Generalized Measurement Error and Occupational Misclassification*

Alden Porter†

September 20, 2021

Abstract

Recent literature has emphasized the importance of changes in occupation, i.e. occupational mobility, for both personal and aggregate outcomes. Despite the abundant literature on occupational classification error, there is no mathematical formalization for how misclassification impacts occupational mobility estimates. This paper fills that gap by generalizing the classical notion of measurement error in a way that can be applied to changes in discrete classification. In this framework mobility probabilities are ambiguously biased, and I provide theoretical results giving conditions under which that bias can be signed. In this context regressions on the mobility variable will also be ambiguously biased. I apply these results to the Current Population Survey (CPS) and show that misclassification in occupation leads to overestimation of the occupational mobility rate, and rising measurement error leads to a spurious rise in raw occupational mobility estimates from 2005 onward.

*I am grateful to Pascual Restrepo, Adam Guren, Robert G. King, and Johannes Schmieder for their guidance at different stages of this project.

†Boston University, Department of Economics, porteraw@bu.edu

1 Introduction

Occupational choice has been a topic of interest for economists dating back at least to Roy [1951]. Occupation determines one's earnings, lifestyle, healthcare access, et cetera. Indeed, occupation is one of the most important things people choose in their lives. Choice of occupation is also not static, people can and do change their occupation. This naturally raises questions about what causes changes in occupation, and what the implications of those changes are.

Despite there being good reasons to study occupational mobility, practical data concerns often make doing so quite challenging. It is well known that occupations are difficult to observe in practice, and are subject to substantial measurement error. In survey data for example, occupations are determined by surveyor coding of written responses to questions like "what is your primary activity?" or by self selection into occupational categories. In the first case, the text the individual submits must be allowed to change for occupational mobility to occur, however this leaves open the possibility for misinterpretation of changing text and misclassification of the occupation. In the second case, individuals who are "on the fence" may assign themselves to different occupation bins depending on the tasks they are currently performing, even though their overall role remains the fixed.

By way of example, consider an individual who works as a dishwasher. Their regular tasks may involve busing tables, washing dishes and doing prep work. Though their occupation remains unchanged and we expect them to perform the same mix of tasks in the long run, the actual tasks they perform may vary slightly in the short run. The individual may be assigned to do prep work one month, and table bussing the next month. They may think of themselves as a "dishwasher", "prep-cook" or "bus-boy" depending on the recent tasks they've been performing. This short run variation does not reflect genuine changes in the expected future tasks they will perform. Never the less, it may change their responses to questions about their current activity or change the occupational label they assign themselves.

This paper contributes to the literature on occupational mobility by developing a novel measurement error framework. In this framework I consider how measurement error in discrete classifications affects estimates of classification changes. I then apply this framework to the Current Population Survey (CPS), and show that the rise in missing answers found by Fujita et al. [2020] seems to cause occupational mobility to rise spuriously. This framework is necessary because occupational mobility is a discrete variable constructed from other discrete variables. In the classical sense, dating back to Frisch [1934], measurement error is thought of as a continuous white noise term entering linearly into a continuous equation. More recently, econometricians have made strides in understanding different types of discrete measurement error or "classification error". However these models tend to rely on continuous running variable as the source of the underlying measurement error.[1] Occupational mobility, however, is a discrete indicator for changes in a discrete variable. Its study necessitates a corresponding notion of measurement error. To my knowledge this is the first paper to study classification error with this characteristic in mind.

The theoretical results presented in this paper are applicable to more than just occupations. Indeed, they are applicable to any context in which there is a discrete indicator for a change in some discrete classification. One could imagine this framework being applicable to changes in industry, physical region, and education since in most data-sets these are measured with some sort of discrete classification.

I apply this framework to the context of occupations by studying the raw occupational mobility series in the Monthly Current Population Survey. I contrast the raw series with a number of different adjusted series starting in 1994 and continuing into 2020. I find that the unadjusted occupational mobility probability is increasing starting in 2006 and going

[1]see Chen et al. [2007] for an overview and Sullivan [2009] for an application of this to the context of occupations.

into 2020, however this pattern does not appear in a number of adjusted series including a "missing at random" interpretation of dependent coding questions and a series adjusted a la. Moscarini and Thomsson [2008]. I show that the fraction of missing answers to dependent coding questions, questions used to determine genuine changes in occupation, has been rising since 2005.[2].

I rationalize this finding using my measurement error model and show that, as the degree of measurement error increases, it is likely the measured rate of occupational mobility will too. Applied to the context of the CPS, this explains why the raw occupational mobility series is rising while the raw job to job mobility series is declining. Fujita et al. [2020] shows that there are a rise in missing answers to dependent coding questions in the CPS. Missing answers to dependent coding questions trigger independent coding of occupation, which is associated with greater measurement error. My results imply that raw occupational mobility will rise with more missing answers. At the same time, a missing at random assumption for dependent coding questions leads to selection bias which causes job to job mobility to decline. By using alternate filters developed by Moscarini [2005], I show that the degree of the decline in occupational mobility can be significantly attenuated. It thus seems likely that trends in occupational mobility since 2005 that are measured using the CPS are a result of changes in the data collection process, and should be subject to skepticism.

