .68 and .92 percentage points.

More to the point for this paper, mismatch bias from heaping is present in all four samples but the effects are largest among women, especially low educated women. When one

¹⁴The same general pattern is present for people surveyed in other months though the pattern is not as obvious in all months.

accounts for heaping, the implied marginal effect of a change in the minimum wage increases from .0068 to .0087 and from .0091 to .0116 among women and low educated women respectively. Similarly for men, when one accounts for heaping, the implied marginal effect of a change in the minimum wage increases from .0049 to .0062 and from .0076 to .0093 for men and low educated men respectively.

We leave the debate about whether these are small or large effects to others. Our point here is that heaping biases estimates of interest for labor market outcomes in similar ways as it did for estimates of interest for smoking cessation. In the presence of heaping, estimated coefficients will be biased on any explanatory factor that varies across the same unit of time as the dependent (heaped) variable. Obvious candidates for such variables include average wages, unemployment benefit levels, unemployment rates and a host of other policy variables.

X. Discussion and conclusions

In this paper we have investigated a well-recognized feature of retrospective data on events that occur in the distant past - heaping on some natural unit. We have demonstrated that heaping is ubiquitous across outcomes and across countries. We have shown that when people answer surveys, they use multiple rules to heap responses. Through Monte Carlo simulations and empirical analysis we have shown that the presence of heaping causes bias. Here we have focused on outcomes whose timing is affected by heaping. In this class of outcomes, heaping causes mismatch between the observed behavior and the value of the determinant that caused the true behavior. This mismatch biases estimates of coefficients of interest.

We presented three methods one might use to mitigate mismatch bias. These involve restricting samples to recent events, coarsening the data to redefine the outcome of interest to a broader time interval, and introducing parametric controls to account for the presence and type

of heaping. Of these three, we show that the latter is preferred because it retains sample size and allows analysts to take advantage of more variation in the policy variables of interest.

In the context of retrospective smoking data, there are distinct advantages that retrospective data offer over cross-sectional data - the primary one being that taxes/prices don't change frequently. Retrospective data allow one to model smoking behavior over substantially longer time periods (e.g. in our British data from 1904 to 2001) that allows one to take advantage of more variation. The results in Table 2, Table 4, and Table 5 demonstrate that one can reduce mismatch bias in retrospectively reported data if one estimates a model that includes indicators to flag the natural heaping units and a set of interactions between those indicators and the covariate of interest. We show, in the analysis of smoking cessation, that methods that restrict the sample to recent quitters or that coarsen the data do not resolve the problem when people use two or more heaping rules.

Although we have focused this paper on the analysis of retrospective smoking data and on retrospective employment data, the algorithms we develop and implement apply much more generally to other types of data in which heaping is prevalent. Because longitudinal and cross-sectional data are enriched in so many ways when retrospective data are collected, the algorithms we present here offer the possibility of taking fuller advantage of retrospective data.

References

- Acemoglu, Daron and Angrist, Joshua. (2000). "How Large are the Social Returns to Education? Evidence from Compulsory Attendance Laws," *NBER Macro Annual*, No. 15, 2000, reprinted in *Recent Developments in Growth Theory*, D. Acemoglu, ed., Edward Elgar Publishing Ltd, 2004.
- Alford, B. W. E. (1973). W.D. and H.O. Wills and the Development of the U.K. Tobacco Industry, 1786-1965. New York: Barnes & Noble, 1973.
- Beckett, M., J. DeVanzo, et al. (2001). "The quality of retrospective data - An examination of long-term recall in a developing country." *Journal of Human Resources* **36**(3): 593-625.
- Bound, John, Charles C. Brown, Nancy Mathiowetz. "Measurement Error in Survey Data." In *Handbook of Econometrics* edited by E.E. Learner and J.J. Heckman. Pp. 3705-3843. New York: North Holland Publishing. 2001.
- Card, David and Alan B. Krueger. (1994a). "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania," *American Economic Review*, Vol. 84, No. 4. September 1994, pp. 772-793.
- _____. (1994b). "Minimum Wages and Employment: A Case Study of the Fast-Food Industry in New Jersey and Pennsylvania: Reply," *American Economic Review*, Vol. 90, No. 5, December 1994, pp.1397-1420.
- _____. (1995a). *Myth and Measurement: The New Economics of the Minimum Wage* (Princeton, New Jersey: Princeton University Press), 1995(a).
- _____. (1995b). "Time-Series Minimum-Wage Studies: A Meta-analysis," *American Economic Review* Papers and Proceedings, Vol. 85, No. 2, May 1995(b), pp. 238-243.
- Crockett, A. and Crockett, R. (2006). "Consequences of Data Heaping in the British Religious Census of 1851." *Historical Methods*. (Winter) 39(1): 24-39.
- Douglas, Stratford (1998). "The Duration of the Smoking Habit." *Economic Inquiry* 36 (1):49-64.
- Douglas, S. and Hariharan, G. (1994) "The hazard of starting smoking: estimates from a split population duration model," *Journal of Health Economics*, 13: 213-230.
- Forster, M. and A. M. Jones (2001). "The role of tobacco taxes in starting and quitting smoking: duration analysis of British data." *Journal of the Royal Statistical Society Series a-Statistics in Society* **164**: 517-547.
- Haenszel, William, Shimkin, Michael B., and Miller, Herman P. (1956). "Tobacco Smoking Patterns in the United States." Public Health Monograph No. 45. Washington, DC: US Department of Health, Education, and Welfare, Public Health Service.

- Hausman, J. A., Abrevaya, J., and Scott-Morton, F. M.. (1998). "Misclassification of the dependent variable in a discrete-response setting." Journal of Econometrics **87**(2): 239-269.
- Heitjan, D. F. and D. B. Rubin (1990). "Inference from Coarse Data Via Multiple Imputation with Application to Age Heaping." Journal of the American Statistical Association **85**(410): 304-314.
- Hirsch, Barry T., and Edward J. Schumacher. (2004). "Match Bias in Wage Gap Estimates Due to Earnings Imputation," Journal of Labor Economics. Vol. 22, No. 3, July, pp. 689-722.
- Kenkel, Donald, Lillard, Dean, and Mathios, Alan. (2004). "Accounting for Measurement Error in Retrospective Smoking Data." Health Economics. 13(10): 1031-1044.
- Lillard, Dean. (2001). "A History of State Minimum Wages and Overtime Pay Requirements: 1940-2001." manuscript. Department of Policy Analysis and Management. Cornell University.
- Lillard, Dean R., Bar, Haim, and Wang, Hua. (2008). "Tell Me Again: Using Repeated Reports in Panel Data to Reduce Mismatch Bias." Manuscript. Department of Policy Analysis and Management, Cornell University.
- Lillard, Dean, and Molloy, Eamon. (2008). "Putting People in Their Place: Reducing Mismatch Bias." Manuscript. Department of Policy Analysis and Management, Cornell University.
- Little, R. J. A. (1992). "Incomplete data in event history analysis." In *Demographic applications of event history analysis*, edited by James Trussell, Richard Hankinson, and Judith Tilton. 209-30 pp. Clarendon Press: Oxford, England.
- Neumark, D. and Wascher, W. (1992). "Employment Effects of Minimum and Subminimum Wages: Panel Data on State Minimum Wage Laws," Industrial and Labor Relations Review, Vol. 46, No. 1, October 1992, pp. 55-81.
- _____. (2000). "The Effect of New Jersey's Minimum Wage Increase on Fast-Food Employment: A Reevaluation Using Payroll Records." American Economic Review, Vol. 90, No. 5, December 2000, pp.1362-96.
- _____. (2006). "Minimum Wages and Employment: A Review of Evidence from the New Minimum Wage Research," Unpublished manuscript, 2006.
- Nicolas, A. L. (2002). "How important are tobacco prices in the propensity to start and quit smoking? An analysis of smoking histories from the Spanish National Health Survey." Health Economics **11**(6): 521-535.
- Orzechowski and Walker (2005) The Tax Burden on Tobacco: Historical Compilation 2005. Vol. 40, 2005.

