ECONOMIC RESEARCH
FEDERAL RESERVE BANK OF ST. LOUIS

WORKING PAPER SERIES

International trade and labor reallocation: misclassification errors, mobility, and switching costs

Authors	Maximiliano A. Dvorkin
Working Paper Number	2021-014E
Revision Date	June 2024
Citable Link	https://doi.org/10.20955/wp.2021.014
Suggested Citation	Dvorkin, M.A., 2024; International trade and labor reallocation: misclassification errors, mobility, and switching costs, Federal Reserve Bank of St. Louis Working Paper 2021-014. URL https://doi.org/10.20955/wp.2021.014

Federal Reserve Bank of St. Louis, Research Division, P.O. Box 442, St. Louis, MO 63166

The views expressed in this paper are those of the author(s) and do not necessarily reflect the views of the Federal Reserve System, the Board of Governors, or the regional Federal Reserve Banks. Federal Reserve Bank of St. Louis Working Papers are preliminary materials circulated to stimulate discussion and critical comment.

International trade and labor reallocation: misclassification errors, mobility, and switching costs[*]

Maximiliano Dvorkin[†]

Abstract

International trade has increased at a rapid pace in the last decades, altering production and labor demand in different sectors of the economy. The estimated effects of trade on employment and welfare critically depend on data about workers' reallocation patterns, which is typically plagued with coding errors. I show that the estimated employment and welfare effects of international trade, and the estimated structural parameters of standard models are biased when the analysis uses data subject to misclassification errors. I develop an econometric framework to estimate misclassification probabilities, corrected mobility matrices, and structural parameters, and show that the estimated employment and welfare effects of a trade shock are different from those estimated with uncorrected data, raising an important warning about conclusions drawn from data with coding errors.

Keywords: International trade, labor markets, classification errors, mobility, worker reallocation, structural estimation

JEL Classification: F16, F66, J24, J62, C25.

[*] I thank Jonathan Eaton, Fernando Leibovici, Amanda Michaud, Chris Neely, Theodore Papageorgiou, B. Ravikumar, Hannah Rubinton, Ana Maria Santacreu, Peter Schott, the Editor, Daniel Xu, three anonymous Referees, and seminar and conference participants for helpful comments. Maggie Isaacson provided excellent research assistance. All views and opinions expressed here are the authors' and do not necessarily reflect those of the Federal Reserve Bank of St. Louis or the Federal Reserve System. [†] *Research Division. Federal Reserve Bank of St. Louis. email: maximiliano.a.dvorkin@stls.frb.org*

1 Introduction

The rapid expansion of international trade over the last few decades has altered domestic production of a wide range of goods and reshaped labor demand across different sectors of the economy. The economic literature has extensively studied how trade liberalizations affect workers, a question that continues to be a topic of intense debate in policy circles. Central to this question is how costly it is for workers to reallocate from industries or occupations that are adversely affected by international trade to those that benefit from it. If reallocation is costly for workers due to periods of unemployment, earnings losses, or a mismatch in the skill they supply relative to those demanded, then reallocation would take place gradually over time and the effects of international trade would be heterogeneous, leading to winners and losers.

In an influential paper, Artuç, Chaudhuri, and McLaren (2010) show that the costs workers face to reallocate across different labor markets critically shape the dynamics of adjustment after a trade liberalization and affect workers' welfare in ways that are not directly reflected by earnings alone. As they clearly put, "[t]he welfare effects of trade shocks turn on the nature and magnitude of the costs workers face in moving between sectors". These forces are typically missing in widely-used models studying the welfare effects of international trade, as workers are assumed to be homogeneous and perfectly mobile across sectors. For example, Arkolakis, Costinot, and Rodríguez-Clare (2012) show that trade flows and the trade elasticity are key determinants in the quantification of the gains from trade in many standard models. However, Caliendo, Dvorkin, and Parro (2019) extend their welfare formulas and show that, in addition to the trade flows and the trade elasticity, the matrix of *workers' mobility flows* across industries and the *mobility elasticity* shape the gains from trade in a model where labor reallocation is costly. Thus, to fully understand the effects of international trade on labor markets and workers' welfare, information about how individuals reallocate across industries and occupations becomes critical.

An important limitation in measuring worker mobility across labor markets is the pervasive contamination of workers' industry and occupation information by misclassification errors. In survey data, workers are asked about the characteristics of their job and their employer, and the answers are then mapped into a specific industry or occupation by a professional coder. Small differences in the way individuals answer these questions or differences in the person coding the answers over time for the same individual can lead to errors in the industry and occupation codes assigned. This type of misclassification errors has long been studied in the literature. In particular, Murphy and Topel (1987), study the effects of industry misclassification on estimates of industry earnings and how these change when workers switch industries. Moreover, Moscarini and Thomsson (2007), and Kambourov and Manovskii (2008), study the effects of misclassification errors in measures of industry and occupation mobility and how mobility has evolved over time. All of these studies find evidence of an artificially large amount of mobility that is the result of misclassification errors.

Simonovska and Waugh (2014) emphasize the importance of adequate estimates of the trade elasticity, as this is "one of only two statistics needed to measure the welfare cost of autarky in a large and important class of structural gravity models of international trade", an insight from Arkolakis et al. (2012). Following a similar argument, the key message of the study I present here is that when worker mobility is not properly measured, the estimated effects of international trade will be wrong. In this paper, I highlight the importance of unbiased estimates of the mobility elasticity and mobility flows for calculations of the welfare effects from trade and the dynamics of adjustment to trade shocks.[1]

I start by revisiting the evidence of misclassification in the Panel Study of Income Dynamics (PSID), following Kambourov and Manovskii (2008). I find that, for a very broadly defined group of industries, mobility rates computed using uncorrected data are roughly

[1] While there are potentially many channels that can bias the migration elasticity, such as model misspecification and workers' unobserved characteristics when these are not controlled for, in this paper I focus only on the role of industry misclassification, as other forces leading to parameter bias are well-known and not unique to the context of worker mobility and reallocation.

two times larger than alternative data for which misclassification is likely absent. Errors would be even larger for a finer disaggregation of industry or occupation information. I take this fact as the starting point of my analysis and evaluate how misclassification affects the estimates of the labor market effects of international trade.

The paper's contributions are threefold. First, I use a standard model of costly labor reallocation following Artuç et al. (2010) and introduce a misclassification probability that generates differences between true mobility rates and observed mobility rates. I show that parameters estimated using data with errors are biased. In particular, misclassification biases estimates of the mobility elasticity upwards and the costs of switching industries towards zero.

Second, I propose a method to estimate misclassification probabilities, mobility rates, and the model's structural parameters. This method adapts the algorithm proposed by Arcidiacono and Miller (2011) and uses the Expectation Maximization algorithm, treating workers' true industry (or occupation) as an unobserved state. Misclassification errors not only severely bias measures of occupation and industry mobility, but also measures of wages by industry and the method is able to corrected them as well.

Since misclassification in industry and occupation information are a prevalent feature of the main publicly available U.S. datasets with a panel structure, such as the Panel Survey of Income Dynamics, the National Longitudinal Survey of Youth, the Survey of Income and Program Participation, and the Current Population Survey, the method I propose can be used to correct mobility measures in these and other similar data.[2]

Finally, I show that that differences in estimated mobility flows and structural parameters due to misclassification matter for the calculations of labor market adjustment and welfare

[2] Misclassification errors for industry and occupations have been documented in Murphy and Topel (1987), Moscarini and Thomsson (2007) and Kambourov and Manovskii (2013) for the CPS, Kambourov and Manovskii (2008) and Kambourov and Manovskii (2013) for the PSID, Carrillo-Tudela and Visschers (2023) for the SIPP, and Neal (1999) for the NLSY. Appendix D provides a detailed discussion about misclassification in these surveys and how to apply the econometric method developed in this paper to correct for it.

changes due to trade shocks. Using the estimates of mobility flows corrected of misclassification, I perform different quantitative exercises and show that the estimated employment effects of a trade shock can be as much as 25% different between calibrations that use data corrected from misclassification relative to uncorrected data. In addition, estimations obtained using uncorrected data imply more favorable welfare effects for workers initially attached to manufacturing, the most adversely affected sector to trade, while for the economy calibrated with corrected data, welfare changes are smaller or more negative, depending on the case. Thus, I show how using incorrect data can lead to different conclusions on the effects of a trade shock.

This paper connects with an important recent literature that studies the effects of international trade on workers and their dynamic decisions to work in different labor markets. Recent examples include Artuç and McLaren (2015), Dix-Carneiro (2014), Caliendo et al. (2019), and Traiberman (2019). In all these papers, the adjustment to a trade shock and how welfare changes for different workers depends on estimates of the elasticity of worker mobility across sectors and occupations, and on the estimates of the switching costs or the mobility flow matrix.

Moscarini and Thomsson (2007), Kambourov and Manovskii (2008), and more recently Vom Lehn, Ellsworth, and Kroff (2020), propose methods to estimate the probability of occupation or industry switching (a binary decision) using CPS or PSID data, corrected for misclassification. Critically, estimates of the employment and welfare effects of trade depend on the *whole matrix* of industry or occupation flows, rather than on the binary decision of whether to switch the current industry or occupation. The method I propose here delivers corrected estimates of the mobility matrix, as in Carrillo-Tudela and Visschers (2023),[3] but contrary to them, offers the advantage of generating estimates of individual-level probabilities

[3] Note that Kambourov and Manovskii (2008) also estimate corrected mobility flows, but their analysis focuses more closely on on the switching rate, that is worker's probability of moving out of the current industry or occupation.

of industry and occupation mobility, which I use to correct measures of wages for different industries and occupations.[4]

The paper is organized as follows. Section 2 describes the data and presents evidence of misclassification errors. Section 3 uses a standard model of labor market choice with frictions and shows how, in a simple calibration, misclassification biases measures of mobility and parameter values. Section 4 develops an econometric framework to correct for misclassification and Section 5 uses this method to obtain estimates of mobility flows and structural parameters that are not affected by this error and shows how the estimated effects of a trade shock differ when using corrected and uncorrected data.

2 Inter-industry mobility in the PSID

I build on Kambourov and Manovskii (2008) and analyze worker mobility across industries in the Panel Study of Income Dynamics (PSID). The main difference here relative to Kambourov and Manovskii (2008) is my focus on the patterns of industry mobility by origin and destination. In other words, I focus on the *matrix of mobility flows*, while they focused largely on the probability of an industry or occupation switch.

A well-known problem with self-reported industry (and occupation) information in survey data is the prevalence of misclassification errors. Murphy and Topel (1987), Kambourov and Manovskii (2008, 2013) and Moscarini and Thomsson (2007) study the importance of misclassification in the CPS and the PSID, two of the most important labor market surveys in the United States.[5] Misclassification errors are typically related to how survey respondents

[4] This is particularly valuable for richer models that have occupation or industry-specific human capital and returns to tenure, a feature I abstract of in this paper, where measures of wage changes at the individual level for workers switching labor markets are needed for estimation.