Furthermore, this paper contributes to the literature on occupational mobility by documenting a discrepancy between studies that use different data sources. Studies that rely on survey data may be more prone to error and thus would have upwardly biased estimates of *levels* of occupational mobility, as has been previously documented by Moscarini and Thomsson [2008]. This study further contributes to the literature by showing that *trends* in occupational mobility are also subject to much great noise. This makes it much more diffi-

[2]This is due in part to the introduction of the Respondent Identification Policy as shown by Fujita et al. [2020]

cult to conduct analysis of occupational mobility in countries that don't have occupational information on administrative records, since miscoding of occupation can be frequent and severe in survey data, as documented in Kambourov and Manovskii [2004].

The remainder of the paper is organized as follows: Section 2 reviews the existing literature; section 3 develops the model of measurement error to be used in the paper; section 4 goes over the application of the model of measurement error to the monthly CPS; and section 5 concludes.

2 Literature Review

This paper primarily relates to two existing strands of literature, the first is models of classification error and the second is studies of occupational mobility. With regard to misclassification error Hausman et al. [1998] develops a model of misclassification error a binary response driven by a continuous (and erroneous) running variable, showing that in general probit and logit maximum likelihood estimates will be inconsistent. Lewbel [2007] builds on this finding by showing that this sort of misclassification error in treatment effects can result in attenuation bias, and develops an instrumental variables strategy to obtain identification. Hu [2008] extends the misclassification error model to non-binary discrete random variables and develops an instrumentation procedure that allows for consistent parameter estimates. In the context of occupations, Sullivan [2009] develops a framework in which individual valuations of occupations are treated as a running variable, and applies this framework the National Longitudinal Survey of Youth to estimate the misclassification error in occupations. He finds that around 7% of occupations in the survey are misclassified.

The second strand of literature this paper relates to is general studies of occupational mobility. It seems likely that changes in occupation are important for understanding a wide range of economic phenomena. Topel and Ward [1992] find that changes in employer, which is strongly correlated with changes in occupation, during the start of ones career contribute significantly to wage growth. Furthermore Huckfeldt [2021] finds that changes in occupation contributes significantly to unemployment scarring. Dvorkin [2017] estimate a DSGE model allowing for occupational mobility and find that allowing for occupational mobility is likely important for explaining patterns in wage polarization. This paper contributes to this literature by showing that coefficients in regressions with occupational mobility will, in general, be biased and shows that even in data sets that are "well suited" to studying it like the monthly CPS.

3 A Model of Measurement Error

3.1 Relationship With a Classical Model of Measurement Error

In classical models of measurement error it is generally assumed that the variable of interest x_i^* takes on values in $\mathbb{R}$ and can be written as

$$x_i = x_i^* + \varepsilon_i. \tag{1}$$

Where x_i is the observed value for some index $i \in I$, ${\varepsilon_i}$[3] is a mean zero i.i.d. shock term. When x_i is the dependent variable in a regression coefficients will be unbiased, but standard errors will be biased upwards, when x_i is the independent variable the non-constant regression coefficients will be biased towards zero.

However the underlying variables in occupational mobility are inherently discrete. This necessitates thinking about so called "misclassification error." The literature on misclassification error commonly uses latent variable models to get around this issue. In these models there is some continuous "latent" variable w_i that subject to classical measurement error, and a binary variable $b(w_i)$ which is zero if w_i is less than some constant and 1 if

[3] Note that the index set could contain a person component, a time component, or both.

it is greater than that constant[4]. This approach is taken by a number of papers including Sullivan [2009] and Hausman et al. [1998]. While undoubtedly useful, this approach is not adequate when studying occupational mobility since the underlying variable which causes the mobility indicator to be 1 or 0 is, itself, discrete.

For this reason I opt to use a more abstract notion of measurement error that takes inspiration from the classical case. Notice that by the independence of ε_i for any $j, i \in I$

$$P(x_i^* \leq a, |\varepsilon_j| \leq b) = P(x_i^* \leq a)P(|\varepsilon_j| \leq b) = P(x_i^* \leq a)P(|x_j - x_j^*| \leq b) \tag{2}$$

In other words the distance between the measured and the true value of x at any index j is independent of the true value of x at any index i. Since ε_i is i.i.d. we also have

$$P(|\varepsilon_i| \leq b, |\varepsilon_j| \leq b) = P(|\varepsilon_i| \leq b)P(|\varepsilon_j| \leq b) = P(|x_i - x_i^*| \leq b)P(|x_j - x_j^*| \leq b) \tag{3}$$

For any $i \neq j$.

These properties are easily generalized to more general metric spaces as follows. Let $(\Omega, \mathcal{F}, P)$ be a probability space and $(X, \mathscr{B})$ be a metric space with corresponding borel sets $\mathscr{B}$ and a distance measure $d : X \times X \to \mathbb{R}$. Assume that $x_i : \Omega \to X$ and $x_i^* : \Omega \to X$ are random variables on $(X, \mathscr{B})$ and that x_i^* is the true random variable and x_i is the observed random variable. Assumption 2 generalizes as follows, Let $a, b > 0$ and $B \in \mathscr{B}$ then

$$P(x_i^* \in B, d(x_j, x_j^*) < a) = P(x_i^* \in B)P(d(x_j, x_j^*) < a) \tag{4}$$

Equation 4 states that the distance between the observed and the true variable is independent of the true random variable. Assumption 3 generalizes to:

$$P(d(x_j, x_j^*) < a, d(x_i, x_i^*) < b) = P(d(x_j, x_j^*) < a)P(d(x_i, x_i^*) < b) \tag{5}$$

for $i \neq j$. Equation 5 says that the distances between the observed and true values for any two error draws are independent across the index set.