Peters, H. E. (1988). "Retrospective Versus Panel Data in Analyzing Lifecycle Events." Journal of Human Resources **23**(4): 488-513.

Pudney, Stephen. (2007). "Heaping and Leaping: Survey response behaviour and the dynamics of self-reported consumption expenditure." Manuscript. Institute for Social and Economic Research, University of Essex.

Tauras, John and Frank Chaloupka (2001). "Determinants of Smoking Cessation: An Analysis of Young Adult Men and Women," In *Economics of Substance Abuse*, Michael Grossman and Chee-Ruey Hsieh, editors. Edward Elgar, previously circulated as NBER Working Paper 7262.

Tobacco Manufacturers Association http://www.the-tma.org.uk/page.aspx?page_id=42

Torelli, N. and U. Trivellato (1993). "Modeling Inaccuracies in Job-Search Duration Data." Journal of Econometrics **59**(1-2): 187-211.

US Department of Commerce, Census Bureau (various years). National Cancer Institute Sponsored Tobacco Use Supplement to the Current Population Survey (1992-1996): <http://riskfactor.cancer.gov/studies/tus-cps/>. Data files.

US Department of Commerce, Census Bureau (2001). National Cancer Institute Sponsored Tobacco Use Supplement to the Current Population Survey (1998-1999): <http://riskfactor.cancer.gov/studies/tus-cps/>. Data files

US Department of Commerce, Census Bureau (2004). National Cancer Institute and Centers for Disease Control and Prevention Co-sponsored Tobacco Use Supplement to the Current Population Survey (2001-2002): <http://riskfactor.cancer.gov/studies/tus-cps/>. Data files.

US Department of Commerce, Census Bureau (2006). National Cancer Institute and Centers for Disease Control and Prevention Co-sponsored Tobacco Use Special Cessation Supplement to the Current Population Survey (2003): <http://riskfactor.cancer.gov/studies/tus-cps/>. Data files.

Wright, D.E. and Bray, I. (2003). "A Mixture Model for Rounded Data." The Statistician **52**(1): 3-13.

Wu, L. L., S. P. Martin, et al. (2001). "Comparing data quality of fertility and first sexual intercourse histories." Journal of Human Resources **36**(3): 520-555.

Table 1 Sample statistics

Variables	PSID		CPS-TUS		GSOEP		BHPS	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
<u>Time-varying</u>								
Cessation rate	0.02	(0.15)	0.02	(0.14)	0.02	(0.13)	0.01	(0.10)
Price of cigarettes ^a	2.25	(0.74)	2.08	(0.52)	7.22	(0.80)	3.62	(0.78)
Life events ^b	0.12	(0.33)	0.02	(0.16)			0.02	(0.15)
Health shocks ^b	0.04	(0.19)					0.02	(0.14)
Mental health shocks ^b	0.01	(0.08)						
Age	32.64	(12.03)	34.27	(13.24)	32.48	(12.00)	35.31	(15.66)
Sample period (min/max)	1954	2003	1954	2002	1951	2001	1918	2002
N smokers (Person-year)	85,056		5,194,230		256,864		230,461	
<u>Time-invariant</u>								
Female	0.51	(0.50)	0.47	(0.50)	0.41	(0.49)	0.50	(0.50)
Age when surveyed	46.70	(15.06)	46.68	(16.48)	45.63	(15.54)	48.47	(18.94)
Years smoked	20.85	(13.71)	22.81	(14.83)	21.16	(13.02)	26.98	(17.34)
Ex-smokers	0.50	(0.50)	0.41	(0.49)	0.38	(0.49)	0.31	(0.46)
Fraction heapers (among quitters)	0.53	(0.50)	0.63	(0.48)	0.42	(0.49)	0.55	(0.50)
N smokers (Person)	4,097		236,077		11,820		8,238	
N quitters (Person)	2,035		96,783		4,527		2,554	

Notes:

^aCigarette prices measured as follows:

PSID and CPS-TUS - the average cigarette price in November as measured by Orzechowski and Walker (2005)

GSOEP - the average price per pack (in euros) as reported by the Statistisches Bundesamt publication VI D 46/4-39

BHPS - the real price of a pack of Capstan cigarettes (2008 British pounds) as described in text

^bData on time-varying life events and health shocks vary across the data sets. Life events include people who in a given year retired (PSID), got married (PSID, BHPS), got divorced (PSID), or gave birth (CPS-TUS, PSID). PSID health shock data include having a heart attack, stroke, and being diagnosed with lung disease, heart disease, asthma, cancer, and hypertension. BHPS health data include child birth. PSID mental health shocks identify the age a person first developed emotional problems, learning disabilities, and mental problems.

Table 2 Effects of heaping and sample selection on estimated price effects

Variable	PSID 2003				CPS-TUS			
	Naive	Recent	Coarse	T5	Naive	Recent	Coarse	T5
Cigarette price	0.088*** (0.024)	0.255** (0.123)	-0.025 (0.027)	0.464*** (0.046)	-0.073*** (0.004)	0.563*** (0.004)	-0.274*** (0.005)	0.322*** (0.009)
T5 age				0.561*** (0.061)				0.081*** (0.014)
T5 age*price				-0.108*** (0.023)				-0.022*** (0.006)
T5 time				0.043 (0.122)				1.574*** (0.013)
T5 time*price				-0.018 (0.056)				-0.565*** (0.006)
T5 calendar year				-0.170** (0.074)				0.054*** (0.013)
T5 calendar*price				0.071*** (0.026)				-0.037*** (0.006)
Age at survey				0.048 (0.035)				0.003*** (0.001)
Survey age*price				-0.009*** (0.001)				-0.005*** (0.000)
Implied marginal effect	0.004	0.005	-0.0006	0.105	-0.002	0.010	-0.005	0.027
N	85423	61158	20384	85056	5194230	3931740	1220953	5194230
Sample	All	Recent	5-yr avg.	All	All	Recent	5-yr avg.	All

Table 2 cont.

Variable	GSOEP 2002				BHPS 2002			
	Naive	Recent	Coarse	T5	Naive	Recent	Coarse	T5
Cigarette price	0.088*** (0.008)	1.283*** (0.036)	0.090*** (0.010)	0.232*** (0.002)	0.102*** (0.009)	0.725*** (0.022)	0.012 (0.012)	0.141*** (0.030)
T5 age				0.176 (0.146)				1.046*** (0.080)
T5 age*price				-0.004 (0.004)				-0.190*** (0.021)
T5 time				-0.368** (0.174)				0.570*** (0.117)
T5 time*price				0.007 (0.005)				-0.143*** (0.033)
T5 calendar year				-1.310*** (0.136)				0.125 (0.087)
T5 calendar*price				0.035*** (0.004)				-0.026 (0.023)
Age at survey				0.035*** (0.004)				-0.006** (0.002)
Survey age*price				-0.001*** (0.000)				-0.001** (0.001)
Implied marginal effect	0.0032	0.0431	0.0021	0.0592	0.0039	0.0199	0.0003	0.0075
N	256864	198959	59641	256864	230461	189869	52645	230461
Sample	All	Recent	5-yr avg.	All	All	Recent	5-yr avg.	All

Notes: Robust standard errors in parentheses. Coefficient estimates marked by * significant at 10%; ** significant at 5%; *** significant at 1%. Control variables in each sample include: age, age squared, and sex. Additional variables include: PSID - race (black, other race), years of completed schooling, whether the person experienced a health shock, whether they experienced another life even, state fixed-effects and a linear trend in time (see text for description of health shocks and salient events)

CPS-TUS- for race (black, other race), years of completed schooling, whether a child was born, and state fixed-effects. TUS include data from the 1995, 1996, 19998, 1999, 2001, 2002 and 2003 surveys.

GSOEP - ethnicity (five groups - German is reference group), education of mother and father (school dropout or college graduate relative to secondary school graduate), and whether respondent became widow/widower or divorced in year.