[5] In a recent work, Carrillo-Tudela and Visschers (2023) also study occupational mobility and correct for misclassification using the Survey of Income and Program Participation (SIPP). Their approach to correct mobility measures follows Poterba and Summers (1986). Similarly, Neal (1999) discusses this problem in the National Longitudinal Survey of Youth. Vom Lehn et al. (2020) also study occupation misclassification errors in the CPS data exploiting retrospective and panel data information about individuals surveyed. They offer corrected estimates of the overall switching rate, that is, the probability that a worker changes his/her

answer the questions on “what kind of business or industry” they work in, or “what sort of work” they do, and how these answers are then interpreted and mapped to a code in the classification system by a professional coder. If workers perform a variety of tasks and firms produce different types of products and services, the description that respondents provide to answer these questions may vary from time to time, even if the characteristics of their work did not change. A coder would then assign different industry or occupation codes over time for an individual, leading to misclassification.

As discussed in Kambourov and Manovskii (2008), in 1999 the PSID released a new set of occupation and industry codes for most individuals in the sample in years 1968 to 1980 that had industry and occupation information. These are known as retrospective industry and occupation codes.[6] The retrospective industry and occupation information originated as part of the Working Lives and Mortality in an Aging National Cohort project, funded by the National Institute on Aging and the National Science Foundation, which required industry and occupation information at a level of disaggregation finer than the one available in the original PSID files. To produce this finer level of disaggregation, coders accessed individuals' responses with the written descriptions of their job characteristics from the PSID archives. “To save time and increase reliability, the coder coded all occupations and industries for each person across all required years before moving on to the next case” (PSID, 1999). Kambourov and Manovskii (2008) argue that coding the entire time-series responses on industry and occupation of an individual drastically reduces misclassification errors.

One way to gauge the amount of misclassification errors is comparing the original and retrospective industry information of individuals in the sample. Appendix D shows that, for

occupation, but do not offer corrected measures of the flows by origin or destination. While their online Appendix E presents a corrected mobility flow matrix, this is just the result of a correction of the main diagonal of the matrix (one minus the probability of switching) and a proportional adjustment/imputation of the off-diagonal elements.

[6] Some individuals were excluded from the retrospective coding sample. This includes individuals that had died by 1992 or that reported a main job in fewer than three waves of the survey. See PSID (1999) for more information.

broad industry groups, differences in industry codes assigned to individuals between original and retrospective codes are close to 8%, suggesting that misclassification is significant.[7]

Table 8 in Appendix D shows that misclassification has a modest impact on the estimates of employment shares by industry. However, the measures of industry *mobility* are significantly impacted by misclassification. Table 1 shows matrices of yearly mobility across broad industries using the original and retrospective codes. In both cases, and at this level of disaggregation, industry switching is not very frequent, as can be seen by the high rates along the diagonal of the matrices relative to the mobility rates outside of the diagonal. However, as these matrices show, out-mobility rates are lower using retrospective codes than the original codes. On average, industry out-mobility rates are 10.7% using the original codes and 5.5% using the retrospective codes for the years in which both are available. These numbers align well with those in Kambourov and Manovskii (2008), with the differences stemming mostly from a different level of industry disaggregation and sample selection criteria.[8]

In sum, misclassification errors overstate the amount of industry (and occupation) mobility in the data. I now study the effects of these errors on parameter estimates using a dynamic model of industry switching.

[7] I aggregate industry into six broad groups. These groups correspond to those defined in Artuç et al. (2010) and are slightly more aggregate than one-digit SIC codes. The industries I use are, (1) agriculture, forestry, fisheries and mining; (2) construction; (3) manufacturing; (4) transportation, communications, and other public utilities; (5) wholesale and retail trade; and (6) all other services, including public administration. Appendix D provides more details on the PSID data and the mapping between SIC codes and these broad industry groups. The level of disaggregation in industry classification matters. Differences between original and retrospective codes are larger for more disaggregated industry groups.

[8] For occupations, Kambourov and Manovskii (2008) show mobility rates of 22% and 11% using original and retrospective coding, respectively between 1968 and 1980, which indicates that occupational mobility rates are twice as large using the original coding relative to the retrospective coding. Using CPS data, Murphy and Topel (1987) and Moscarini and Thomsson (2007) also find a considerably large occupation and industry mobility driven by misclassification error.

Table 1: Matrices of industry mobility
(PSID, 1968-1980)

	Using original coding					
	Agric./Mining	Construct.	Manufact.	Transp./Util.	Whole./Retail	Other serv.
Agric./Mining	0.877	0.026	0.039	0.012	0.017	0.029
	(0.011)	(0.005)	(0.006)	(0.004)	(0.004)	(0.005)
Construction	0.014	0.821	0.057	0.030	0.016	0.061
	(0.003)	(0.011)	(0.007)	(0.005)	(0.004)	(0.007)
Manufacturing	0.006	0.012	0.912	0.009	0.028	0.032
	(0.001)	(0.002)	(0.004)	(0.001)	(0.002)	(0.003)
Transport/Util	0.006	0.022	0.026	0.873	0.026	0.046
	(0.002)	(0.004)	(0.004)	(0.009)	(0.004)	(0.006)
Wholesale/Retail	0.004	0.012	0.064	0.015	0.837	0.068
	(0.001)	(0.002)	(0.005)	(0.003)	(0.008)	(0.005)
Other serv.	0.005	0.011	0.026	0.014	0.023	0.921
	(0.001)	(0.001)	(0.002)	(0.001)	(0.002)	(0.003)

	Using retrospective coding					
	Agric./Mining	Construct.	Manufact.	Transp./Util.	Whole./Retail	Other serv.
Agric./Mining	0.927	0.017	0.019	0.007	0.013	0.017
	(0.008)	(0.004)	(0.004)	(0.002)	(0.003)	(0.004)
Construction	0.013	0.911	0.030	0.014	0.014	0.019
	(0.003)	(0.007)	(0.004)	(0.003)	(0.003)	(0.003)
Manufacturing	0.003	0.008	0.956	0.005	0.012	0.016
	(0.001)	(0.001)	(0.003)	(0.001)	(0.001)	(0.002)
Transport/Util	0.006	0.009	0.009	0.939	0.015	0.022
	(0.002)	(0.002)	(0.002)	(0.006)	(0.003)	(0.004)
Wholesale/Retail	0.005	0.009	0.023	0.007	0.922	0.035
	(0.001)	(0.002)	(0.003)	(0.002)	(0.005)	(0.003)
Other serv.	0.003	0.005	0.016	0.006	0.015	0.956
	(0.001)	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)

Note: Author's calculation using PSID data between 1971 and 1980 for the first panel and 1968 and 1980 for the second panel. Fraction of movers in an industry. First year industry by row and second year industry by column. Sample restricted to SRC individuals between 25 and 64 years old employed at the time of the survey in two consecutive years. Standard errors in parenthesis.

3 Biased estimates from misclassification: an example

Given the prevalence of misclassification in workers' industry and occupation information, a natural question is how much it affects quantitative assessments of the effects of trade and the estimates of structural parameters in models of workers' reallocation.

To analyze this, I use the model of labor reallocation in Artuç et al. (2010). To simplify the exposition, in this section I focus only on labor supply and study the problem of a worker that chooses an industry in which to supply labor, taking current and future wages in all industries as given.[9]

Workers begin the period attached to an industry, which is the consequence of past labor supply decisions, and inelastically supply labor in that market, obtaining the market wage. At the end of each period, workers optimally decide the industry to supply labor in the following period. Reallocation is subject to frictions in the form of switching costs. These frictions make the problem of the worker dynamic, as workers trade off the sunk costs of switching with benefits of future wages that accrue over time.

In addition, I assume that workers decisions are influenced by idiosyncratic shocks that affect the costs or benefits of switching to a different labor market. As commonly assumed in the literature, I let these shocks follow a Type-I Extreme Value distribution. Workers are also heterogeneous in their permanent characteristics, like demographic factors or educational attainment, which I capture with a vector of characteristics τ, defining a worker's type.[10]

The recursive problem of the worker with characteristics τ that begins the period in labor market j and has a realization of preference shocks ϵ, is defined in the following way,

$$V_t^j(\tau, \epsilon) = w_t^j(\tau) + \max_{\ell} \left\{ \beta E_{\epsilon'} \left[V_{t+1}^{\ell}(\tau, \epsilon') \right] - \kappa^{j\ell}(\tau) + \nu \epsilon_t^{\ell} \right\},$$

[9] The assumption of perfect foresight for workers is just to simplify the exposition. I relax this assumption later in the analysis following Artuç et al. (2010).

[10] In Section 5.2 and Appendix B, I relax the assumption of time-invariant worker's characteristics τ, allowing, for example, workers' aging.

where $V_t^j(\tau)$ is the value function of the worker, β is the discount factor, $\kappa^{j\ell}(\tau)$ are the origin-destination-specific costs of switching industry, which may differ by workers' characteristics, $w_t^j(\tau)$ is the wage workers with characteristics τ get in labor market j, and ν is a parameter that scales the variance of the preference shocks ϵ and is inversely related to the elasticity of mobility with respect to wage changes.

The first term in the right is the flow utility from consumption. Workers are hand-to-mouth and spend all their labor earnings in the consumption of the numeraire good. The second term describes the worker's choice at the end of the period. The worker chooses the labor market that maximizes future expected lifetime utility, taking into account the costs of switching and preference shocks.

As is well-known, the assumption of i.i.d. Type-I Extreme Value shocks for ϵ leads to a simpler expression for the *ex-ante* value function, $v_t^j(\tau) = E_\epsilon\left[V_t^j(\tau,\epsilon)\right]$, such that,

$$v_t^j(\tau) = w_t^j(\tau) + \nu \log\left[\sum_{\ell=1}^{J} \exp\left(\beta v_{t+1}^{\ell}(\tau) - \kappa^{j\ell}(\tau)\right)^{1/\nu}\right]. \tag{1}$$

Moreover, the probability that the worker chooses labor market ℓ at the end of the period also has a tractable expression given by,

$$\mu_t^{j\ell}(\tau) = \frac{\exp\left(\beta v_{t+1}^{\ell}(\tau) - \kappa^{j\ell}(\tau)\right)^{1/\nu}}{\sum_{k=1}^{J} \exp\left(\beta v_{t+1}^{k}(\tau) - \kappa^{jk}(\tau)\right)^{1/\nu}}. \tag{2}$$

By the law of large numbers, this also represents the fraction of workers that move from j to ℓ each period, such that if $L_t^j(\tau)$ is the total number of workers in labor market j at time t, then the distribution of workers evolves as, $L_{t+1}^{\ell}(\tau) = \sum_{j=1}^{J} L_t^j(\tau)\,\mu_t^{j\ell}(\tau)$.

In the absence of misclassification errors, the matrix $\mu_t(\tau)$ can be mapped directly to the matrix of mobility flows in the data analyzed in the previous section. This is the usual way to proceed in many studies in the literature. However, when there are coding errors in the data, the econometrician does not perfectly observe workers' industry choices. Let s be the true industry of the worker and j the one observed in the data, and define $\alpha_{j|s}(\tau)$ as the probability that the labor market of worker with characteristics τ is recorded as j when the true labor market is s.[11]

Assumption 1: The distribution of misclassification errors depends only on the worker's true industry in the current period and his/her observed characteristics τ.[12]

This assumption, which is commonly used in the misclassification literature (i.e. Feng and Hu (2013)), limits the amount of temporal dependence in misclassification. Any past history of occupation choices and coding errors does not affect misclassification in the current period. This assumption seems reasonable if a worker's industry (or occupation) is coded independently from any past information each period, as is the case in many surveys, including the way original codes were assigned in the PSID.