[4]There are of course extensions to the non binary case, yet the intuition remains the same.

This generalization now makes it easy to define measurement error in the case of occupational mobility. If d is the discrete metric and X is a finite set then for $\varepsilon = \frac{1}{2}$ equation 4 becomes

$$P(x_i^* \in B, x_j = x_j^*) = P(x_i^* \in B)P(x_j = x_j^*) \tag{6}$$

Where $B \subset X$. In other words, the probability distribution on x^* is independent of whether or not the observed value x is equal to the true value of x for any index. Equation 5 becomes:

$$P(x_j = x_j^*, x_i = x_i^*) = P(x_j = x_j^*)P(x_i = x_i^*) \tag{7}$$

In other words, the probability that x is misclassified at j does not affect the probability x is misclassified at i when $i \neq j$. For the remainder of this paper, I will assume that 6 and 7 hold.

3.2 Error In Movements Between Classifications

Suppose there is a population of individuals $i \in I$ where I is some index set, and that time is discrete and denoted by $t \in \mathbb{N}$. Each person has some true class given by k_{it}^* and some observed class given by k_{it}. This class could denote occupation, industry, or any other discrete feature that is time varying. The econometrician is interested in the probability that an individual changes class between two periods, i.e. $P(k_{i,t}^* \neq k_{i,t-1}^*)$, however she can only estimate $P(k_{i,t} \neq k_{i,t-1})$. Note that 6 and 7 imply:

$$P(k_{i,t}^* \neq k_{i,t-1}^*, k_{i,t} = k_{i,t}^*, k_{i,t-1} = k_{i,t-1}^*) = P(k_{i,t}^* \neq k_{i,t-1}^*)P(k_{i,t} = k_{i,t}^*, k_{i,t-1} = k_{i,t-1}^*). \tag{8}$$

Since under those assumptions the three events $k_{i,t}^* \neq k_{i,t-1}^*, k_{i,t} = k_{i,t}^*$ and $k_{i,t-1} = k_{i,t-1}^*$ are independent. Let

$$\eta_{it} = P(k_{i,t} \neq k_{i,t-1}, k_{i,t} \neq k_{i,t}^* \text{ or } k_{i,t-1} \neq k_{i,t-1}^*) \tag{9}$$

$$\theta_{it} = P(k_{i,t} = k_{i,t-1}, k_{i,t} \neq k_{i,t}^* \text{ or } k_{i,t-1} \neq k_{i,t-1}^*) \tag{10}$$

denote the probability than an individual is recorded as moving (η_{it}) or staying (θ_{it}) and their class is misrecorded in t or $t-1$. Finally let $p^*_{i,t} = P(k^*_{i,t} \neq k^*_{i,t-1})$ be the true mobility probability and $p_{i,t} = P(k_{i,t} \neq k_{i,t-1})$ be the observed mobility probability. The following is true:

Proposition 1. *$p^*_{it} < p_{it}$ if and only if $p^*_{it} < \frac{\eta_{i,t}}{\eta_{i,t}+\theta_{i,t}} = P(k_{i,t} \neq k_{i,t-1} | k_{i,t} \neq k^*_{i,t}$ or $k_{i,t-1} \neq k^*_{i,t-1})$.*

Proof. By the law of total probability:

$$
\begin{aligned}
p_{it} =& P(k_{i,t} \neq k_{i,t-1} | k_{i,t} = k^*_{i,t} \text{ and } k_{i,t-1} = k^*_{i,t-1}) P(k_{i,t} = k^*_{i,t} \text{ and } k_{i,t-1} = k^*_{i,t-1}) + \\
& P(k_{i,t} \neq k_{i,t-1}, k_{i,t} \neq k^*_{i,t} \text{ or } k_{i,t-1} \neq k^*_{i,t-1}) \\
=& P(k^*_{i,t} \neq k^*_{i,t-1} | k_{i,t} = k^*_{i,t} \text{ and } k_{i,t-1} = k^*_{i,t-1}) P(k_{i,t} = k^*_{i,t} \text{ and } k_{i,t-1} = k^*_{i,t-1}) + \\
& P(k_{i,t} \neq k_{i,t-1}, k_{i,t} \neq k^*_{i,t} \text{ or } k_{i,t-1} \neq k^*_{i,t-1})
\end{aligned}
$$

Equation 8 implies $P(k^*_{i,t} \neq k^*_{i,t-1} | k_{i,t} = k^*_{i,t}$ and $k_{i,t-1} = k^*_{i,t-1}) = p^*_{it}$, so we can write

$$
p_{it} = (1 - \theta_{i,t} - \eta_{i,t}) p^*_{it} + \eta_{i,t} \tag{11}
$$

By the fact that $\theta_{i,t} + \eta_{i,t} = P(k_{i,t} \neq k^*_{i,t}$ or $k_{i,t-1} \neq k^*_{i,t-1})$. But then

$$
\begin{aligned}
& p^*_{it} < \frac{\eta_{i,t}}{\eta_{i,t} + \theta_{i,t}} \\
& \Leftrightarrow (\eta_{i,t} + \theta_{i,t}) p^*_{it} < \eta_{i,t} \\
& \Leftrightarrow (1 - (1 - \theta_{i,t} - \eta_{i,t})) p^*_{it} < \eta_{i,t}. \\
& \Leftrightarrow p^*_{it} < (1 - \theta_{i,t} - \eta_{i,t}) p^*_{it} + \eta_{i,t} = p_{it}
\end{aligned}
$$

□

Proposition 1 states that the true probability of changing class p^*_{it} is less than the observed probability if and only if the true probability is less than the probability of observing a move conditional on there being an error. Proposition 1 implies that it is theoretically possible for the mobility rate to be understated, but only if the value of $\theta_{i,t}$ is high, e.g. the probability of recording someone as non-mobile given their occupation is misreported is high.