BHPS - race (black, Indian, other race), whether a person gave birth in that year, and whether they got married or began to cohabitate with someone in that year.

The implied marginal effect for the 5-year average sample are divided by five to convert them to one-year implied effects. A set of full results are available on request.

Table 3 Effects of heaping and sample selection on estimated price effects

Variable	PSID 2003			CPS-TUS		
	Naive+	Coarse+	T5+	Naive+	Coarse+	T5+
Cigarette price	0.225** (0.109)	0.385*** (0.057)	0.465*** (0.047)	0.548*** (0.005)	0.336*** (0.015)	0.339*** (0.009)
T5 age	-0.225 (0.144)		0.437*** (0.081)	0.036* (0.021)		0.026 (0.019)
T5 age*price	0.085** (0.038)		-0.095*** (0.028)	-0.007 (0.008)		-0.004 (0.008)
T10 age		-0.010 (0.097)	0.269** (0.118)		-0.248*** (0.018)	0.107*** (0.024)
T10 age*price		0.040 (0.036)	-0.032 (0.043)		-0.043*** (0.009)	-0.036*** (0.011)
T5 time			0.165 (0.162)			0.985*** (0.019)
T5 time*price			-0.063 (0.070)			-0.301*** (0.009)
T10 time		-0.186 (0.132)	-0.342 (0.293)		1.977*** (0.018)	0.938*** (0.023)
T10 time*price		0.077 (0.061)	0.140 (0.127)		-1.105*** (0.009)	-0.411*** (0.011)
T5 calendar year	-0.755*** (0.138)		-0.158 (0.174)	-0.314*** (0.023)		0.003 (0.027)
T5 calendar*price	0.275*** (0.042)		0.058 (0.083)	0.099*** (0.008)		-0.021* (0.013)
T10 calendar year		-0.473*** (0.079)	0.016 (0.185)		-1.120*** (0.019)	0.144*** (0.030)
T10 calendar*price		0.236*** (0.032)	0.006 (0.089)		0.468*** (0.009)	-0.047*** (0.014)
Age at survey		0.066 (0.049)	0.046 (0.035)		0.000 (0.001)	0.002** (0.001)
Survey age*price		-0.014*** (0.001)	-0.009*** (0.001)		-0.006*** (0.000)	-0.005*** (0.000)
Implied marginal effect	0.004	0.019	0.084	0.010	0.017	0.030
N	61158	20384	85056	3931740	1220953	5194230
Sample	Recent	5-yr avg.	All	Recent	5-yr avg.	All

Table 3 continued

Variable	GSOEP 2002			BHPS 2002		
	Naive+	Coarse+	T5+	Naive+	Coarse+	T5+
Cigarette price	0.259*** (0.008)	0.077*** (0.010)	0.040*** (0.007)	0.743*** (0.025)	-0.188*** (0.040)	0.132*** (0.030)
T5 age	0.490 (1.017)		0.095 (0.187)	0.079 (0.372)		0.804*** (0.106)
T5 age*price	-0.012 (0.024)		-0.002 (0.005)	-0.008 (0.074)		-0.145*** (0.028)
T10 age	0.087 (1.415)		0.195 (0.265)	0.729 (0.453)	0.107 (0.101)	0.469*** (0.136)
T10 age*price	-0.0025 (0.034)		-0.0066 (0.007)	-0.136 (0.091)	-0.067** (0.027)	-0.086** (0.037)
T5 time			-0.341 (0.219)			0.490*** (0.148)
T5 time*price			0.008 (0.006)			-0.129*** (0.040)
T10 time		-17.008*** (0.628)	0.421 (0.341)		0.215 (0.175)	-0.069 (0.275)
T10 time*price		0.380*** (0.016)	-0.016* (0.010)		-0.175*** (0.053)	0.043 (0.082)
T5 calendar year			-0.342 (0.209)			0.407* (0.216)
T5 calendar*price			0.007 (0.006)			-0.125** (0.060)
T10 calendar year		17.976*** (0.565)	-1.374*** (0.254)		-0.585*** (0.088)	-0.234 (0.227)
T10 calendar*price		-0.436*** (0.014)	0.039*** (0.007)		0.111*** (0.023)	0.099 (0.063)
Age at survey		0.031*** (0.007)	0.031*** (0.004)		-0.011*** (0.003)	-0.006*** (0.002)
Survey age*price		-0.001*** (0.000)	-0.001*** (0.000)		0.000 (0.001)	-0.001* (0.001)
Implied marginal effect	0.044	0.018	0.054	0.0202	-0.0057	0.0064
N	198959	59641	256864	189869	52645	230461
Sample	Recent	5-yr avg.	All	Recent	5-yr avg.	All

Notes: Robust standard errors in parentheses. Coefficient estimates marked by * significant at 10%; ** significant at 5%; *** significant at 1%. Other control variables same as listed in note to Table 2. A set of full results are available on request.

Table 4 Sample statistics - non employment analysis

Variables	Women		Women w/ed. <12		Men		Men w/ed.<12		
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	
<u>Time-varying</u>									
Not employed	0.19	(0.39)	0.20	(0.40)	0.18	(0.38)	0.17	(0.38)	
Minimum wage	6.49	(0.65)	6.51	(0.66)	6.56	(0.70)	6.54	(0.69)	
Survey month	0.08	(0.28)	0.08	(0.28)	0.08	(0.28)	0.08	(0.28)	
June	0.08	(0.28)	0.08	(0.28)	0.08	(0.28)	0.08	(0.28)	
September	0.09	(0.28)	0.09	(0.28)	0.09	(0.28)	0.09	(0.28)	
December	0.09	(0.28)	0.09	(0.28)	0.09	(0.28)	0.09	(0.28)	
January	0.08	(0.27)	0.08	(0.27)	0.08	(0.27)	0.08	(0.27)	
Age	38.27	(9.95)	38.04	(10.27)	37.92	(9.98)	36.72	(10.10)	
Years completed schooling	11.57	(4.66)	9.21	(4.69)	11.51	(5.01)	8.72	(5.00)	
Number of children	1.72	(1.31)	1.85	(1.35)	1.57	(1.32)	1.59	(1.37)	
Year	1995.06	(6.99)	1995.07	(7.21)	1993.99	(7.48)	1994.29	(7.45)	
N (Person-year)	354331		208819		419749		237494		
<u>Time-invariant</u>									
Age when surveyed	41.72	(11.34)	40.93	(11.60)	42.91	(11.42)	41.21	(11.36)	
Black	0.32	(0.47)	0.36	(0.48)	0.23	(0.42)	0.28	(0.45)	
Hispanic	0.01	(0.11)	0.01	(0.11)	0.01	(0.12)	0.02	(0.13)	
N (Person)	4457		2913		4223		2704		

Notes: Employment/non employment measured from month started first reported job. Sample period restricted to January 1970 to December 2002.

Minimum wage is the higher of the state or federal minimum wage as reported in the *Monthly Labor Review* article "State Labor Legislation Enacted in 19xx" published each January from 1970-2007.