In this way, I can represent the observed mass of workers in labor market j at time t as,

$$\tilde{L}_t^j(\tau) = \sum_{s=1}^{J} L_t^s(\tau)\, \alpha_{j|s}(\tau),$$

where $\tilde{L}_t^j$ denotes the *observed* mass of workers in labor market j. The observed number of workers moving from market j to ℓ is,

$$\tilde{L}_t^j(\tau)\tilde{\mu}_t^{j\ell}(\tau) = \sum_{s=1}^{J}\sum_{s'=1}^{J} L_t^s(\tau)\, \alpha_{j|s}(\tau)\, \alpha_{\ell|s'}(\tau)\mu_t^{ss'}(\tau), \tag{3}$$

where $\tilde{\mu}_t^{j\ell}(\tau)$ is the mobility matrix of the observed flows between j and ℓ, and s, s' are the true industries of workers in period t and $t+1$, respectively.

[11] Poterba and Summers (1986) define the matrix with elements $\alpha_{j|s}(\tau)$ as the matrix of error rates.

[12] Note that it is not necessary for the vector of workers' characteristics τ to be time-invariant. See also footnote 10.

Similarly, observed wages may differ from actual wages according to,

$$\tilde{w}_t^j(\tau)\,\tilde{L}_t^j(\tau) = \sum_{s=1}^{J} \alpha_{j|s}(\tau) w_t^s(\tau) L_t^s(\tau), \tag{4}$$

where $\tilde{w}_t^j(\tau)$ are observed wages.

3.1 The impact of misclassification errors on parameter estimates

Since the goal is to illustrate the effects of misclassification, I assume a stationary economy in which wages are invariant over time and workers do not differ in observable characteristics τ.[13] As discussed in Artuç et al. (2010), variability in mobility flows and wages over time are needed to separately identify ν and κ, but this variability is absent in a stationary economy. In the example I present here, the value of these parameters is identified by a non-symmetric economy with differences in the level of wages across industries and the functional form assumption of the flow utility.[14]

Here I use the six industries discussed in the previous section. I calibrate wages such that the shares of employment in each sector approximate employment shares listed in the sixth column of Table 2 in Artuç et al. (2010).[15] As is commonly assumed in the literature, the cost of staying in the same industry are normalized to zero, $\kappa^{jj} = 0$, for all j, and for simplicity, here I assume switching costs are homogeneous, such that $\kappa^{j\ell} = \kappa$ for all $j \neq \ell$.[16] I use two sets of parameter values in this example. First, I use $\nu = 1.8$ and $\kappa = 7$. Second, I use $\nu = 2.2$ and $\kappa = 10$. In both cases I make $\beta = 0.97$. These values correspond to

[13] In the estimation in Section 5, I relax both these assumptions.

[14] Note that different specifications of the flow utility, like multiplying by a constant, impact the values of ν and κ. Despite this caveat, this example illustrates the impact of misclassification errors on parameter estimates, keeping all else equal.

[15] The model I use abstracts from amenities that may influence worker's flow utility and industry choice. However, calibrated wages do not differ substantially from those listed in the fourth column of Table 2 in Artuç et al. (2010).

[16] Appendix B shows the results for the case of heterogeneous switching costs that vary by origin and destination.

estimates in Section 5 discussed later and the first set is close to the preferred estimates of Artuç et al. (2010) (Table 3, Panel IV).[17]

Finally, I assume that misclassification errors are $\alpha_{j|s} = 1 - (J-1)\psi$, for $j = s$, and $\alpha_{j|s} = \psi$, for $j \neq s$, where ψ is a small scalar representing the importance of misclassification error.[18]

To estimate parameters ν and κ, I use equation (9) in Artuç et al. (2010), which links model parameters with mobility probabilities and wages that can be measured in the data. I re-arrange that equation and reproduce it here for convenience,[19]

$$\left[\log\left(\tilde{\mu}^{ij}\right) - \log\left(\tilde{\mu}^{ii}\right)\right] - \beta\left[\log\left(\tilde{\mu}^{ij}\right) - \log\left(\tilde{\mu}^{jj}\right)\right] = \mathbb{C}_1 + \mathbb{C}_2\left(\tilde{w}^{j} - \tilde{w}^{i}\right). \tag{5}$$

The left-hand side of equation (5) is a function of observed mobility rates between two labor markets, i and j, and the discount factor β, which I take as known by assumption and is not estimated. The regressors in the equation are wage differences across labor markets, which in a non-symmetric economy will be different from zero, and a constant term. In the absence of misclassification, the model relates structural parameters and the estimates in the regression in the following way, $\mathbb{C}_1 = -(1-\beta)\frac{\kappa}{\nu}$ and $\mathbb{C}_2 = \frac{\beta}{\nu}$. In this example there is no need to simulate data: the calibrated wages and parameters deliver values for v^j and μ^{ij} using (1) and (2). With misclassification, I use the moments implied by (3) and (4) to obtain $\tilde{\mu}^{ij}$ and $\tilde{w}^{j}$, which are the variables an econometrician that ignores misclassification would use in the regression.

[17] Artuç et al. (2010) use data from the Annual Social and Economic Supplement of the CPS, also known as the March CPS. As I discuss in Appendix D.3, there are important problems in measuring mobility with the March CPS and Artuç et al. (2010) propose a simple correction method. However, their correction method targets mobility moments from the National Longitudinal Survey of Youth, which as I discuss in the Appendix, are subject to similar misclassification errors as in the monthly CPS and the PSID data.

[18] These assumptions just simplify the exposition, as the goal here is to illustrate the effects of the errors. The method discussed later in Section 4 is more general and allows for richer heterogeneity in the probability of misclassification and model parameters. Moreover, the method is suitable for a non-stationary environment, with the exception of α, which I assume is time-invariant.

[19] Appendix B shows the derivation of this expression from the model.

Table 2 shows the estimated value of parameters for an economy with no misclassification error ($\psi = 0$) in column 2, and different levels of misclassification error in columns 3 to 5. Not surprisingly, in an economy with no error, estimated values coincide with the true value of parameters and the regression has an R-squared of one. On average, measured mobility in this economy is between 4% and 8%, depending on the example, values that are close to the uncorrected data discussed in the previous section for the earlier years in the sample.

Table 2: The effect of misclassification on parameter estimates

Panel A				
True parameters	No error	misclassification		
$\nu = 1.8$, $\kappa = 7.0$	($\psi = 0$)	($\psi = 0.001$)	($\psi = 0.005$)	($\psi = 0.01$)
Estimated ν	1.8	1.6	1.2	0.8
Estimated κ	7.0	6.0	3.8	2.4
Estimated ratio κ/ν	3.9	3.8	3.4	2.4
R-squared	1.0	0.999	0.996	0.990
Avg. outflow prob.	0.08	0.09	0.12	0.17
Panel B				
True parameters	No error	misclassification		
$\nu = 2.2$, $\kappa = 10.0$	($\psi = 0$)	($\psi = 0.001$)	($\psi = 0.005$)	($\psi = 0.01$)
Estimated ν	2.2	1.8	1.0	0.6
Estimated κ	10.0	7.8	3.8	2.1
Estimated ratio κ/ν	4.5	4.3	3.8	2.0
R-squared	1.0	0.998	0.987	0.973
Avg. outflow prob.	0.04	0.05	0.09	0.13

Note: Parameters estimated by OLS according to equation (5). Each observation is a sector of origin-destination, where origin is different from destination. The example uses 6 sectors, and there are 30 observations in each regression. Panel A shows results for structural parameter values $\nu = 1.8$, $\kappa = 7.0$ and different levels of misclassification errors. Panel B shows results for $\nu = 2.2$, $\kappa = 10.0$.

As the value of ψ increases, there is more misclassification in the model. In this particular example with six industries, the probability of being misclassified is five times the value of ψ. Thus, a value of ψ equal to 0.005 implies a probability of 2.5% that a worker will be classified in an industry different than the one she is working in. For all values of ψ greater than zero reported in table 2, the regression does not take into account that the model may be subject to misclassification error. In this case, the R-squared deviates from its theoretical value of one since equation (5) is not derived from a model with misclassification. Moreover, the measured (average) mobility in the last row of the table increases with larger ψ. Column 4 shows the results for an economy with a moderate amount of error ($\psi = 0.005$). Measured mobility is 12% and 9% in Panels A and B, respectively, roughly fifty to one hundred percent larger than the true mobility rate. This aligns well with the discussion in the previous section on differences in mobility in the data for the original and retrospective coding. In this case, estimated values for κ and ν (related to mobility costs and mobility elasticity) are around half their true value in both panels, with the exception of the estimates of κ in Panel B, which is estimated around one third the actual value.[20]

Why does misclassification bias the parameter values towards zero? On the one hand, this economy exhibits an artificially large amount of mobility due to misclassification. If mobility was uniformly larger across all labor markets such that the left-hand side of equation (5) increases by the same magnitude due to misclassification, estimates of the constant term would be less negative and closer to zero, which can be rationalized by a lower value of the ratio κ/ν. Table 2 show how the estimated ratio κ/ν falls as the error increases. This agrees with Artuç et al. (2010), page 1017, when they argue that coding errors will result in an "underestimate of the ratio" κ/ν.

However, the effects of misclassification on the estimate of parameter ν are more nuanced. Artuç et al. (2010) argue that "coding errors [...] will definitely provide an overestimate of ν" as wage differences would appear less relevant for mobility relative to the preference shocks. At first sight, this intuition seems reasonable as the structural model implies more mobility when ν is larger, all else equal.

Yet, the effect on estimates for $\mathbb{C}_2$, and by extension ν, depends on how misclassification

[20] In Appendix B, I show that this conclusion also extends to the case of heterogeneous switching costs that vary by origin and destination. Additionally, it is possible to use the Delta method to derive standard errors for parameter estimates. With no misclassification, standard errors are zero in this simple example, but increase as ψ increases.

affects: (1) the correlation between the dependent variable (differences in mobility rates) and the regressor (differences in wages), and (2) the variance of the regressor. Appendix A shows that the dependent variable using uncorrected data can be decomposed into the correct value and an additive measurement error which is correlated with the (correct) regressor. Thus, since the error in the dependent variable is correlated with the regressor, it departs from the standard case of "classical measurement error", leading to inconsistent estimates. Moreover, the appendix shows that the regressor constructed with uncorrected data can also be decomposed into the correct variable and an additive measurement error which is negatively correlated with the correct regressor, leading to a lower variance.[21] In this way, estimates for $\mathbb{C}_2$ with uncorrected data are biased upwards, and the estimated value for ν underestimates the true value of the parameter. Similarly, the appendix shows that the estimate for $\mathbb{C}_1$ using uncorrected data is biased upwards, and given the bias in ν this leads to a downward bias in κ.

It is worth highlighting that welfare calculations depend on wages, mobility, and parameter ν. In the context of this simple model, welfare, or expected lifetime utility, for a representative individual in industry i would be equal to $v^i = \frac{1}{1-\beta}\left(w^i - \nu \log\left(\mu^{ii}\right)\right)$. Thus, misclassification errors also bias model-implied measures of welfare and how welfare changes with, for example, trade shocks.[22]

While this section illustrates to what extent misclassification biases estimates of structural parameters, it does not estimate misclassification probabilities and parameters in the data. I do this next.

[21] Therefore, this also departs from the "classical" case of measurement error in the regressor, as the lower variance of the observed (uncorrected) regressor does not lead to the typical attenuation bias, but to a magnification bias.