In the case of occupations this is precisely the problem that M&T try to address, the point in time probability estimates for the mobility series are going to be systematically overstated because of measurement error. By looking at the levels of the month on month mobility series in the CPS it seems very likely that $p^*_{it} < P(k_{i,t} \neq k_{i,t-1} | k_{i,t} \neq k^*_{i,t} \text{ or } k_{i,t-1} \neq k^*_{i,t-1})$. Following M&T and assuming that errors are more likely for the population of "suspicious movers" we can take $P(k_{i,t} \neq k_{i,t-1} | \text{Suspicious})$ as a rough estimate for $P(k_{i,t} \neq k_{i,t-1} | k_{i,t} \neq k^*_{i,t} \text{ or } k_{i,t-1} \neq k^*_{i,t-1})$. As this is an order of magnitude larger than p_{it}, about .4 compared to .03, it seems more likely than not the conditions for proposition 1 hold. Seen another way, if the true mobility probability were higher than the mobility probability of suspicious movers, then workers would switch occupation more than $12 \times .4 = 4.8$ times per year[5]. Thus, the probability levels of the mobility series in the CPS appear to be overstated.

I now analyze what happens to the observed mobility probability if there is a rise in occupational miscoding. Let $\zeta_{i,t} = P(k_{i,t} \neq k^*_{i,t} \text{ or } k_{i,t-1} \neq k^*_{i,t-1})$ denote the probability of the occupation being incorrectly coded in t or $t-1$. Note that $\zeta_{i,t} = \eta_{i,t} + \theta_{i,t}$, we can write equation 11 as

$$p_{i,t} = p^*_{i,t} - \theta_{i,t} p^*_{i,t} + (1 - p^*_{i,t})\eta_{i,t} = p^*_{i,t} + \zeta_{i,t}[(1 - p^*_{i,t}) \frac{\eta_{i,t}}{\eta_{i,t} + \theta_{i,t}} - p^*_{i,t} \frac{\theta_{i,t}}{\eta_{i,t} + \theta_{i,t}}]. \tag{12}$$

Hence a rise in the probability of miscoding $\zeta_{i,t}$ will cause a rise in the observed probability exactly when

$$(1 - p^*_{i,t}) \frac{\eta_{i,t}}{\eta_{i,t} + \theta_{i,t}} > p^*_{i,t} \frac{\theta_{i,t}}{\eta_{i,t} + \theta_{i,t}} \Leftrightarrow \frac{\eta_{i,t}}{\theta_{i,t}} > \frac{p^*_{i,t}}{1 - p^*_{i,t}}$$

Seen this way, rise in miscoding causes a rise in observed mobility probabilities if the mobility

[5]Suspicious movers are defined a la M&T, and are individuals with blank answers to dependent coding questions. See section 4 for details.

probability under miscoding ($\eta_{i,t}$) is relatively higher than the true mobility probability ($p^*_{i,t}$) when compared to the respective staying probabilities. In the case of occupations in the CPS the mobility probabilities with blank answers are so much higher than the regular point estimates this is almost certainly the case.

3.3 Bias In Regression Coefficients

This section explores the potential for bias in regression coefficients under classification error. Let X be a set of regressors, p be the corresponding vector of observed mobility probabilities and p^* be the vector of true mobility probabilities. Consider the linear probability model given by a regression of X on p of the form:

$$p = \beta X + \varepsilon \tag{13}$$

Where ε is an i.i.d. error term. We can write the estimated coefficient β as

$$\beta = (X'X)^{-1}X'p = (X'X)^{-1}X'(1-\zeta) \odot p^* + (X'X)^{-1}X'\eta \tag{14}$$

Where ζ is the vector of $\zeta_{i,t}$, η is the vector of $\eta_{i,t}$ and $\odot$ is the hadamard product. Note that under our current assumptions the error probabilities ζ and η are in general correlated with X. The regression will in general be biased and it is not possible to determine the direction of the bias. To see this practically, suppose one hopes to understand the affect of task distance on mobility as Gathmann and Schönberg [2010] hope to do. If "close" occupations are more likely to be confused for one another by respondents or survey takers, occupational mobility will be spuriously higher for close occupations. This implies the impact of task distance on occupational mobility will be *negatively* biased rather than biased towards zero. To see the importance of this distinction, note that regressing inverse distance on mobility probabilities will be *positively* biased as a result of this.

4 Application to the Monthly CPS

In this section I show how the theory described above fits into occupational mobility in the monthly CPS. I find that there is a large increase post 2006 in the probability of an occupation switch. I also document a large increase in "suspicious" observations, i.e. observations with missing answers to dependent coding questions, during that same period. Decomposing the mobility series by whether or not the observation is suspicious, I find that the post-2006 increase is driven entirely by rising suspicious observations.