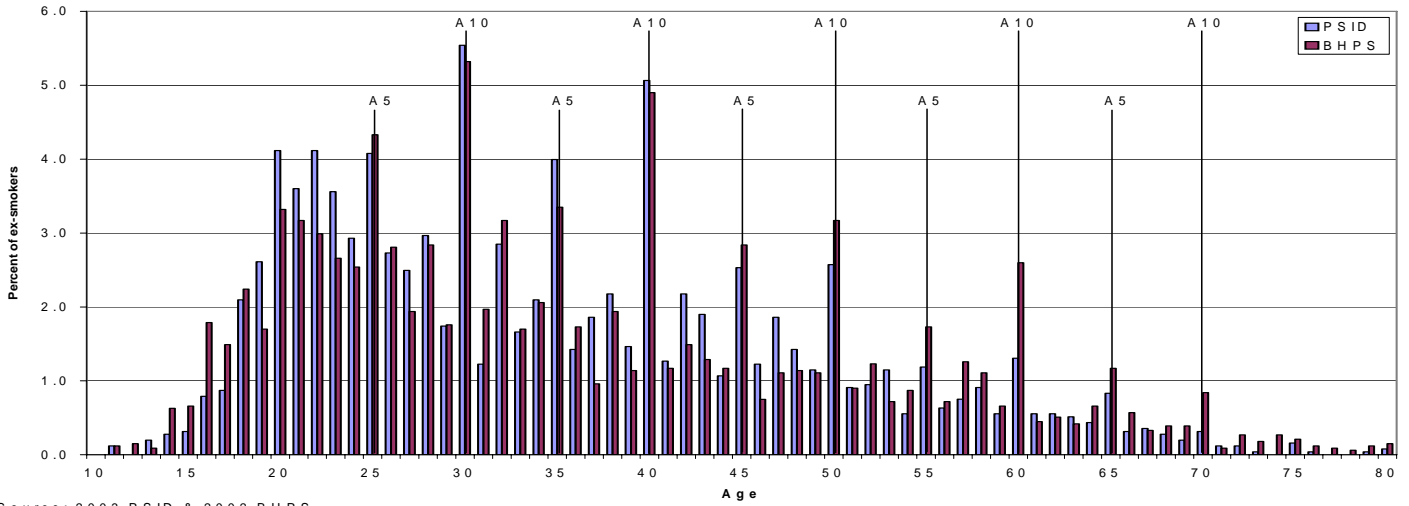
Table 5 Effects of minimum wage on probability of being not employed

Variable	Women				Men			
	Naive	Heaping	Naive	Heaping	Naive	Heaping	Naive	Heaping
Minimum wage	0.039** (0.018)	0.049*** (0.019)	0.053** (0.026)	0.066** (0.026)	0.030 (0.022)	0.038* (0.022)	0.047* (0.025)	0.058** (0.024)
Survey month		-0.051* (0.028)		-0.057* (0.034)		0.014 (0.031)		0.028 (0.040)
January		-0.027*** (0.024)		-0.043 (0.027)		-0.056** (0.022)		-0.049 (0.030)
June		0.138*** (0.050)		0.204*** (0.055)		0.034 (0.026)		-0.011 (0.034)
September		0.231* (0.136)		0.320* (0.171)		0.079 (0.104)		0.004 (0.123)
December		-0.179** (0.090)		-0.246** (0.109)		-0.036 (0.080)		0.020 (0.089)
Minimum wage* Survey month		0.009** (0.004)		0.010** (0.005)		-0.002 (0.005)		-0.002 (0.006)
January		0.003 (0.004)		0.006 (0.004)		0.009*** (0.003)		0.009* (0.005)
June		-0.019** (0.008)		-0.028*** (0.009)		0.000 (0.004)		0.005 (0.005)
September		-0.037* (0.022)		-0.047* (0.028)		-0.006 (0.016)		0.003 (0.020)
December		0.027** (0.013)		0.034* (0.017)		-0.003 (0.011)		-0.009 (0.014)
Implied marginal effect	0.0068	0.0087	0.0092	0.0116	0.0049	0.0062	0.0076	0.0093
As % of not employed	0.0353	0.0451	0.0472	0.0592	0.0281	0.0353	0.0437	0.0536
Percent not employed	0.1930	0.1930	0.1966	0.1966	0.1756	0.1756	0.1742	0.1742
N	354331	354331	208819	208819	419749	419749	237494	237494
Sample	All		Education<=12		All		Education<=12	

Notes: Robust standard errors in parentheses. Coefficient estimates marked by * significant at 10%; ** significant at 5%; *** significant at 1%. Control variables in each sample include: age, age squared, race (black, hispanic), years of education, age when surveyed in 2003, state fixed-effects, and a linear trend for time (year). A set of full results are available on request.

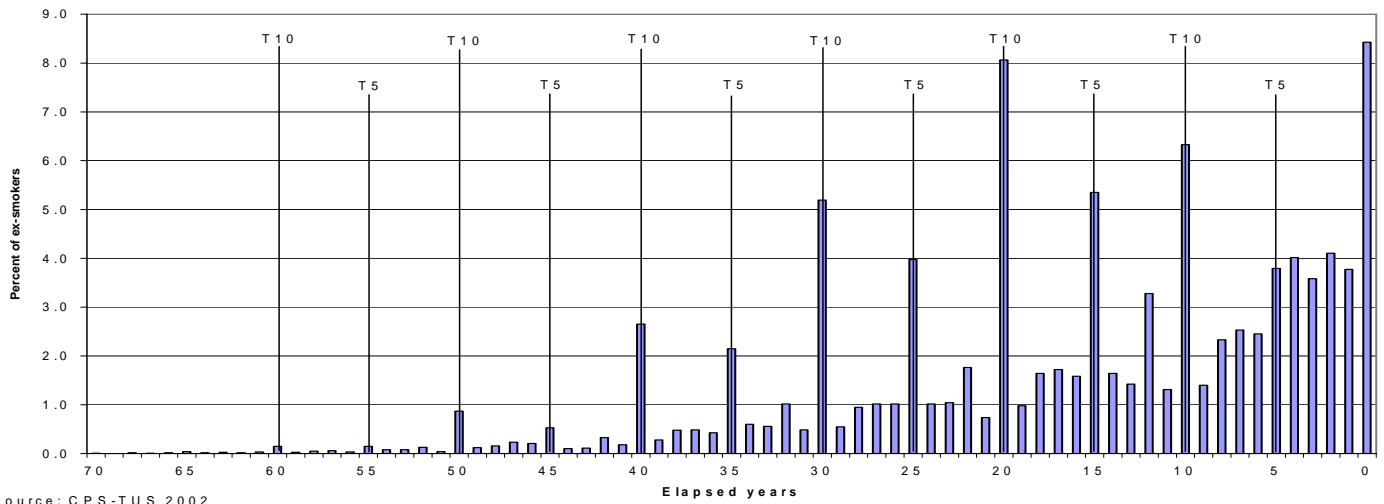
Figure 1 Age, time, and calendar year heaping in measures of time since quit