[22] See Appendix G for a derivation of welfare formulas.

4 Main econometric framework

I propose an empirical model to estimate the probability of misclassification, mobility matrices, employment shares and wages corrected from misclassification, and structural parameters.[23] In this section I specify the method in general terms and in the next section I adjust it to fit the characteristics of the available data. In Section 4.1, I specify the econometric model, which does not have to be linked to a specific structural model, such as the one in the previous section. Since estimation of structural parameters, denoted by θ, using the general model can be very demanding, in Section 4.2, I adapt the ideas in Arcidiacono and Miller (2011) and propose a two-stage estimation method. In the first stage, some endogenous variables, such as the matrix of mobility probabilities or conditional choice probabilities, are themselves treated as parameters that can can be estimated at a low computational costs using the EM algorithm. Then, the second stage uses these estimates to pin-down values for the structural parameters θ.

4.1 General model

The sample consists of N individuals such that $n = \{1, \ldots, N\}$, and the econometrician observes workers' industry (or occupation) original codes and/or retrospective codes for T periods. I assume that retrospective codes have no errors, following Kambourov and Manovskii (2008) and the discussion in Section 2. In addition, the econometrician observes wages. Let $\alpha_{j_t|s_t}(\tau_t)$ be the probability that individual n with observed characteristics τ_t and a true occupation/industry s_t has an observed occupation/industry j_t in period t. Thus, $\alpha_{j_t|s_t}(\tau_t)$ reflects the probability of misclassification. By Assumption 1, the misclasssification probability in period t only depends on the worker's true industry/occupation in period t

[23] While most of the analysis and the application in this paper focuses on industry mobility, the method I develop here applies to misclassification of other variables, like occupations. Thus, in this section I refer to both industry and occupation misclassification and how to correct for it.

and worker's characteristics.[24]

Let $p_{s_1}(\tau, \theta)$ be the probability that a worker has true occupation/industry s_1 in the first period in which she is observed in the sample, and and $p_{s_{t+1}|s_t}(\tau, \theta)$ be the probability that a worker with true occupation/industry s_t switches to industry/occupation s_{t+1} the following period. These are functions of structural parameters listed in the vector θ. In the context of the model discussed in Section 3, $p_{s'|s}(\tau, \theta)$ are the conditional choice probabilities of the structural model, which before I denoted by $\mu^{ss'}(\tau)$ and the vector θ is comprised of ν and κ. Note however that the estimation proposed here is more general and does not have to be linked to a specific structural model, thus the alternative notation.

In addition, I assume that observed log-wages for an individual n are subject to classical measurement error or to i.i.d. income shocks realized at the beginning of the period, such that $log(\omega^n) = \log(w_s(\theta, \tau)) + \eta^n$.[25] w_s is the common component of wages for all individuals with characteristics τ in labor market s. In the context of the structural model of Section 3, these are the wages of the model, which may be assumed exogenous as in the previous section, or in a richer general equilibrium framework, they may depend on structural parameters.

Let d^n_{jt} be a dummy variable that assumes the value one if worker is in observed industry/occupation j in period t and zero otherwise. Then, the contribution to the likelihood of observing the sequence (τ^n, d^n, ω^n) for individual, n, conditional on parameters, is,

$$\begin{aligned}\mathcal{L}(\tau^n, d^n, \omega^n|\theta, \alpha) \quad = \quad & \sum_{s_1=1}^{J}\sum_{s_2=1}^{J}\cdots\sum_{s_T=1}^{J}\prod_{j_1=1}^{J}\prod_{j_2=1}^{J}\cdots\prod_{j_T=1}^{J}\Bigg[\ \left[p_{s_1}(\tau_1^n, \theta)\,\alpha_{j_1|s_1}(\tau_1^n) f(\log(\omega_{j_1}^n / w_{s_1}(\theta, \tau_1^n)))\right] \times \\ & \left[p_{s_2|s_1}(\tau_2^n, \theta)\,\alpha_{j_2|s_2}(\tau_2^n) f(\log(\omega_{j_2}^n / w_{s_2}(\theta, \tau_2^n)))\right] \times \ldots \times \\ & \left[p_{s_T|s_{T-1}}(\tau_T^n, \theta)\,\alpha_{j_T|s_T}(\tau_T^n) f(\log(\omega_{j_T}^n / w_{s_T}(\theta, \tau_T^n)))\right]\ \Bigg]^{d_{j_1}^n \times d_{j_2}^n \times \ldots \times d_{j_T}^n}\end{aligned} \tag{6}$$

Note that, for retrospective coding information $\alpha_{j_t|s_t}(\tau_t^n) = 1$ if $j_t = s_t$ and zero otherwise. In each period, the contribution to the likelihood is the joint probability that an individual

[24] Importantly, Assumption 1 implies that misclassification probabilities are time-invariant.

[25] An important assumption is that i.i.d. income shocks are not persistent and are not known at the time of selecting labor markets or that they do not affect the labor market choice decision.

with actual industry/occupation s has observed industry/occupation j, respectively, and the likelihood of the observed wage given the true industry/occupation. The log-likelihood of the sample is, $\sum_{n=1}^{N} \log\left[\mathcal{L}^n(\tau^n, d^n, \omega^n|\theta, \alpha)\right]$.

Assumption 2: Structural parameters θ are a differentiable function of time and are independent of misclassification.

Assumption 2 is key for identification of the coding error probabilities α and the structural parameters θ. In particular, it implies that the evolution of actual mobility rates, whether directly observed or not, is smooth (with no jumps) and independent of coding errors, ruling out, for example, that mobility rates change suddenly or that they differ arbitrarily between observations with original coding and those with retrospective coding.[26]

Note that the estimation allows for industry/occupation mobility, as well as wages and employment shares, to vary over time in a flexible way.[27] In particular these objects can be differentiable functions of time if the vector of parameters θ contain time trends. In the application of the next section I allow for these trends.

A structural model, like the one presented in Section 3, specifies a functional form for the conditional choice probability $p_{s'|s}(\tau, \theta)$. By making additional functional form assumptions for the density f and the initial industry/occupation allocation $p_{s_1}(\tau, \theta)$, which could also be modeled structurally, it is possible to find values for parameters θ and α that maximize the likelihood of the sample. Note, however, that a direct maximization of all these parameters can be quite demanding as the number of parameters to estimate may grow geometrically with the number of occupations/industries, and it involves maximizing a non-linear, and potentially not well-behaved, function over many dimensions. In particular, the sum inside

[26] This follows closely Kambourov and Manovskii (2008), as their estimated probit models had common parameters across periods with parameteric trends. Kambourov and Manovskii (2008) did not estimate a structural model of occupational/industry choice, but a statistical one. The parameters of that model would play a similar role as parameters α and θ used here.

[27] Given my identification assumptions, I cannot allow for a fully flexible, non-parametric, time-variation in the patterns of mobility as they would not be identified separately from misclassification error (or its changes over time).

the logarithm of the likelihood does not permit simpler expressions and all parameters must be maximized jointly. In the next subsection I adapt the insights from Arcidiacono and Miller (2011) and use the Expectation-Maximization (EM) algorithm to estimate the parameters of the model in a very tractable manner.

4.2 Expectation-Maximization algorithm

The EM algorithm is a widely used technique to estimate parameters by maximum likelihood, and is commonly used in models with latent variables. In a recent influential paper, Arcidiacono and Miller (2011) propose a two-step approach using the EM algorithm to estimate parameter values of dynamic discrete choice models with unobserved heterogeneity. While my setup does not map directly to their class of models, it is closely related as I treat the true industry/occupation of a worker as a latent state unobserved by the econometrician.

In a similar fashion as Arcidiacono and Miller (2011), in the first stage I treat the functions $p_s(\tau, \theta)$ and $p_{s'|s}(\tau, \theta)$ as primitives to be estimated together with misclassification parameters $\alpha_{j|s}(\tau)$, and parameters linked to wages. In a second stage, I use the estimates from the first stage to estimate switching costs and mobility elasticity.

First stage. The starting point of the EM algorithm is to treat the unobserved –or latent state– as observed, assigning a probability to each individual in the sample of being in each particular state. Define $\mathcal{L}^n(s_1^n = \tilde{s}_1, \ldots, s_T^n = \tilde{s_T}|\theta, \alpha) = \mathcal{L}(\tau^n, d^n, \omega^n|\theta, \alpha, s_1^n = \tilde{s}_1, \ldots, s_T^n = \tilde{s_T})$ to be the contribution of observations from individual n to the likelihood conditional on parameters and on the individual's true industries/occupations being $\{\tilde{s}_1, \ldots, \tilde{s_T}\}$, which using (6) is simply,

$$
\begin{aligned}
\mathcal{L}^n(s_1^n = \tilde{s}_1, \ldots, s_T^n = \tilde{s}_T|\theta, \alpha) \quad = \quad & \prod_{j_1=1}^{J} \prod_{j_2=1}^{J} \cdots \prod_{j_T=1}^{J} \Bigg[\left[p_{\tilde{s}_1}(\tau_1^n, \theta)\, \alpha_{j_1|\tilde{s}_1}(\tau_1^n) f(\log(\omega_{j_1}^n / w_{\tilde{s}_1}(\theta, \tau_1^n)))\right] \times \ldots \times \\
& \left[p_{\tilde{s}_t|\tilde{s}_{t-1}}(\tau_t^n, \theta)\, \alpha_{j_t|\tilde{s}_t}(\tau_t^n) f(\log(\omega_{j_t}^n / w_{\tilde{s}_t}(\theta, \tau_t^n)))\right] \times \ldots \times \qquad (7) \\
& \left[p_{\tilde{s}_T|\tilde{s}_{T-1}}(\tau_T^n, \theta)\, \alpha_{j_T|\tilde{s}_T}(\tau_T^n) f(\log(\omega_{j_T}^n / w_{\tilde{s}_T}(\theta, \tau_T^n)))\right] \Bigg]^{d_{j_1}^n \times d_{j_2}^n \times \ldots \times d_{j_T}^n}
\end{aligned}
$$

Using equations (6) and (7), then define,

$$
\begin{aligned}
q^n(\tilde{s}_1,\ldots,\tilde{s}_T) &= \frac{\mathcal{L}^n(s_1^n=\tilde{s}_1,\ldots,s_T^n=\tilde{s}_T|\theta,\alpha)}{\mathcal{L}(\tau^n,d^n,\omega^n|\theta,\alpha)} && (8)\\
q^n_{s_t=s} &= \sum_{s_1=1}^{J}\sum_{s_2=1}^{J}\cdots\sum_{s_{t-1}=1}^{J}\sum_{s_{t+1}=1}^{J}\cdots\sum_{s_T=1}^{J} q^n(s_1,\ldots,s_{t-1},s,s_{t+1},\ldots,s_T) && (9)\\
q^n_{s_{t+1}=s',s_t=s} &= \sum_{s_1=1}^{J}\sum_{s_2=1}^{J}\cdots\sum_{s_{t-1}=1}^{J}\sum_{s_{t+2}=1}^{J}\cdots\sum_{s_T=1}^{J} q^n(s_1,\ldots,s_{t-1},s,s',s_{t+2},\ldots,s_T) && (10)
\end{aligned}
$$

Equation (8) is the probability that individual n has in true industries/occupations $\{\tilde{s}_1,\ldots,\tilde{s}_T\}$, given the sequence of observations for n. Equations (9) and (10) are the joint and conditional probabilities that individual n is in true occupation s' at time $t+1$ and s at time t, respectively. Note that, for observations with retrospective coding, $q^n_s = 1$ if $s_t = j_t$ and zero otherwise.