4.1 Data Description

The data I use for the analysis below is monthly CPS panel data from 1996 to 2017 which was retrieved from the Center for Advancement of and Research in Economics at the Kansas City Fed. I use this data source because it has a out of the box personal identifier[6] and also because it has all three dependent occupational coding variables used to identify suspicious observations in M&T 2008[7]. To keep the analysis consistent with M&T, the sample consists of the first four months of observations of men aged 17-64 inclusive who I can match over time.[8]

I use post-1994 data because, after 1994, the CPS implemented a "dependent coding procedure". Under this new procedure occupations were recorded for an individual in their first month of sampling and then in follow up samples they were asked a series of questions to determine if a change in occupation was likely. If the answer to a dependent coding question

[6]The personal identifier I use in this analysis is called kc_pid in the data set, which I validate based on age, sex and race.

[7]I also performed this analysis with data from IPUMS to get nearly identical results, I elected to use the Kansas city data because IPUMS is missing the second dependent coding question

[8]For the actual replicaton of M&T I impose the additional restriction that individuals should be employed for at least two consecutive months. Including or removing this restriction does not affect my results.

indicated a change or was left blank, the respondent were asked for their new occupation. This redesign significantly reduced the number of spurious occupational transitions in the CPS, and focusing on the period after the redesign allows for results that are more easily comparable over time. I focus my attention on the period after 1996 specifically because, prior to that, there are large periods of missing data. The dependent coding questions used to determine a likely change in occupation are as follows:

1. Last month, it was reported that you worked for (employer's name). Do you still work for (employer's name) (at your main job)?
2. Have the usual activities and duties of your job changed since last month?
3. Last month you were reported as (a/an) (occupation) and your usual activities were (description). Is this an accurate description of your current job?

A well known issue with studying occupational mobility in the CPS is that the occupational classification system changes every 10 years. There are thus two changes to the occupational system in my sample, once going into 2003 and once going into 2011. I follow the literature and drop observations in a 2 month window around the change in order to prevent spurious spikes in the mobility series.[9] I also run my results making occupational classification consistent over time using the *occ1990dd* occupational coding system from Autor and Dorn [2013] manually updated to include the 2010 census codes. However imposing this coding scheme does not affect the results.

4.2 Methodology

I define two variables of interest. The first variable is an indicator for whether or not a person's primary occupation changed between two months of their participation in the

[9] I also drop observations from June 2015 as there is a large unexplained spike in the data on this date.

survey[10] denoted MOB, and formally defined as

$$MOB_{i,t} = \begin{cases} 0 & \text{if } k_{i,t-1} = k_{i,t} \\ 1 & \text{if } k_{i,t-1} \neq k_{i,t}. \end{cases}$$

where $k_{i,t}$ is person i's occupation at time t. The second variable is an indicator for whether or not the dependent coding question had a blank answer when it should not have, which is called a suspicious observation. Formally I denote this with the indicator variable $suspicious_{i,t}$. $suspicious_{i,t}$ takes on a value of 1 if the answer to the first coding question is blank; the answer to the second question is blank and the answer to the first question is "yes"; or the answer to the third question is blank, the answer to the first question is yes, and the second is no. The indicator is zero otherwise.

For comparability with M&T 2008, and to analyze their method's effectiveness, I replicate their procedure on my sample. In brief, the procedure sets $MOB_{i,t} = 0$ if there is a suspicious transition and no change in: industry, whether or not the person looked for work in the past 4 weeks, or what class of worker they are. This procedure also sets to zero any suspicious observation which had an unusual pattern of occupational changes[11]. For my analysis I plot the point in time mobility probability estimates ($p_{i,t}$) using the different cleaning procedures I have described and compare the results.

4.3 Results

I first plot the raw estimates of $P(MOB_{i,t} = 1)$ and $P(suspicious_{i,t} = 1)$ in figure 1. The raw mobility series already has an implausibly high level going up to the mid 2000s. Taken

[10] When replicating M&T 2008 I only look at transitions between months 2 and 3 because they use the "trajectory" of occupations to try tease out which suspicious transitions will represent a true change in occupation. My main results use .

[11] For details see M&T 2008, I follow their post 1994 procedure setting $MOB_{i,t}$ for flags 3,10,11,12 and 13 to be zero.

at face value this would suggest that, in a year, the odds an individual stays in the same occupation is around 30%.[12] The series then increases dramatically, almost doubling between 2005 and 2010. This rise appears to be matched by an increase in the frequency of blank answers to dependent coding questions post 2006. In light of the theoretical results above, it seems very unlikely the increase in occupational mobility reflects a genuine economic shift.

Figures 3 and 4 decompose the raw mobility series into mobility probability conditional on suspicious and non-suspicious observations and applies a 12 month moving average. One can see immediately see how a rise in suspicious observations mechanically increases the mobility probability estimate. Note that the law of total probability gives

$$\begin{aligned} P(MOB_{i,t} = 1) = P(MOB_{i,t} = 1 | suspicious_{i,t} = 1) P(suspicious_{i,t} = 1) \\ + P(MOB_{i,t} = 1 | suspicious_{i,t} = 0)[1 - P(suspicious_{i,t} = 1)]. \end{aligned}$$

Since the level of the two conditional series is so vastly different[13], small changes in the weight ($P(suspicious_{i,t} = 1)$) cause large movements in the raw mobility series. This directly relates to Fujita et al. [2020] who find that there was a large rise in missing answers to the first dependent coding question following the introduction of the Respondent Identification Policy (RIP) in 2008. This policy allows individuals to opt out of sharing their answers with household members in subsequent surveys. In particular they can opt out of sharing their employer name, which automatically generates blank responses to the "same job" dependent coding question if they personally are not around to complete the survey in subsequent months.