A. BHPS & PSID: Age last smoked regularly



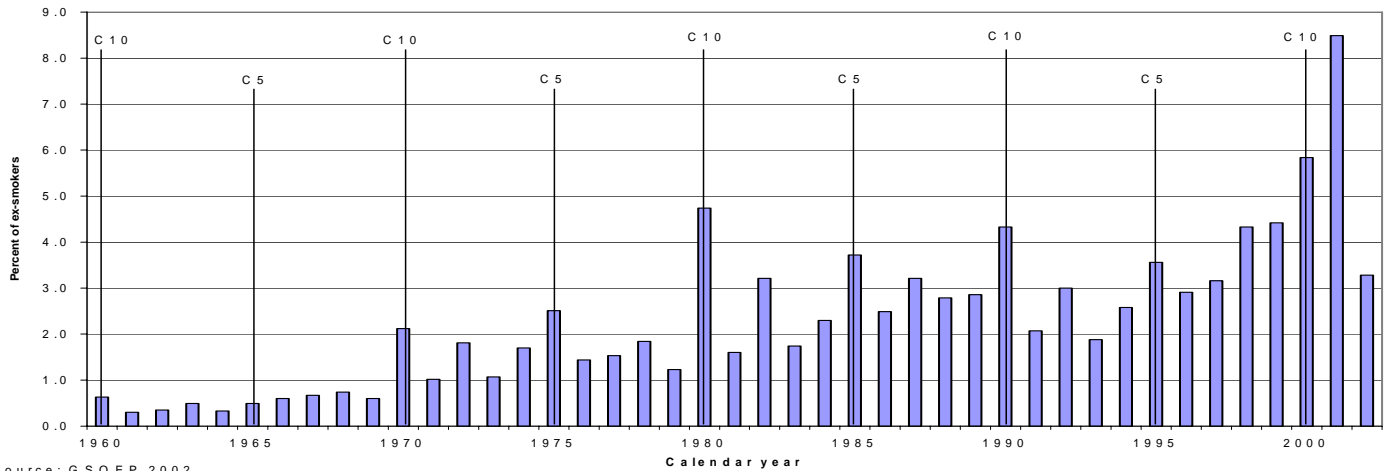
Source: 2003 PSID & 2002 BHPS

B. CPS-TUS Time since last smoked



Source: CPS-TUS 2002

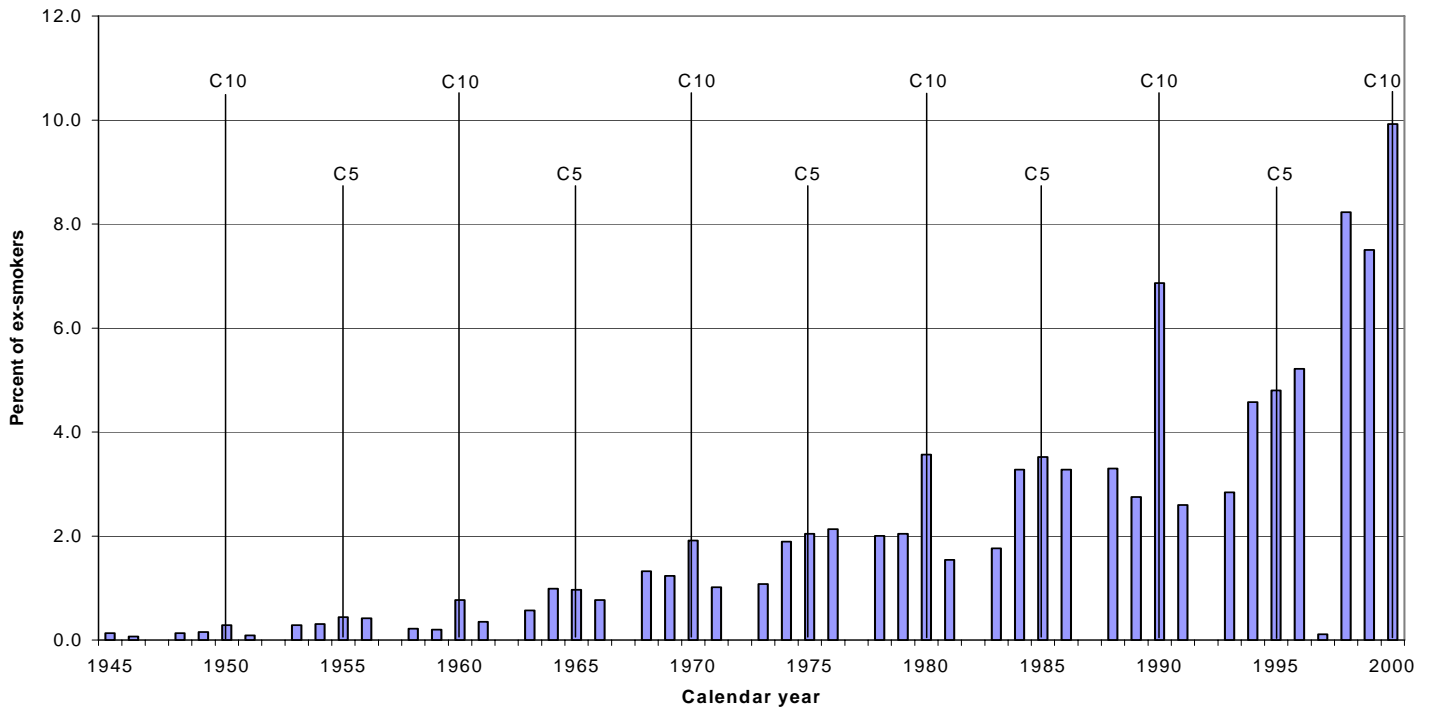
C. GSOEP: Calendar year quit



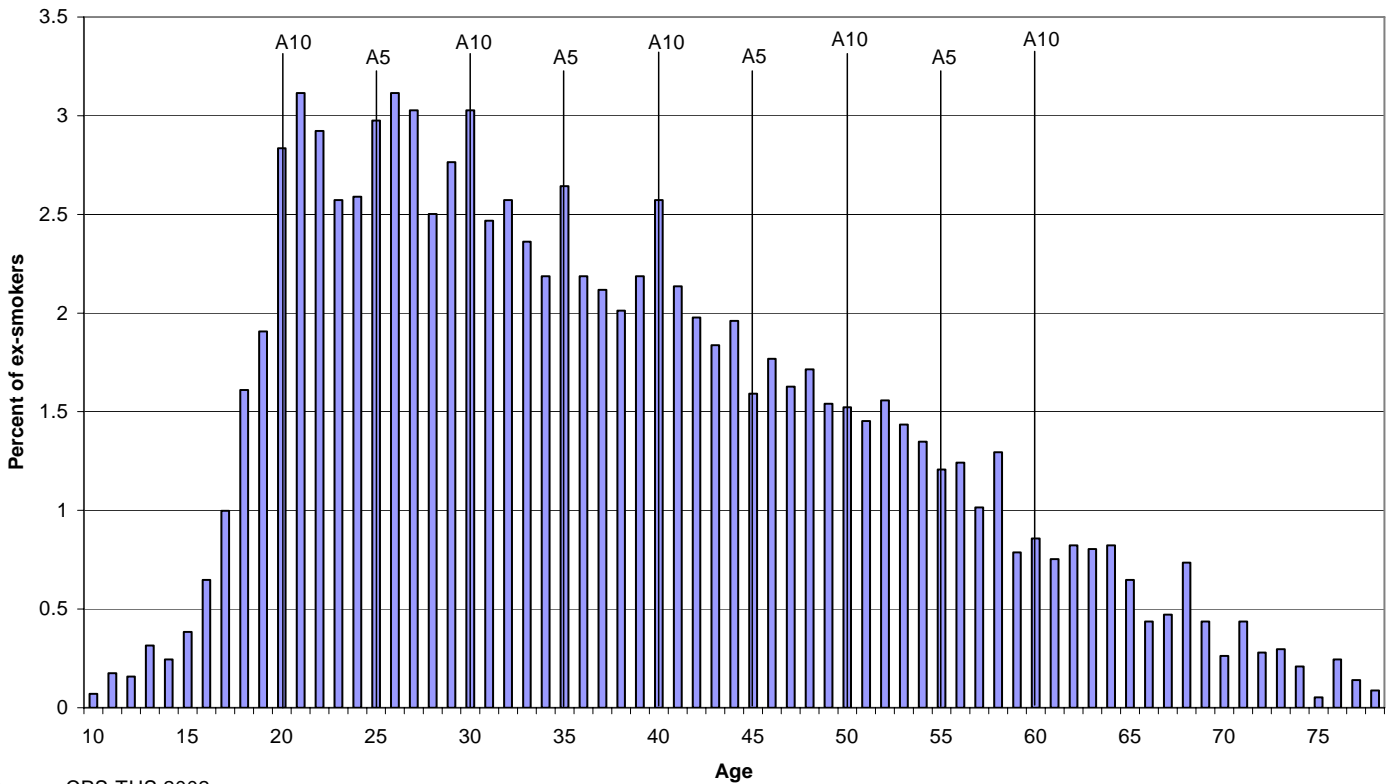
Source: GSOEP 2002

Figure 2 Calendar year heaping and age heaping in marginal distributions of time since quit in CPS-TUS

**A. CPS-TUS Calendar year heaping in distribution of implied calendar year quit
(marginal distribution - minus time and age heapers)**