The maximization step treats the unobserved states as observed, weighting each observation by $q^n_{s_t=s}$. In addition, I treat $p_{s_1}(\tau,\theta)$, $p_{s'|s}(\tau,\theta)$, and $\omega_s(\tau,\theta)$ as parameters themselves in this first step of the estimation. Then, define,

$$
\begin{aligned}
\hat{\mathscr{L}}(\mathbf{d},\boldsymbol{\tau},\boldsymbol{\omega}|p_{s_1},p_{s'|s},w_s,\alpha) = {} & \sum_{n=1}^{N}\sum_{s_1=1}^{J}\sum_{s_2=1}^{J}\cdots\sum_{s_T=1}^{J}\sum_{j_1=1}^{J}\sum_{j_2=1}^{J}\cdots\sum_{j_T=1}^{J}\Bigg[d^n_{j_1}\times d^n_{j_2}\times\ldots\times d^n_{j_T}\times\\
& q^n(s_1,\ldots,s_T)\times\log\bigg(\left[p_{s_1}(\tau_1^n,\theta)\,\alpha_{j_1|s_1}(\tau_1^n)f(\log(\omega^n_{j_1}/w_{s_1}(\theta,\tau_1^n)))\right]\times\\
& \left[p_{s_2|s_1}(\tau_2^n,\theta)\,\alpha_{j_2|s_2}(\tau_2^n)f(\log(\omega^n_{j_2}/w_{s_2}(\theta,\tau_2^n)))\right]\times\ldots\times\\
& \left[p_{s_T|s_{T-1}}(\tau_T^n,\theta)\,\alpha_{j_T|s_T}(\tau_T^n)f(\log(\omega^n_{j_T}/w_{s_T}(\theta,\tau_T^n)))\right]\bigg)\Bigg]. \qquad (11)
\end{aligned}
$$

Using (11), we can easily compute optimal values for parameters at each step of the algorithm. Conditional on q^n, the optimal values for parameters are,

$$\begin{aligned}
p_{s_1=s}(\tau,\theta) &= \frac{\sum_{i=1}^{N}\sum_{j_1=1}^{J} d_{j_1}^{n}\, q_{s_1=s}^{n}\mathbb{I}(\tau_1^n=\tau)}{\sum_{i=1}^{N}\sum_{\tilde{s}=1}^{J}\sum_{j_1=1}^{J} d_{j_1}^{n}\, q_{s_1=\tilde{s}}^{n}\,\mathbb{I}(\tau_1^n=\tau)}, && (12)\\
p_{s_{t+1}=s'|s_t=s}(\tau,\theta) &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T-1}\sum_{j_{t+1}=1}^{J}\sum_{j_t=1}^{J} d_{j_t}^{n} d_{j_{t+1}}^{n}\, q_{s_{t+1}=s',s_t=s}^{n}\,\mathbb{I}(\tau_{t+1}^n=\tau)}{\sum_{i=1}^{N}\sum_{t=1}^{T-1}\sum_{j_{t+1}=1}^{J}\sum_{j_t=1}^{J}\sum_{\tilde{s}'=1}^{J} d_{j_t}^{n} d_{j_{t+1}}^{n}\, q_{s_{t+1}=\tilde{s}',s_t=s}^{n}\,\mathbb{I}(\tau_{t+1}^n=\tau)} && (13)\\
\alpha_{j|s}(\tau) &= \frac{\sum_{i=1}^{N}\sum_{t=1}^{T} d_{j_t=j}^{n}\, q_{s_t=s}^{n}\,\mathbb{I}(\tau_t^n=\tau)}{\sum_{i=1}^{N}\sum_{t=1}^{T}\sum_{\tilde{j}=1}^{J} d_{j_t=\tilde{j}}^{n}\, q_{s_t=s}^{n}\,\mathbb{I}(\tau_t^n=\tau)} && (14)
\end{aligned}$$

where $\mathbb{I}(\tau^n=\tau)$ is an indicator that is equal to one if $\tau^n=\tau$.[28] Moreover, $\hat{w}_s(\tau)$ can be estimated flexibly using weighted means for different τ, using q_s^n as weights.[29]

The algorithm iterates between the expectation step, updating q_s^n conditional on parameters, and the maximization step, updating parameters conditional on q_s^n. At each step the likelihood increases until a maximum is attained.

Second stage. We can use the fist stage estimates for $p_{s_1}(\tau,\theta)$, $p_{s'|s}(\tau,\theta)$, and $w_s(\tau,\theta)$ in a second stage to obtain estimates for structural parameters θ. As discussed in Section 6.2. of Arcidiacono and Miller (2011), it is possible to form a likelihood function to estimate these parameters, but alternative estimators would also be consistent. In particular, we can use the estimator in Artuç et al. (2010) discussed in Section 3, with the variables corrected for misclassification, $\hat{w}_s(\tau)$ and $\hat{p}_{s'|s}(\tau)$.[30]

Four aspects are worth highlighting. First, depending on the data available and the identification assumptions, it is possible to modify the likelihood function to accommodate different characteristics of the sample.[31] However, the method proposed here would be essentially the same. This is particularly useful when working with other data, such as the CPS, the SIPP and the NLSY, as discussed in Appendix D and used in the next section.

[28] It is possible to make parametric assumptions for the estimators of the choice probabilities and misclassification probabilities. This may be useful in cases where the sample size is small and many transitions from j to j' have very few observations.

[29] Alternatively, if w_s is assumed to be a linear function of parameters, these can be estimated via weighted Ordinary Least Squares.

[30] See Artuç and McLaren (2015) for an alternative estimator of structural parameters.

[31] In other words, the method can be adapted to data without the retrospective coding questions in the PSID that allows identification of parameters in a different way.

Second, the method not only estimates mobility matrices corrected for misclassification, but also a corrected measure of wages. Wages are a key ingredient in the estimation of structural parameters parameters and, as shown in Appendix A, the measurement error in observed wages is negatively correlated with actual wages, leading to biases in the estimates of structuctural parameters if using uncorrected wage data. To the best of my knowledge, this is the first paper that proposes a unified method to obtain measures of occupation/industry mobility *and wages* corrected for misclassification. Third, even for a small number of industries/occupations, direct estimation via maximum likelihood of model parameters can be computationally very demanding. Nonlinear optimization routines typically require a very large number of evaluations to compute not only levels of the likelihood function but also gradients. On the contrary, the optimization step in the EM algorithm for the model proposed here is quite simple to implement, and under fairly common assumptions, the maximization step has a closed-form solution. Fourth, the first stage of the estimator recovers the matrices of industry/occupation mobility, together with wages and industry/occupation shares, corrected for misclassification. For many questions, these variables are interesting on their own and in Appendix C I show that these parameters are identified from moments in the data. The second stage shows how to use them to obtain estimates of structural parameters of a particular model. While different models would lead to different estimators used in the second stage, the first stage would be the same. Thus, the method developed here can be used across different models where misclassification affects observed variables.

5 Estimated mobility, wages, and model parameters

I now employ the method discussed in the previous section and estimate the probability of misclassification error and measures of inter-industry mobility corrected for misclassification. For this I use the PSID and CPS data discussed in Section 2 and Appendix D. For the estimation I restrict the sample to males between 25 and 64 years old, employed at the time of the survey.[32] In addition I focus on the six broad industry groups presented previously. For simplicity, I assume misclassification probabilities differ by industry of origin and destination, but not by individuals' demographic characteristics (or time period).

Estimating $\hat{p}_s(\tau)$, $\hat{p}_{s'|s}(\tau)$, and $\hat{w}_s(\tau)$ non-parametrically using a bin estimator typically requires a large sample, as some transitions tend to be infrequent for some groups of individuals. The usual approach in the literature is to impose a flexible parametric model to estimate these objects. I use a multinomial logit model for $\hat{p}_s(\tau)$ and $\hat{p}_{s'|s}(\tau)$ with a third order polynomial time trend, a polynomial of order two in age and a dummy for some college education or more. In the case of $\hat{p}_{s'|s}(\tau)$, these models are estimated independently for each origin s with the same set of controls.[33] In addition, I assume a log-linear model for wages with Gaussian errors with the same demographic controls and a linear time trend.[34]

First, I obtain estimates for misclassification probabilities, together with corrected employment shares, wages and mobility rates using the PSID and the sample period 1968 to 1980, which contains both original and retrospective industry information.[35] Table 3 shows the estimated probability of being classified in industry j (in columns) when the true industry is s (in rows). This corresponds to the estimates of $\alpha_{j|s}$. Misclassification can be inferred by the elements outside the diagonal of the matrix.

[32] I drop observations with missing industry information.

[33] I estimate p_s using a multinomial logit regression as follows. First, I replicate each individual observation in the sample J times, assigning a value of $s = \{1, ..., J\}$ to each of the replicas. In this way, I have J copies of each observation in the original sample, each with a different value of s. Then, I attach the q_s^n that corresponds to each s of the replica to each observation n and estimate a multinomial logit regression for s, which has J possible categories, using q_s^n as "sample weights" for each observation. To this multinomial logit, I add as controls a third order polynomial time trend, a polynomial of order two in age and a dummy for some college education or more. The resulting predicted probabilities from this estimated multinomial logit are $p_s(\tau)$. For $p_{s'|s}$ the logic is similar, but we need $J \times J$ replicas and the multinomial logit is for s' conditional on each value of s, using $q_{s'|s}^n$ from equation (10) as individual n sample weight.

[34] Note that the log-linear model for wages is used for the estimation of structural wage parameters, but this does not imply any assumptions on the functional form of the flow utility of the structural model.

[35] I initialize the EM algorithm with a symmetric matrix of misclassification error with a parameter $\psi = 0.005$, that is, I assume a small error. See the discussion in Appendix C for details on identification. The starting values for the algorithm for $\hat{p}_s(\tau)$, $\hat{p}_{s'|s}(\tau)$, and $\hat{w}_s(\tau)$ are the ones estimated from a model with no misclassification.

On average, the probability of misclassification is 3.2%, but there is substantial heterogeneity across different industries, ranging from around 2% in agriculture, manufacturing, and other services, to over 5% in retail and wholesale trade. Moreover, a large number of manufacturing workers are misclassified as working in other services, and most of the misclassified workers in retail and wholesale trade are coded into other services and manufacturing, according to the estimates for $\alpha_{j|s}$. Thus, estimates suggest that misclassification errors are heterogeneous and non-symmetric.

It is worth noting that the estimates of $\alpha_{j|s}$ are highly correlated with the discrepancies in coding shown in Table 8, in the appendix, with a correlation coefficient of 0.87. That is, industries with high values of $\alpha_{j|s}$ outside of the diagonal, are also industries with more discrepancies between original and retrospective coding. Note, however, that the estimation of $\alpha_{j|s}$ does not directly use the values of the discrepancy, thus this correlation provides some external validation to the estimates for α.