The introduction of RIP is undoubtedly part of the story as one can plainly see a sharpening of the rise in blank answers starting in 2008. However it appears that the upward trend in blank answers starts prior to this (a fact which the authors discuss in their paper)

[12] $(1 - .03)^{12} = .306$

[13] $P(MOB_{i,t} = 1 | suspicious_{i,t} = 1) \approx .3$ and $P(MOB_{i,t} = 1 | suspicious_{i,t} = 0) = .02$

hence there seems to be another source of increased measurement error that is affecting the CPS in this period.

It is certainly tempting to simply drop observations with blank answers as the level of that series (being around .02) is far more reasonable. However, this approach may result in a series that artificially declines due to selection bias. There is no inherent reason to believe that suspicious observations are selected in the same way as non-suspicious observations, furthermore the nature of that selection bias may change over time.[14] Therefore alternative cleaning procedures should be used to determine the true trend of the occupational mobility series.

One such alternative procedure is implemented by Moscarini [2005].[15] Figure 2 plots my estimate of the month to month mobility probability using their series extended into 2017. Both the upward and downward trends see in figures 1 and 3 following the mid-2000s are attenuated to non-existent in this series. This result implies that much of the observed change in occupational mobility during this time period is spurious, and a result of rising measurement error.

5 Conclusion

Models of discrete choice are becoming more popular in economics, and as they do understanding the pitfalls associated with discrete choice statistics becomes more important. This paper contributes to our understanding of discrete choice modelling by demonstrating and analyzing a form of potential bias that occurs in real world discrete choice settings. I have shown general conditions under which estimates of mobility between discrete categories will be biased, and shown what direction that bias is likely to take. I have also shown that

[14]See Fujita et al. [2020] for further discussion.

[15]see 4.2 for a description

changes in measurement errors can cause substantial problems in this setting, and applied my theoretical framework to the Monthly CPS to show how these sorts of errors can manifest themselves in the real world. I have provided evidence that increases in measurement error led to a spurious rise in estimates of occupational mobility, and provided examples of existing techniques that could be used to address this issue. Future work could apply my framework by directly estimating error probabilities, and constructing counterfactual series based on this.

Bibliography

David H. Autor and David Dorn. The growth of low-skill service jobs and the polarization of the us labor market. *The American Economic Review*, 103(5):1553–1597, 2013. ISSN 00028282. URL `http://www.jstor.org/stable/42920623`.

Xiaohong Chen, Han Hong, and Denis Nekiplov. Measurement error models. Technical report, Stanford University, 2007.

Maximiliano Dvorkin. Skills, Occupations, and the Allocation of Talent over the Business Cycle. 2017 Meeting Papers 1527, Society for Economic Dynamics, 2 2017. URL `https://ideas.repec.org/p/red/sed017/1527.html`.

Ragnar Frisch. Statistical confluence study, 1934.

Shigeru Fujita, Giuseppe Moscarini, and Fabien Postel-Vinay. Measuring employer-to-employer reallocation. Technical report, Yale University, Department of Economics, July 2020.

Christina Gathmann and Uta Schönberg. How general is human capital? a task-based approach. *Journal of Labor Economics*, 28(1):1–49, 2010. ISSN 0734306X, 15375307. URL `http://www.jstor.org/stable/10.1086/649786`.

J.A. Hausman, Jason Abrevaya, and F.M. Scott-Morton. Misclassification of the dependent variable in a discrete-response setting. *Journal of Econometrics*, 87(2):239 – 269, 1998. ISSN 0304-4076. doi: https://doi.org/10.1016/S0304-4076(98)00015-3. URL `http://www.sciencedirect.com/science/article/pii/S0304407698000153`.

Yingyao Hu. Identification and estimation of nonlinear models with misclassification error using instrumental variables: A general solution. *Journal of Econometrics*, 144:27–61, 05 2008. doi: 10.1016/j.jeconom.2007.12.001.

Christopher Huckfeldt. Understanding the scarring effects of recessions. *Working Paper*, March 2021.

Gueorgui Kambourov and Iourii Manovskii. A cautionary note on using (march) cps and psid data to study worker mobility. Technical report, National Bureau of Economic Research, 2004. URL `http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.364.960&rep=rep1&type=pdf`.

Arthur Lewbel. Estimation of average treatment effects with misclassification. *Econometrica*, 75(2):537–551, 2007. doi: 10.1111/j.1468-0262.2006.00756.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0262.2006.00756.x`.

Giuseppe Moscarini. Job matching and the wage distribution. *Econometrica*, 73(2):481–516, 2005. doi: 10.1111/j.1468-0262.2005.00586.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0262.2005.00586.x`.

Giuseppe Moscarini and Kaj Thomsson. Occupational and job mobility in the us. *The Scandinavian Journal of Economics*, 109(4):807–836, 2008. doi: 10.1111/j.1467-9442.2007.00510.x. URL `https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9442.2007.00510.x`.