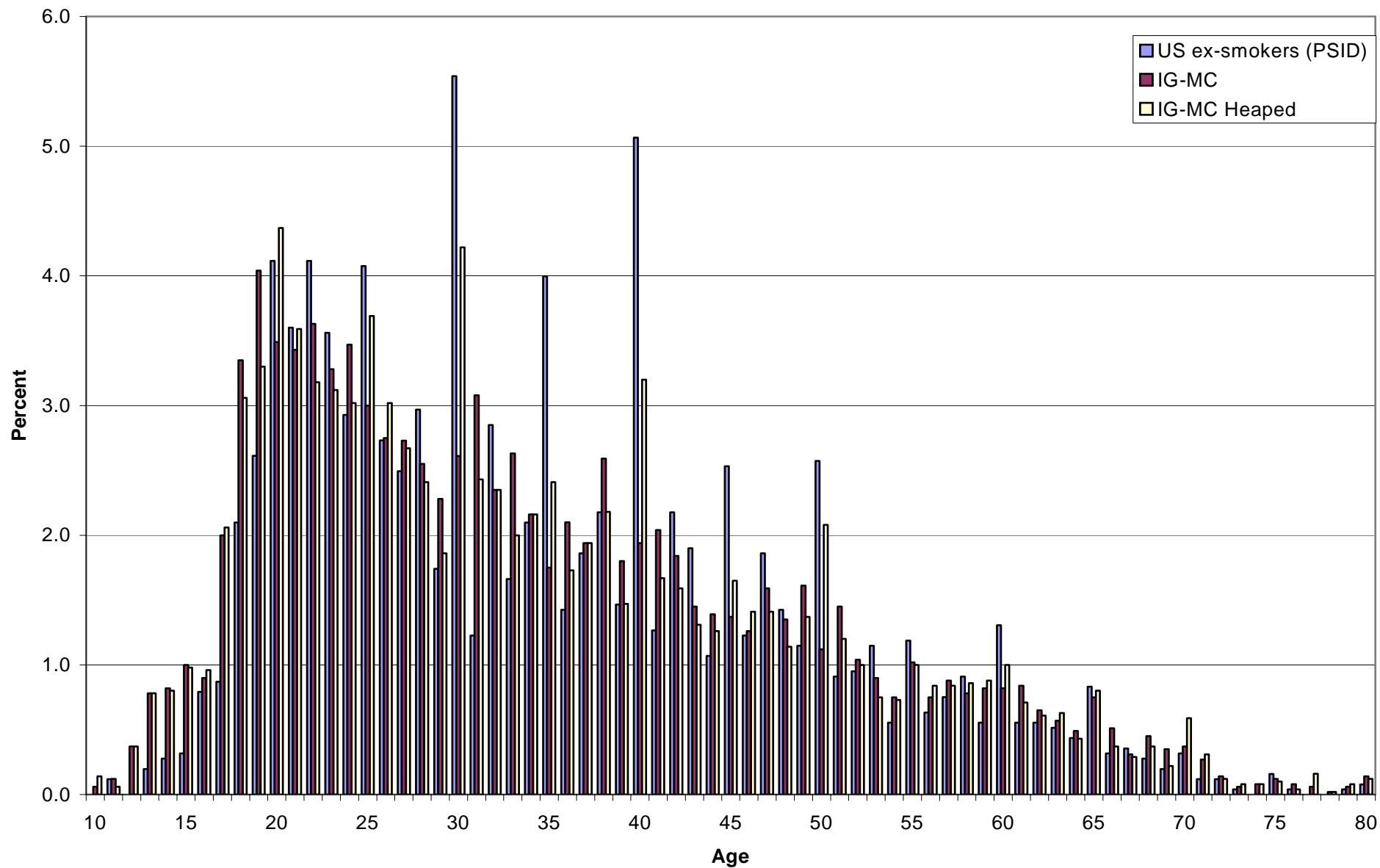


**B. CPS-TUS Age heaping in distribution of implied age quit
(marginal distribution - minus calendar and chronological time heapers)**



Source: CPS-TUS 2002

Figure 3 Distribution of age last smoked regularly - Empirical distribution and Monte Carlo draws from Inverse Gaussian with and without heaping