Table 3: Estimated misclassification probability matrix, $\alpha_{j|s}$

		Observed industry					
		Agric./Mining	Construct.	Manufact.	Transp./Util	Wholes./Retail	Other serv.
actual industry	Agric./Mining	0.979	0.006	0.003	0.002	0.003	0.008
		(0.008)	(0.004)	(0.006)	(0.001)	(0.002)	(0.004)
	Construction	0.004	0.957	0.010	0.008	0.002	0.019
		(0.003)	(0.014)	(0.006)	(0.004)	(0.002)	(0.007)
	Manufacturing	0.002	0.004	0.974	0.004	0.006	0.011
		(0.001)	(0.001)	(0.004)	(0.001)	(0.002)	(0.002)
	Transport/Util	0.001	0.009	0.005	0.962	0.006	0.017
		(0.001)	(0.005)	(0.003)	(0.009)	(0.003)	(0.006)
	Wholesale/Retail	0.001	0.000	0.028	0.007	0.944	0.019
		(0.001)	(0.001)	(0.006)	(0.002)	(0.009)	(0.005)
	Other serv.	0.001	0.005	0.006	0.007	0.006	0.976
		(0.000)	(0.002)	(0.002)	(0.002)	(0.002)	(0.004)

Note: Estimated via EM algorithm using PSID data 1971-1980 for males. The table shows the probability that a worker with true industry s (in rows) has an observed industry j (in columns). See text for details on the estimation method. Block bootstrap standard errors in parenthesis using 500 random samples of individuals (with replacement).

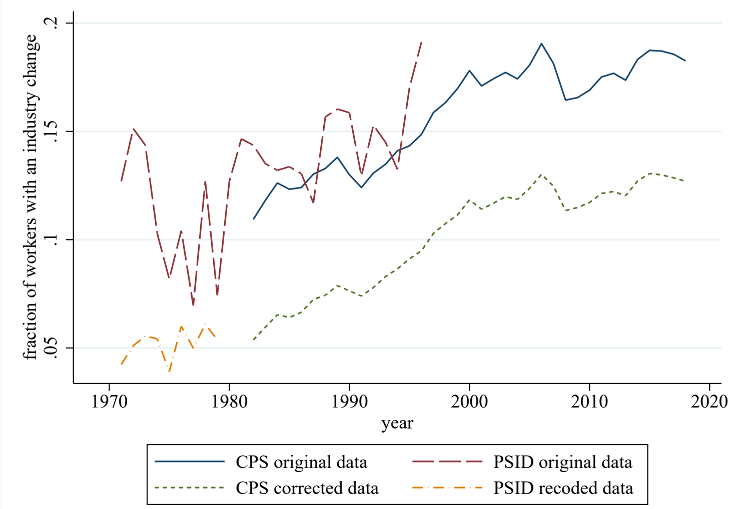
Conditional on the estimated values for α, I profit from the much larger sample size of the CPS and use monthly data for outgoing rotation groups between 1982 to 2018, linking individuals one year apart. I estimate corrected employment shares, wages and mobility rates, averaging across months to obtain a yearly measure, and using these I obtain estimates of structural parameters.[36]

As discussed earlier, misclassification artificially inflates measures of mobility. Figure 1 summarizes overall mobility rates computed with original and corrected CPS data. The solid and short-dashed lines in the figure show, respectively, overall mobility rates for males computed using original data and corrected (estimated) data. Uncorrected mobility rates in the PSID and the CPS are of similar magnitude and follow the same trends during the periods they overlap. As highlighted by Kambourov and Manovskii (2008), industry mobility increased between 1968 and 1997, but appears to have stabilized by early 2000s at around 17% per year. Corrected mobility rates, on the other hand, are lower. The dahsed-dotted line shows overall industry mobility computed using retrospective data up to 1980 in the PSID. Comparing with uncorrected mobility measures for the same period, the difference is a mobility rate that is 80% higher in the uncorrected data. The short-dashed line shows the corrected mobility estimated with CPS data, conditional on the estimated values for $\alpha_{j|s}$. The trends over time, as expected, are similar, but the level is lower, with a mobility rate that stabilized at around 12% per year since 2000.

Average measures of mobility mask substantial heterogeneity in the evolution of mobility across industries. Table 4 shows average industry mobility matrices for four different years using the corrected CPS data. In agriculture, manufacturing, and transportation and utilities, the probability of moving to a different industry increased sharply over time, while in the other sectors, the increase was more moderate. Moreover, the outflows from manufacturing, which had a high concentration of male employment in the early 1980s, translated in large mobility flows into wholesale and retail trade and other services, where the probability more than doubled. The increase in the outflows from wholesale and retail trade were also

[36] Appendix D discusses also the SIPP and NLSY data and how the method developed in Section 4 can be used to correct measures in these data as well.

Figure 1: Probability of broad industry switching
corrected measure vs. original data



Note: Original data refers to yearly industry mobility rate computed using original PSID and CPS codes. Recoded data refers to mobility rates computed using recoded PSID data. Corrected data computes industry mobility rates using CPS data corrected for misclassification estimated via EM algorithm. See the text for details on the estimation method.

heterogeneous, with switches mostly into other services and transportation and utility.

Note that, the different estimates of $\hat{p}_{s'|s}$ in the first panel of Table 4, are closely aligned with the mobility matrix computed using retrospective codes in the PSID data in Table 1, providing confidence on the estimates of corrected rates. Table 9 in Appendix D shows mobility rates computed with uncorrected CPS data. For some industries, such as wholesale and retail trade, the differences in mobility rates using corrected and uncorrected data are substantial.

5.1 Structural parameters

In the previous subsection I estimated corrected mobility matrices and wages. Now I use these corrected measures to estimate structural parameters of the model developed in Section 3.[37]

[37] It is worth remarking that, while different structural models would lead to different structural parameter estimates, the corrected matrices and wages would be the same. In this way, the corrected mobility and wages can be used across different structural models.

Table 4: Estimated matrices of industry mobility, $\hat{p}_{s'|s}$
(Corrected CPS data)

Panel A: 1982-1984

	Agric./Mining	Construction	Manufacturing	Transport/Util	Wholesale/Retail	Other serv.
Agric./Mining	0.884	0.005	0.031	0.008	0.033	0.040
	(0.016)	(0.005)	(0.008)	(0.004)	(0.007)	(0.009)
Construction	0.006	0.903	0.022	0.013	0.026	0.030
	(0.002)	(0.013)	(0.006)	(0.006)	(0.004)	(0.007)
Manufacturing	0.003	0.005	0.950	0.006	0.024	0.013
	(0.001)	(0.002)	(0.005)	(0.002)	(0.003)	(0.003)
Transport/Util	0.002	0.002	0.007	0.982	0.005	0.002
	(0.001)	(0.003)	(0.004)	(0.007)	(0.003)	(0.004)
Wholesale/Retail	0.006	0.017	0.060	0.007	0.882	0.028
	(0.001)	(0.002)	(0.007)	(0.003)	(0.009)	(0.005)
Other serv.	0.002	0.009	0.013	0.001	0.013	0.962
	(0.001)	(0.002)	(0.003)	(0.001)	(0.003)	(0.005)

Panel B: 1989-1991

	Agric./Mining	Construction	Manufacturing	Transport/Util	Wholesale/Retail	Other serv.
Agric./Mining	0.865	0.014	0.037	0.015	0.044	0.024
	(0.015)	(0.007)	(0.009)	(0.004)	(0.008)	(0.005)
Construction	0.005	0.885	0.029	0.009	0.035	0.036
	(0.002)	(0.013)	(0.007)	(0.004)	(0.003)	(0.008)
Manufacturing	0.004	0.008	0.930	0.007	0.031	0.020
	(0.001)	(0.002)	(0.005)	(0.001)	(0.003)	(0.003)
Transport/Util	0.003	0.005	0.014	0.958	0.012	0.009
	(0.001)	(0.003)	(0.004)	(0.007)	(0.004)	(0.005)
Wholesale/Retail	0.007	0.017	0.056	0.011	0.869	0.040
	(0.001)	(0.002)	(0.006)	(0.003)	(0.008)	(0.005)
Other serv.	0.002	0.010	0.018	0.004	0.019	0.947
	(0.001)	(0.002)	(0.003)	(0.002)	(0.002)	(0.005)

Panel C: 1999-2001

	Agric./Mining	Construction	Manufacturing	Transport/Util	Wholesale/Retail	Other serv.
Agric./Mining	0.795	0.032	0.035	0.024	0.063	0.052
	(0.017)	(0.009)	(0.008)	(0.006)	(0.008)	(0.010)
Construction	0.007	0.858	0.033	0.016	0.044	0.043
	(0.002)	(0.013)	(0.006)	(0.005)	(0.003)	(0.008)
Manufacturing	0.006	0.015	0.878	0.012	0.050	0.040
	(0.001)	(0.002)	(0.007)	(0.002)	(0.004)	(0.004)
Transport/Util	0.004	0.011	0.023	0.901	0.026	0.034
	(0.001)	(0.004)	(0.004)	(0.009)	(0.004)	(0.007)
Wholesale/Retail	0.008	0.026	0.057	0.021	0.831	0.058
	(0.001)	(0.002)	(0.006)	(0.003)	(0.009)	(0.005)
Other serv.	0.003	0.013	0.022	0.011	0.026	0.925
	(0.001)	(0.002)	(0.002)	(0.002)	(0.003)	(0.005)

Panel D: 2014-2016

	Agric./Mining	Construction	Manufacturing	Transport/Util	Wholesale/Retail	Other serv.
Agric./Mining	0.791	0.031	0.045	0.032	0.051	0.050
	(0.015)	(0.008)	(0.007)	(0.006)	(0.006)	(0.008)
Construction	0.013	0.839	0.028	0.024	0.042	0.053
	(0.003)	(0.013)	(0.005)	(0.005)	(0.004)	(0.008)
Manufacturing	0.010	0.023	0.844	0.019	0.043	0.062
	(0.001)	(0.003)	(0.009)	(0.002)	(0.005)	(0.006)
Transport/Util	0.009	0.021	0.028	0.849	0.035	0.058
	(0.002)	(0.005)	(0.004)	(0.011)	(0.005)	(0.008)
Wholesale/Retail	0.011	0.029	0.042	0.027	0.819	0.071
	(0.001)	(0.002)	(0.006)	(0.003)	(0.010)	(0.006)
Other serv.	0.004	0.011	0.022	0.016	0.025	0.922
	(0.001)	(0.002)	(0.002)	(0.002)	(0.002)	(0.005)

Note: Fraction of movers in an industry that switch to a different industry. First year industry by row and second year industry by column. Estimated by EM algorithm using CPS data, see text for details. Sample restricted to males between 25 and 64 years old employed at the time of the survey in two consecutive years. Block bootstrap standard errors in parenthesis from 500 random samples with replacement.

As discussed previously, misclassification biases the estimates of structural parameters. I assume individuals face a common cost of switching $\kappa^{ij} = \kappa$ for $i \neq j$, and I begin by estimating a model with no worker heterogeneity, similar to the baseline model in Artuç et al. (2010). I use a more general version of (5), that allows for time variation in wages and mobility probabilities, and that individuals form expectations rationally. In particular, I follow Artuç et al. (2010) and estimate parameters κ and ν using,[38]

$$\left(\log\left[\tilde{\mu}_t^{ij}\right] - \log\left[\tilde{\mu}_t^{ii}\right] - \beta\left(\log\left[\tilde{\mu}_{t+1}^{ij}\right] - \log\left[\tilde{\mu}_{t+1}^{jj}\right]\right)\right) = \frac{\beta}{\nu}\left(\tilde{w}_{t+1}^{j} - \tilde{w}_{t+1}^{i}\right) - \frac{1-\beta}{\nu}\kappa^{ij} + e_{t+1}^{ij}, \tag{15}$$

where mobility rates, $\tilde{\mu}_t^{ij}$, and wages, $\tilde{w}_{t+1}^{j}$ are the observed variables in the regression. In the case the econometrician abstracts from misclassification errors, $\tilde{\mu}_t^{ij}$ and $\tilde{w}_{t+1}^{j}$ can be directly computed from the data as is typically the case in the literature. With misclassification, mobility rates and wages must be corrected for misclassification.