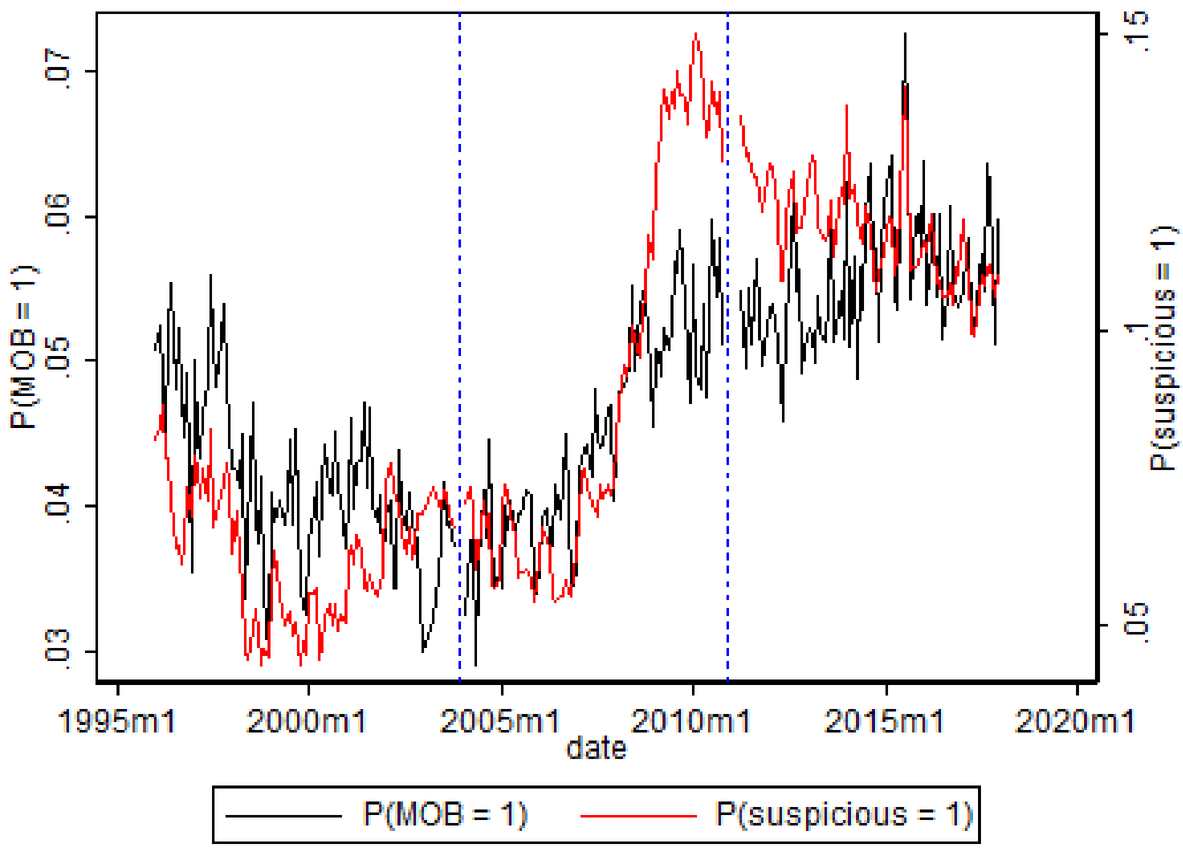
A. D. Roy. Some thoughts on the distribution of earnings. *Oxford Economic Papers*, 3(2):135–146, 1951. URL `https://EconPapers.repec.org/RePEc:oup:oxecpp:v:3:y:1951:i:2:p:135-146`.

Paul Sullivan. Estimation of an occupational choice model when occupations are misclassified. *The Journal of Human Resources*, 44(2):495–535, 2009. ISSN 0022166X. URL `http://www.jstor.org/stable/20648906`.

Robert H. Topel and Michael P. Ward. Job mobility and the careers of young men. *The*

Quarterly Journal of Economics, 107(2):439–479, 1992. ISSN 00335533, 15314650. URL http://www.jstor.org/stable/2118478.

Figure 1: Mobility Probability Over Time, Uncleaned



Estimates computed from the monthly CPS. The red line and right-hand axis correspond to probability an individual is flagged as having a suspicious answer to a dependent coding question. The black line is the occupational mobility rate. The vertical blue lines correspond to years with occupation coding changes.