Source: 2003 PSID and authors' calculations

Figure 4 Hypothetical smoking histories and mismatch bias

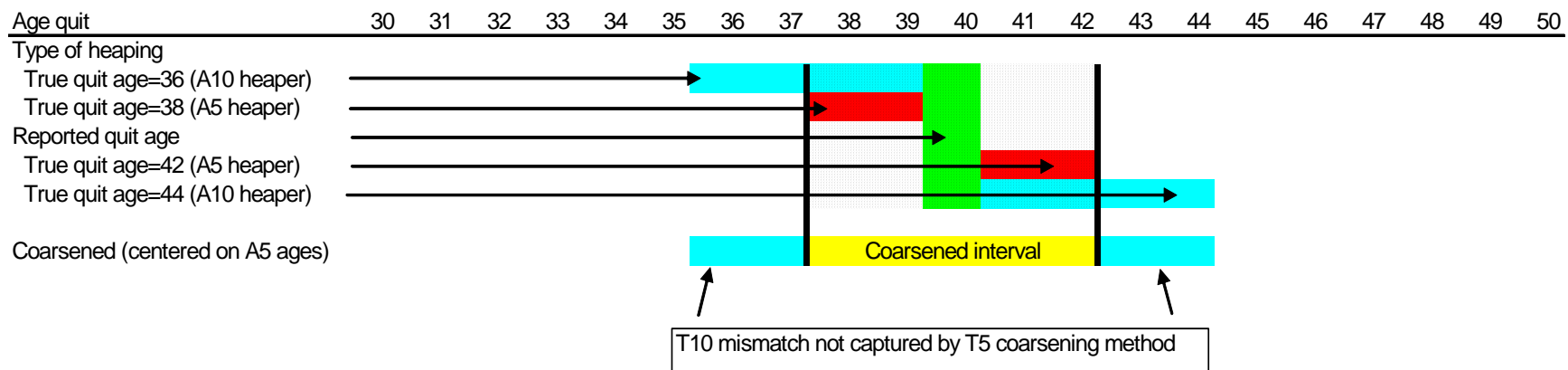
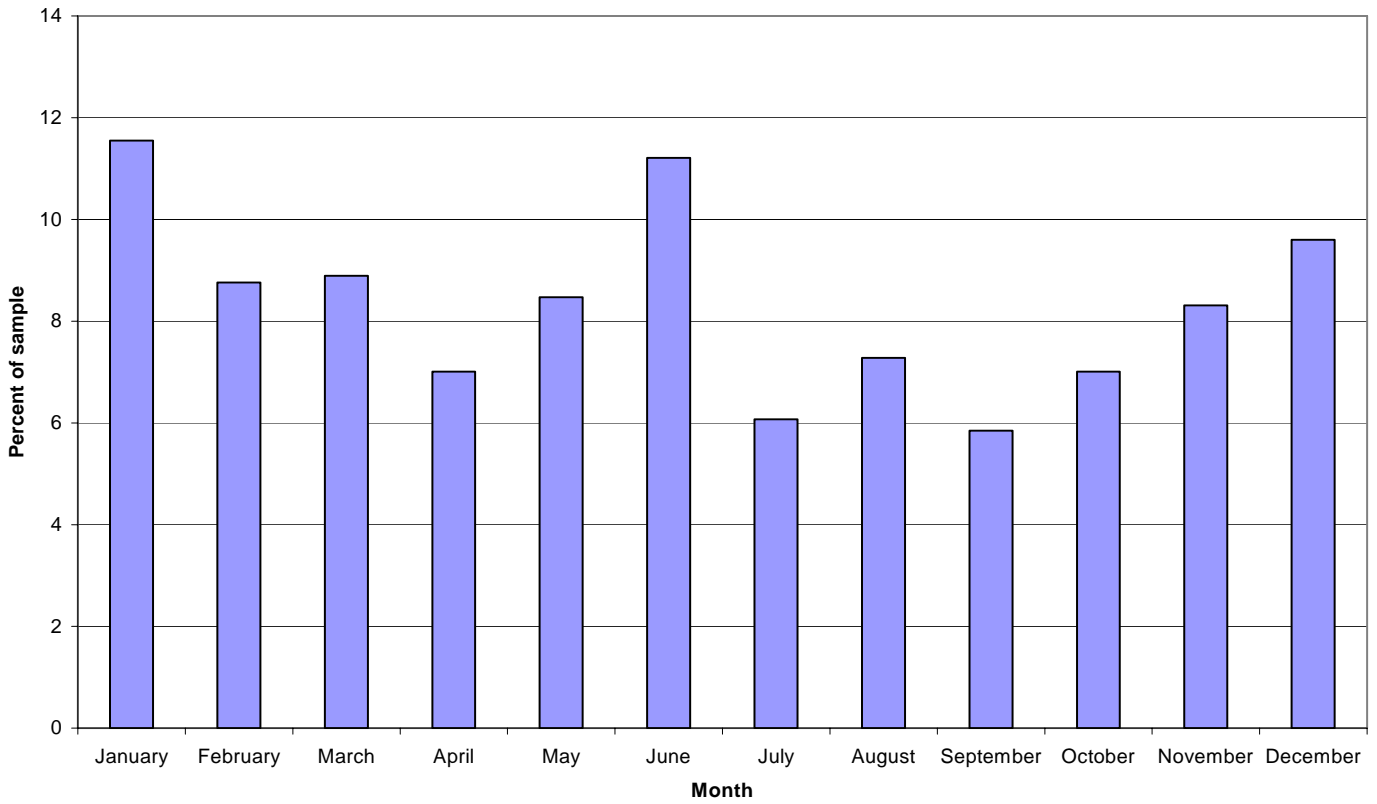
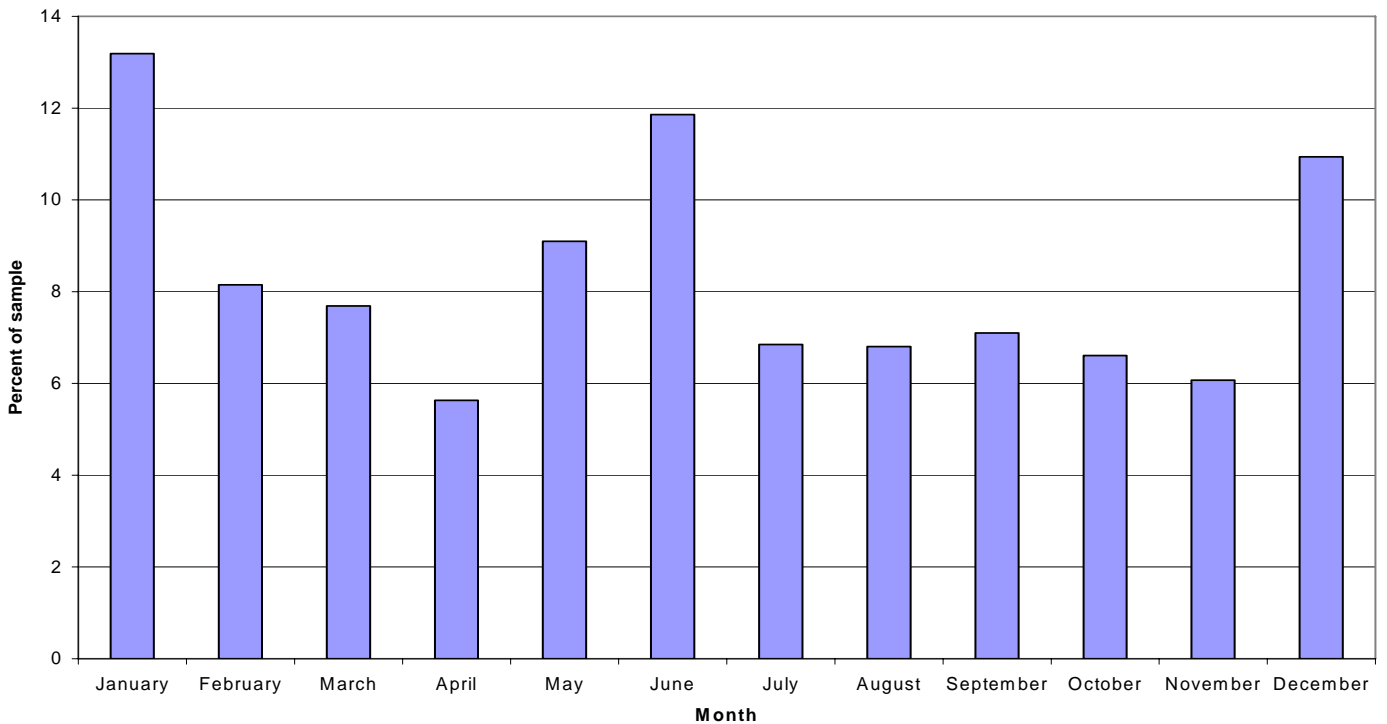


Figure 5. Distribution of month last worked among unemployed men and women

A. Month last worked - unemployed men

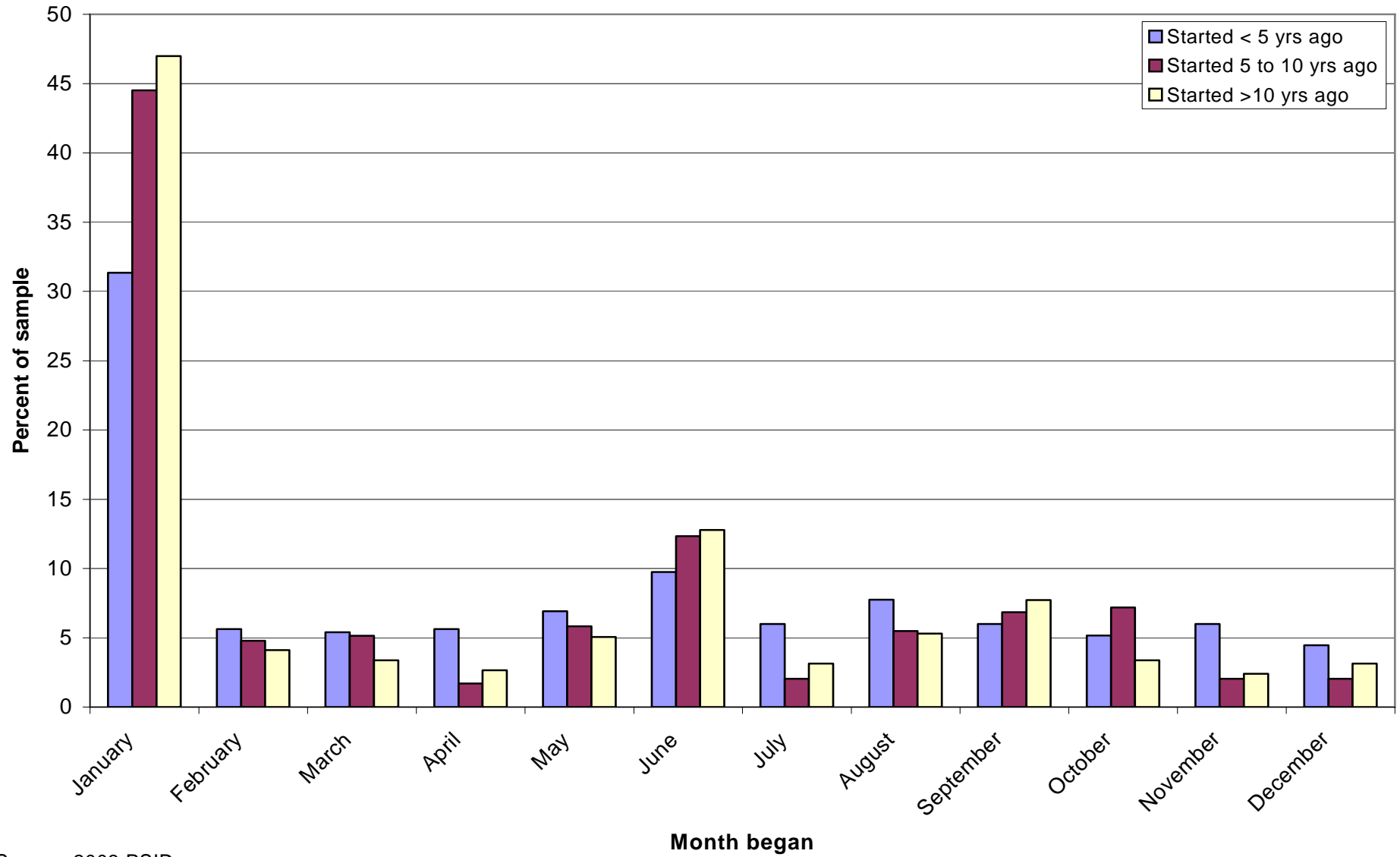


B. Month last worked - unemployed women



Source: 2003 PSID

Figure 6 Month started first job, by years since started (sample surveyed June 2003)