Equation (15) is derived from a moment condition and, under the null hypothesis, the error term e_{t+1}^{ij} is independent of any variable in the information set of individuals up to time t. Assuming a known value for β, parameters κ and ν can be estimated using an instrumental variable regression using, for example, lagged values of wages and mobility rates as instruments. Table 5 shows estimated values of parameters. Column 2 shows estimates of κ and ν using mobility and wages obtained with original CPS industry data for the years 1981 to 2018, which abstract from misclassification. Similarly, column 3 shows estimates using the mobility and wage measures corrected for misclassification estimated before. The estimates using original coded data suggest large costs of switching industries, around 7

[38] Appendix B shows the the derivation of this equation from the model equilibrium conditions. For the estimation wages are normalized so the average wage in the sample is equal to one. Moreover, I assume $\beta = 0.97$. Note that, the continuation values, or expected future utilities, are not estimated. The intuition for this is closely connected to the "inversion" in Hotz and Miller (1993), where they show that one can use conditional choice probabilities to estimate structural parameters in this class of models. Finally, note that it is possible to propose other estimators for ν and κ, which may be useful in cases where workers flows between two industries are small or zero, such as estimation by Pseudo Poisson Maximum Likelihood.

times the mean wages, as mean wages are normalized to one.[39] The estimate for ν using original coded data suggest that idiosyncratic forces affecting workers' mobility decisions are also large, with a standard deviation about four times the average wage, since switching decisions are influenced by the difference between two idiosyncratic shocks.[40] However, note that the effective costs of switching for a worker are the differences between the cost κ and the realization of the idiosyncratic ϵ shock multiplied by parameter ν. The effective cost of switching faced by those that make a transition is much smaller than what the value of κ alone suggests.[41]

Note that the estimates of κ and ν in column 2 Table 5 are quite similar to those obtained by Artuç et al. (2010). Their preferred estimates in Table 3, panel IV, shows values for κ of 6.6 and ν of 1.88.[42]

As discussed in Section 3 and shown in Appendix A, misclassification biases the estimates of the costs κ towards zero. The estimates of κ in column 3 of Table 5, which use data corrected for misclassification, are fifty percent larger than those estimated using the original CPS codes in columns 2. In addition, the results of the table show that misclassification biases the estimates for ν also towards zero and the estimates using corrected data show values for ν that are approximately twenty percent larger than those obtained using the

[39] The standard errors of parameter estimates are obtained via block bootstrap. See Cameron and Trivedi (2005), page 377, for a description of the block bootstrap technique used here, also called panel bootstrap. This bootstrap relies on large N and short panel dimension. It treats each individual in the microdata as a block with many observations over time, and samples, with replacement, over blocks. In other words, I generate 500 random samples with replacement from the PSID and the CPS microdata. For each of these samples I generate a new set of estimates for α, mobility, wages and structural parameters, computing moments out of these 500 values. In this way, standard errors take into account the uncertainty from estimates used in second stages of the estimation, such as the use of α when correcting the CPS data, and the constructed regressors, such as wages and mobility matrices in the estimation of κ and ν. See Murphy and Topel (2002) for a discussion.

[40] The variance of the idiosyncratic shock ϵ multiplied by ν is $\nu^2 \frac{\pi^2}{6}$.

[41] It is important to highlight that the ϵ shocks "may be viewed as either a preference shock or a shock to the cost of moving, with no way to distinguish between the two." (Kennan & Walker, 2011, page 219) Then, while at first sight these cost appear very large, the costs effectively faced by individuals that actually move are substantially lower as these are the individuals that select on the most favorable realization of the shocks. That is, those that effectively move face much lower costs than average.

[42] Artuç et al. (2010) obtain additional estimates under different specifications and data adjustments.

Table 5: Estimates of structural parameters κ and ν

	Using original coding	Corrected for misclassification
Elasticity parameter, ν	1.77	2.16
	(0.06)	(0.24)
Cost of industry switching, κ	6.95	10.54
	(0.17)	(0.97)

Note: Estimated using equation (15) and yearly information about industry mobility and industry wages, assuming $\beta = 0.97$. Column 3 uses industry mobility and wages corrected for misclassification and estimated by EM algorithm using CPS data 1982-2018. See text for details. IV regressions instrumented using two-year lagged values of wages and mobility rates. Block bootstrap standard errors in parenthesis using 500 random samples with replacement.

original coding without adjustment.[43]

The estimation in Table 5 does not control for worker's characteristics. In Appendix H, I show that mobility patterns differ by workers characteristics, like age and education. Since the composition of workers has changed over time in the U.S. economy, the patterns of industry mobility may be affected by changes in the population. Moreover, the *observed* mobility flows of some groups of workers may be more affected by misclassification than that of others due to difference in industry choices by workers with different characteristics. In Appendix H, I estimate structural parameters that vary with workers' characteristics and the same conclusion applies: misclassification overstates industry mobility and bias parameter estimates. The estimates of κ and ν in Appendix H are lower that those in Table 5, both using original and corrected data.[44]

[43] A small fraction of observation in the CPS (1.2% in my final sample) has their industry information imputed. I also conducted the correction and estimation dropping these observations and the results are very similar to the ones here, with estimates of ν and κ only slightly larger.

[44] The estimated values of κ and ν in Table 5 are large when compared with average wages in the economy and this may be linked, in part, to the assumption of homogeneous workers, as the model is trying to explain the differences in average mobility rates across industries with differences in *average* industry wages, biasing upwards the estimates of ν. When controlling for worker heterogeneity, the differences in the wage of workers with different characteristics across industries and their patterns of mobility is important for the estimates of structural parameters as reflected in Appendix H.

5.2 Employment and welfare effects of a trade shock

While there is evidence of an important bias in parameter estimates and measures of mobility, the effect of this bias on the estimated employment and welfare effects of a trade shock is not immediately clear. To better understand this, in Appendix E, I perform different exercises using different versions of the trade model in Section 3 to compare the evolution of labor across industries and the impact on workers' welfare in the United States to a trade shock that permanently affects wages across sectors in an asymmetric way. Here I discuss the findings of exercise using a general equilibrium model of trade with costly labor reallocation and relegate the details and the result for the other exercises to the appendix.

I use a version of the model in Caliendo et al. (2019) with fewer countries and sectors and I engineer a trade shock as an increase in TFP in manufacturing the rest of the world. In particular, in period $t = 1$ agents learn that TFP for manufacturing goods from the rest of the world increases by 8% each period for 10 years. This captures the large increase in international trade of the recent decade and in particular the large expansion of exports from China into advanced economies since the mid-1990s.[45]

I compare the effects of this trade shock in two economies, one calibrated using labor allocations and mobility matrices obtained with originally coded data and with $\nu = 1.8$, and one calibrated using corrected data and with $\nu = 2.2$.[46] Figure 2 shows the evolution of the change in the manufacturing employment share, in the left panel, and the changes in period-one welfare due to the shock, in the right panel, for these two economies, which differ only in the estimates of parameter ν, initial mobility flows and initial employment shares.

[45] In the model, this shock doubles the amount of imports of the United States over this period, while at the same time U.S. exports also expand. The model allows for trade deficits and, overall, the trade deficit of the United States increases but less than the expansion in imports. All of these movements are consistent with the evolution of aggregate trade variables of the United States in the early 2000s, before the Great Recession.

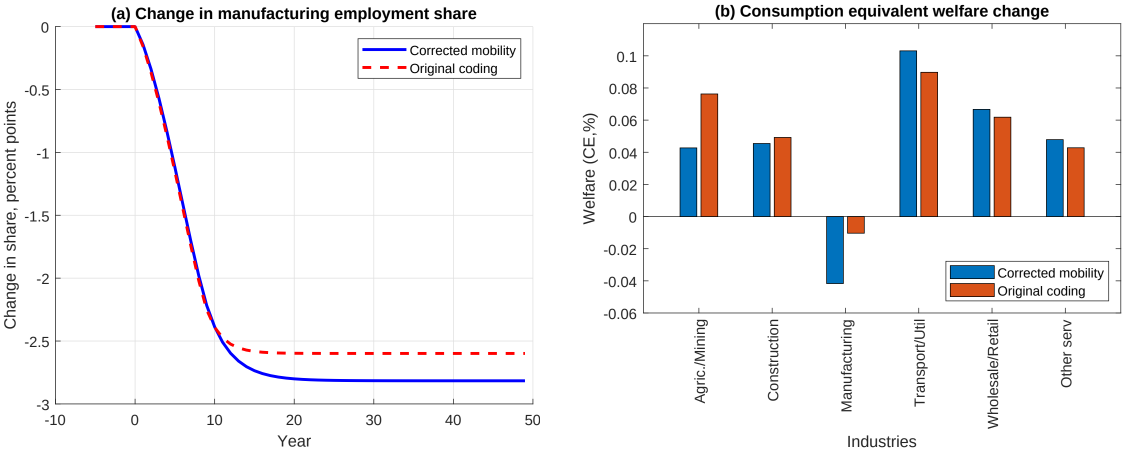
[46] In particular, I calibrate the initial value of the shares of employment by sector for the United States to those in the CPS data for the year 2000. Similarly, I use the matrix of workers' mobility by industry as estimated using CPS data for the year 2000. For one economy I use the data as originally coded and for the other I use the corrected values estimated previously. See Appendix E for further details.

The economy calibrated using corrected mobility information data displays larger effects on manufacturing employment over the long run than the economy that uses originally coded data.[47] In the initial years of the transition, differences in employment are small, but by year 20, misclassification errors magnify the effects on employment by 0.25 percentage points, or around a 10% difference of the total effect. Measures of welfare are also substantially affected by misclassification. Overall, consumption equivalent welfare for workers initially attached to manufacturing declines only mildly due to the trade shock in the economy calibrated with originally coded data. However, welfare falls more notably in the economy that uses corrected mobility measures, with the magnitude of the effect being roughly four times larger. The opposite is true for the other sectors of the economy, which show a larger increase in welfare in the economy that uses corrected mobility measures, with the exception of agriculture. These differences in the welfare effects are not only the result of differences in the evolution of wages, but also on the option value of moving (Artuç et al., 2010), which is closely linked to mobility rates.

While the previous section showed how parameter estimates are biased when using uncorrected data, the key insight from this exercises, and the ones in the appendix, is that the quantification of the *dynamic adjustment of labor* and the *welfare effects* of a trade shock are severely biased when using data with misclassification errors. Since many works in the trade literature using using publicly available data for the United States do not correct mobility flows, one should be cautious in taking the estimated magnitudes at face value.