Figure 2: Mobility Probability Over Time, Cleaned According to M&T 2008

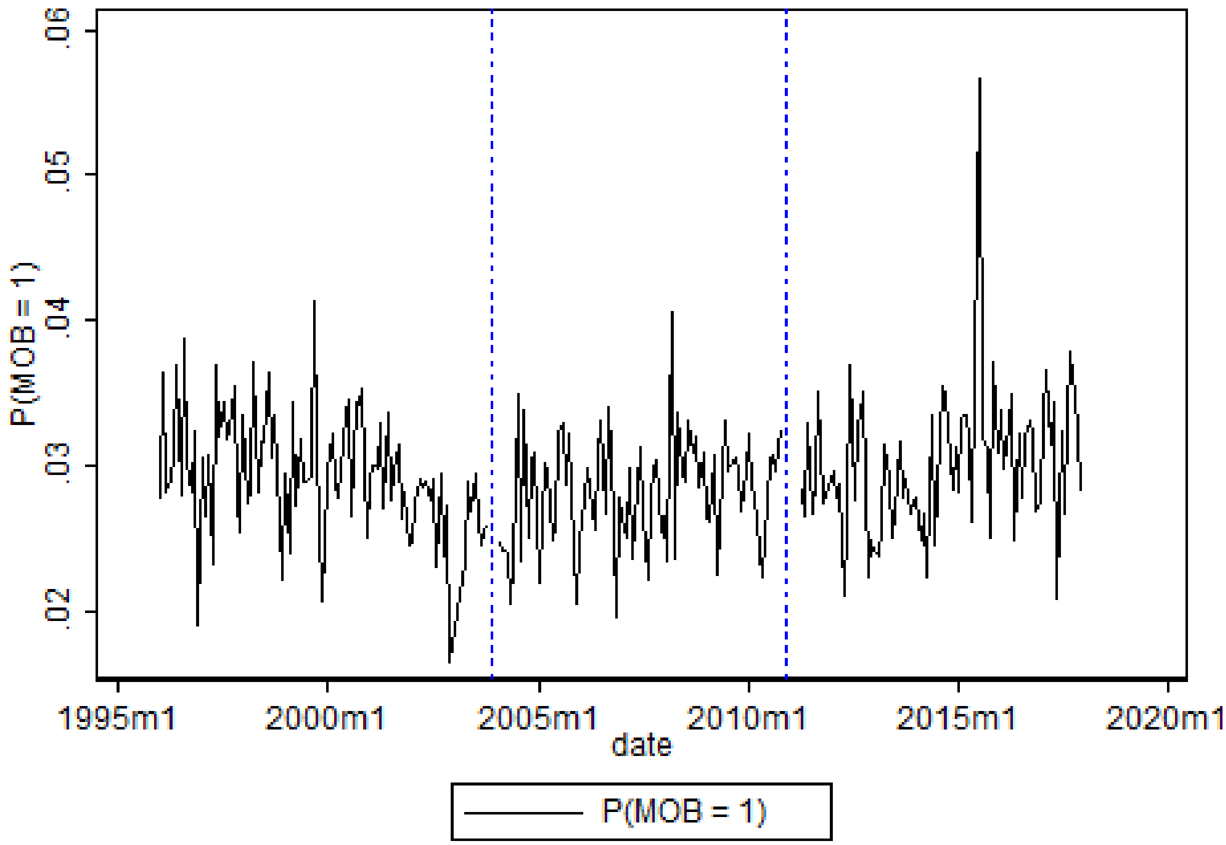


Figure 3: Probability MOB = 1 Given Suspicious = 0, 12 Month Moving Average

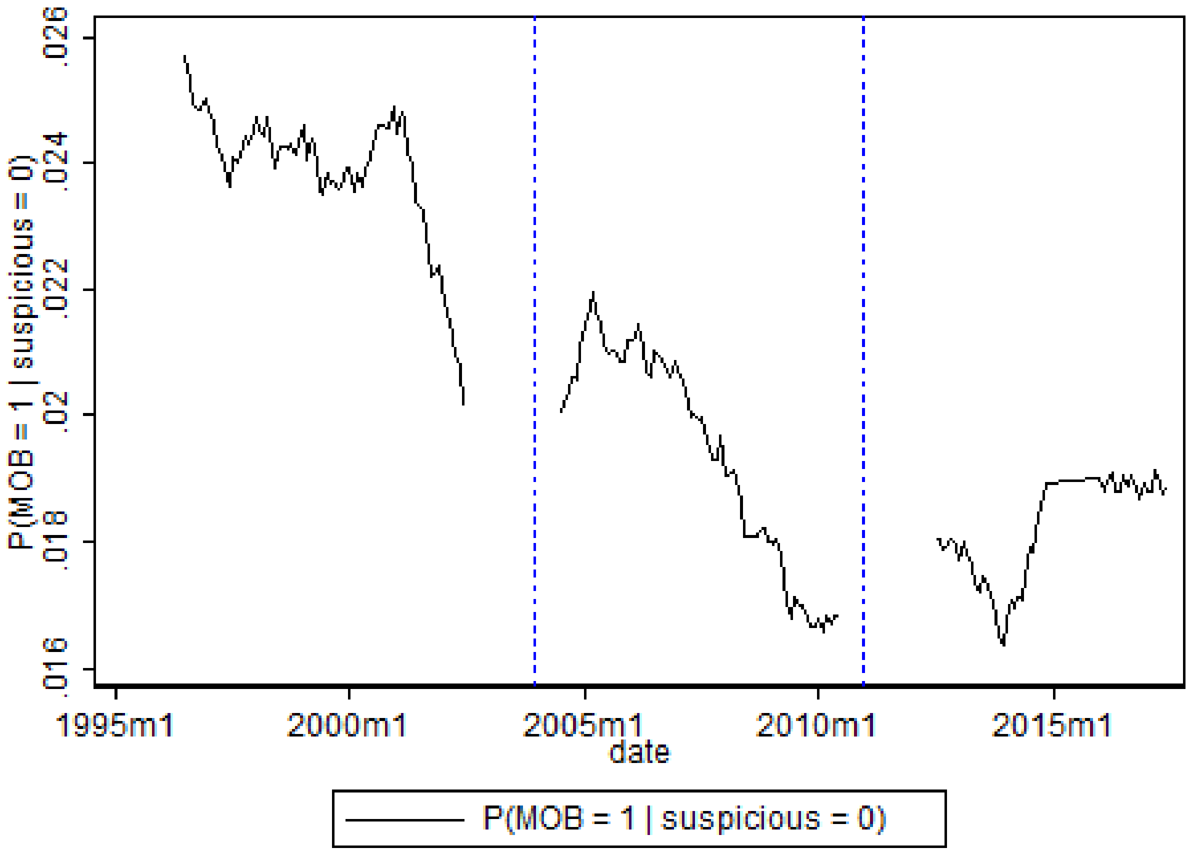


Figure 4: Probability MOB = 1 Given Suspicious = 1, 12 Month Moving Average