Source: 2003 PSID

Appendix - Data sources

Panel Study of Income Dynamics (www.psidonline.isr.umich.edu)

The PSID began in 1968 with a sample of 5,000 households, representing a disproportionate number of low-income individuals. All current PSID families contain at least one member who was either part of the original 5,000 families or born to a member of one of these families. Although the original sampling scheme disproportionately selected individuals from low-income families, a representative sample of the United States population can be obtained by excluding the original over-sample from the data or by applying sample weights. Starting in 1997 the PSID administers its survey every other year. Only the head and "wife" (a PSID term designating the household member with whom the head has a "significant" relationship) are asked about their cigarette consumption. Retrospective smoking questions were asked in 1986, 1990 (for those age 65+), 1999, 2001, 2003, and 2005. In this paper we only analyze data from the 2003 wave of the PSID. We do so primarily to limit the analysis to data from a single cross-section because the retrospective smoking data in our data sources are from a single cross-section.¹⁵ We analyze the 2003 PSID smoking data because it is closest in calendar time to the data from the UK and Germany. Results from other waves are similar with the exception of the 1986 and 1990 samples.¹⁶ After we restrict the sample to respondents who answered the relevant smoking questions, the sample size for the smoking cessation model is

¹⁵Lillard, Bar, and Wang (2008) analyze separately the role heaping plays when multiple observations on the same smoker are available in longitudinal data.

¹⁶Estimated price coefficients are much larger in the 1990 sample of older PSID sample members. By contrast, we find almost no relationship between price and the probability of a quit in the 1986 sample. In the 1999, 2001, and 2005 waves of the data the coefficient estimates on price are almost identical to the ones we present for 2003.

4,097 person and 85,056 person-age year observations. Just under half of this sample (2,035 people) quit smoking by 2003.

British Household Panel Survey (www.iser.essex.ac.uk/ulsc/bhps/)

The BHPS is an annual survey of each person age 16 or older in a sample of more than 5,000 British households (approximately 10,000 individual interviews). Launched in 1991, the BHPS follows all individuals in the original households as they form their own households. All individuals age 16 and older in the new households are also surveyed. In 1999 all current and former smokers were asked to report the age they first smoked. In 2002 all former smokers were asked the age they last smoked regularly. After the sample is restricted to respondents who answered the relevant questions, the sample size for the smoking cessation model is 8,238 person and 230,461 person-age year observations. Of these, 2,554 smokers quit during the period we observe them.

Current Population Survey - Tobacco Use Supplements (CPS-TUS)

The Tobacco Use Supplements to the Current Population Survey, sponsored by the National Cancer Institute and administered as part of the U.S. Census Bureau's continuing labor force survey, have been collected since 1955 (Haenszel, Shimkin, Miller 1956, Hartman *et al.* 2002). In the more recent CPS-TUS surveys, data on smoking behavior of a large, nationally representative sample of about 240,000 individuals 15 years of age and older is collected in a three-month survey cycle. These cycles were conducted in September 1992, January and May 1993; September 1995, January and May 1996; September 1998, January, and May 1999; and June and November 2001 and February 2002. A separate TUS "Special Topics supplement" was administered in 2003. After the sample is restricted to respondents who answered the relevant questions, the sample size for the smoking cessation model is 236,077 person and

5,194,230 person-age year observations. Across all of these surveys we observe 96,783 CPS-TUS smokers quit.

German Socio-Economic Panel (www.diw.de/english/sop/index.html)

The GSOEP is a longitudinal survey of households begun in 1984. The GSOEP began with a sample of 6,000 households in the Western States of Germany representing a disproportionate number of non-German migrant-workers. The GSOEP attempts to collect information from all household members ages 16 and older. Contemporaneous smoking data was collected in 1998, 1998, and 2001. Retrospective smoking histories were gathered in the 2002 survey. After the sample is restricted to respondents who answered the relevant questions, the sample size for the smoking cessation model is 11,820 person and 256,864 person-age year observations. In this sample, 4,527 smokers quit by 2002.

Other covariates

We also draw data on other demographic characteristics of the individuals in each of the above data sets. The data include both time-invariant and time-varying characteristics. The time-invariant characteristics include sex and race. The type and number of available time-varying characteristics differs across data sources. In each data set we construct as many measures of time-varying characteristics as possible. These time-varying data include highest grade completed, marital status (single, separated, divorced, widowed), marital events (got married, separated, got divorced), number of children, birth of a child, and indicators of shocks to one's health. These health shocks measure the age a person said he was when he had his first heart attack, stroke, or was told he had diabetes, lung cancer, high blood pressure, or heart disease. In general we chose to construct time-varying measures that could be defined before

each survey was first administered so that we could take full advantage of the long history the retrospective data make available.

We construct measures of time-varying covariates in the PSID, CPS-TUS, BHPS, and GSOEP using available information on the age of biological children, on retrospective dates marriages began (or ended), on the age a person was when he suffered health shocks (PSID only), and on educational attainment (PSID only).

For each person in our samples, we coded the birth of a biological child based on their exact date of birth when available and by simple subtraction when only the current age of the child was available. The PSID and BHPS had information on the month and year that up to three marriages started or ended. In the BHPS there was information also on when a person began to live together with a partner. We used that information as well. In the PSID we also used retrospective information on health events. These health shocks measure the age a person said he was when he had his first heart attack, stroke, or was told he had diabetes, lung cancer, high blood pressure, or heart disease. Finally, in the PSID we imputed each person's progression through schooling using contemporaneous information across all years a person participated in the PSID, retrospective information on grades a respondent had repeated in school, the year a person was last enrolled in school, and dates of high school and/or college graduation. For people who began or completed schooling before the PSID started we worked backwards from the year they reported having left school (by dropping out or by graduating). If the person reported having repeated a grade we included that information. Otherwise we assumed a smooth progression through school.

Construction of Capstan cigarette price

In the UK cigarette prices are measured as the real price of one pack (20 cigarettes) of the British brand Capstan (in 2008 British pounds). We constructed a time series of the price of Capstan from 1904 to 2002 using data on the price of a pack of Capstan and on the price of Benson and Hedges Special Filter Golds. These data were provided to the authors by the (British) Tobacco Manufacturers Association (TMA). The TMA compiled these data from Alford (1973), the HMC&E Annual reports, and, from 1997 onwards, their own internal data. The series on the price of Capstan has a gap from April 1975 to October 1997. For the years 1976 to 1978, we substitute the average retail price of 20 standard tipped cigarettes with a constant percentage markup (calculated from the years 1971-1975) because the price of Capstan is higher than the average cigarette price. To impute a price for Capstan for the years 1979 to September 1997 we use a time series for the price of a pack of Benson & Hedges Special Filter Golds (provided by the Tobacco Manufacturers Association). A regression of the price of Capstan on the price of Benson & Hedges Special Filter Golds in the months for which both series are observed yields a coefficient of .988. We substitute the price of Capstan predicted from the price of Benson and Hedges for the years 1979 to 1997.¹⁷

¹⁷In unpublished results, we also use a measure of the price of tobacco products from the UK Inland Revenue Service that runs from 1920 to present. Results from these models are qualitatively similar.