[47] While the value of ν is larger when using corrected data, mobility flows also change and there is less mobility than using originally coded data. In the background, this would translate, for example, into higher estimates of mobility costs. If costs are higher, a permanent decline in wages in manufacturing will translate into more gradual outflows from manufacturing and fewer inflows into manufacturing, leading to a lower long-run employment share, as shown. Panel (b) in the figure provides additional intuitions for this. The decline in welfare is more pronounced in manufacturing using corrected data than originally coded data, and the model predicts that industries that deliver lower levels of lifetime utility will have lower levels of employment, explaining why the employment share in manufacturing is lower when using corrected data.

Figure 2: Effects of a trade liberalization in the general equilibrium model



Note: Evolution of the changes in the manufacturing employment share in the general equilibrium model with an increase in TFP in the rest of the world. $\nu = 1.8$ for original coding, and $\nu = 2.2$ for the data corrected for misclassification. In both cases the shock increases TFP by 8% per year for ten years in manufacturing in the rest of the world.

6 Conclusion

Misclassification errors are prevalent in industry and occupational mobility data and typically exaggerate worker flows across labor markets. The literature has studied this problem, focusing on its causes and possible ways to correct overall mobility measures. In this paper I focus on the *consequences* of misclassification errors for estimates of structural parameters and on the estimates of the effects of international trade on labor markets.

Many recent models of international trade with segmented labor markets and costly reallocation critically depend on information about workers' reallocation patterns. In this paper I show how misclassification bias estimates of structural parameter values in this class of models and propose an econometric framework to jointly estimate misclassification probabilities, corrected mobility matrices and wages, and structural parameters in a unified way. The method I propose is not unique to the PSID or the CPS and can be applied to correct industry and occupational mobility in other survey data.

My results show that estimated employment and welfare effects of a trade shock are different if the analysis uses originally coded data or corrected data. An important conclusion from this paper is that the estimates of the dynamic adjustment of labor and welfare effects of a trade shock are biased when using mobility flows that are subject to misclassification errors, raising an important warning for future research where workers' industry and occupation reallocation is a central part of the study.

In this paper I focused only on the role of industry (and occupation) misclassification in biasing parameter estimates and quantitative exercises. However, other well-known mechanisms, such as model misspecification, omission of relevant controls, and workers' unobserved characteristics, will also bias parameter estimates in usual ways studied in the literature, not necessarily specific to dynamic models of worker mobility and reallocation.

References

Abrevaya, J., & Hausman, J. A. (1999). Semiparametric estimation with mismeasured dependent variables: an application to duration models for unemployment spells. *Annales d'Economie et de Statistique*, 243–275.

Arcidiacono, P., & Miller, R. A. (2011). Conditional choice probability estimation of dynamic discrete choice models with unobserved heterogeneity. *Econometrica*, *79*(6), 1823–1867.

Arkolakis, C., Costinot, A., & Rodríguez-Clare, A. (2012). New trade models, same old gains? *American Economic Review*, *102*(1), 94–130.

Artuç, E., Chaudhuri, S., & McLaren, J. (2010). Trade shocks and labor adjustment: A structural empirical approach. *American Economic Review*, *100*(3), 1008–45.

Artuç, E., & McLaren, J. (2015). Trade policy and wage inequality: A structural analysis with occupational and sectoral mobility. *Journal of International Economics*, *97*(2), 278–294.

Bureau, U. C. (Ed.). (2015). *Current population survey interviewing manual.* Washington, D.C.: U.S. Dept. of the Census [i.e. Dept of Commerce], U.S. Census Bureau.

Caliendo, L., Dvorkin, M., & Parro, F. (2019). Trade and labor market dynamics: General equilibrium analysis of the China trade shock. *Econometrica*, *87*(3), 741–835.

Cameron, A. C., & Trivedi, P. K. (2005). *Microeconometrics: methods and applications.* Cambridge university press.

Carrillo-Tudela, C., & Visschers, L. (2023). Unemployment and endogenous reallocation over the business cycle. *Econometrica*, *91*(3), 1119–1153.

Dix-Carneiro, R. (2014). Trade liberalization and labor market dynamics. *Econometrica*, *82*(3), 825–885.

Eaton, J., & Kortum, S. (2002). Technology, geography, and trade. *Econometrica*, *70*(5), 1741–1779.

Feng, S., & Hu, Y. (2013). Misclassification errors and the underestimation of the U.S. unemployment rate. *American Economic Review*, *103*(2), 1054–70.

Flood, S., King, M., Rodgers, R., Ruggles, S., Warren, J. R., & Westberry, M. (2021). *Integrated Public Use Microdata Series, Current Population Survey: Version 9.0 [dataset]. Minneapolis, MN: IPUMS, 2021.* https://doi.org/10.18128/D030.V9.0.

Fujita, S., Moscarini, G., & Postel-Vinay, F. (2020). *Measuring employer-to-employer reallocation* (Tech. Rep.). National Bureau of Economic Research.

Hausman, J. A., Abrevaya, J., & Scott-Morton, F. M. (1998). Misclassification of the dependent variable in a discrete-response setting. *Journal of econometrics*, *87*(2), 239–269.

Hotz, V. J., & Miller, R. A. (1993). Conditional choice probabilities and the estimation of dynamic models. *The Review of Economic Studies*, *60*(3), 497–529.

Kambourov, G., & Manovskii, I. (2008). Rising occupational and industry mobility in the United States: 1968–97. *International Economic Review*, *49*(1), 41–79.

Kambourov, G., & Manovskii, I. (2013). A cautionary note on using (March) Current

Population Survey and Panel Study of Income Dynamics data to study worker mobility. *Macroeconomic Dynamics*, *17*(1), 172–194.

Kennan, J., & Walker, J. R. (2011). The effect of expected income on individual migration decisions. *Econometrica*, *79*(1), 211–251.

Moscarini, G., & Thomsson, K. (2007). Occupational and job mobility in the us. *Scandinavian Journal of Economics*, *109*(4), 807–836.

Murphy, K. M., & Topel, R. H. (1987). Unemployment, risk, and earnings: Testing for equalizing wage differences in the labor market. In K. M. Lang & J. S. Leonard (Eds.), *Unemployment and the structure of labor markets* (pp. 103–140). Basil Blackwell.

Murphy, K. M., & Topel, R. H. (2002). Estimation and inference in two-step econometric models. *Journal of Business & Economic Statistics*, *20*(1), 88–97.

Neal, D. (1999). The complexity of job mobility among young men. *Journal of Labor Economics*, *17*(2), 237–261.

Poterba, J., & Summers, L. (1986). Reporting errors and labor market dynamics. *Econometrica*, 1319–1338.

PSID. (1999). *A Panel Study of Income Dynamics: 1968-1980 retrospective occupation-industry files documentation.* Survey Research Center. Institute for Social Research. The University of Michigan. Accessed from https://data.nber.org/psid/supp/occind.txt.

Simonovska, I., & Waugh, M. E. (2014). The elasticity of trade: Estimates and evidence. *Journal of International Economics*, *92*(1), 34–50.

Timmer, M. P., Dietzenbacher, E., Los, B., Stehrer, R., & De Vries, G. J. (2015). An illustrated user guide to the world input–output database: the case of global automotive production. *Review of International Economics*, *23*(3), 575–605.

Traiberman, S. (2019). Occupations and import competition: Evidence from Denmark. *American Economic Review*, *109*(12), 4260–4301.

Vom Lehn, C., Ellsworth, C., & Kroff, Z. (2020). Reconciling occupational mobility in the current population survey. *IZA Discussion Paper*.

A Estimation bias

The literature has studied the effects of misclassification errors on parameter estimates and found that estimators that abstract from misclassification errors in the data lead to inconsistent estimates of population parameters (Abrevaya & Hausman, 1999; Hausman, Abrevaya, & Scott-Morton, 1998). Therefore, it is not surprising that in the examples presented in Section 3, parameters κ and ν are biased due to misclassification. However, the *direction* of the bias needs additional discussion.

In this appendix I use a simple version of the model developed in Section 3 to show how misclassifiaction errors bias parameter estimates. In particular, I assume an economy with two industries, labeled A and B, where wages are higher in industry B. In addition, to simplify some expressions, I assume that the discount factor is very close to one and that the probability of being classified in a different industry is equal to ψ, which is identical across industries, and not too large, such that errors in the data are not excessively large, which is the empirically relevant case as discussed in Section 2 and Appendix D. Thus, $(1-\psi)$ is the probability of being correctly classified in the industry of work.

Structural parameters with no misclassification errors can be estimated using equation (19), which is derived in Appendix B from the model's equilibrium conditions, and extends equation (5) by allowing variables to fluctuate over time. With misclassification, in the case of two industries and β close to one, we have,[48]

$$\log\left(\tilde{\mu}_{t+1}^{BB}\right) - \log\left(\tilde{\mu}_{t}^{AA}\right) = \tilde{\mathbb{C}}_1 + \tilde{\mathbb{C}}_2\left(\tilde{w}_{t+1}^{B} - \tilde{w}_{t+1}^{A}\right) + e_{t+1},$$

where the variables with a tilde denote the observed variables in the data, which differ from the actual variables (without a tilde) due to misclassification, and e_{t+1} is an error that is uncorrelated with wages.[49]

Regression coefficients under no misclassification are functions of the true structural parameters: $\mathbb{C}_1 = -(1-\beta)\frac{\kappa}{\nu}$ and $\mathbb{C}_2 = \frac{\beta}{\nu}$. Note that estimates $\tilde{\mathbb{C}}_1$, and $\tilde{\mathbb{C}}_2$ obtained using observed (uncorrected) data may differ from $\mathbb{C}_1$, and $\mathbb{C}_2$ if misclassification induces a bias in parameter estimates.

By OLS, regression coefficients are,

$$\tilde{\mathbb{C}}_2 = \frac{Cov\left(\log\left(\tilde{\mu}_{t+1}^{BB}\right) - \log\left(\tilde{\mu}_{t}^{AA}\right), \left(\tilde{w}_{t+1}^{B} - \tilde{w}_{t+1}^{A}\right)\right)}{Var\left(\tilde{w}_{t+1}^{B} - \tilde{w}_{t+1}^{A}\right)}; \quad \tilde{\mathbb{C}}_1 = E\left[\log\left(\tilde{\mu}_{t+1}^{BB}\right) - \log\left(\tilde{\mu}_{t}^{AA}\right)\right] - \tilde{\mathbb{C}}_2 E\left[\tilde{w}_{t+1}^{B} - \tilde{w}_{t+1}^{A}\right].$$

Thus, to understand how misclassification affects parameter estimates, we need to analyze how the different moments in these expressions change relative to the case of no misclassification. Using equation (4) in the case of two sectors, we have,

$$\begin{aligned}
\tilde{w}_{t+1}^{A} &= w_{t+1}^{A}\left(1-\psi\,\pi_{t+1}^{A}\right) + w_{t+1}^{B}\,\psi\,\pi_{t+1}^{A} \\
\tilde{w}_{t+1}^{B} &= w_{t+1}^{B}\left(1-\psi\,\pi_{t+1}^{B}\right) + w_{t+1}^{A}\,\psi\,\pi_{t+1}^{B},
\end{aligned}$$

[48] Technically, β cannot be equal to one as a solution to the worker's problem with infinite horizon would not exist. In this Appendix, I make this assumption to simplify the expressions needed to build intuitions about the source of the estimation bias. In no other part of the paper or the quantitative application I use this assumption. With a finite horizon, i.e. finite worklife, the assumption of $\beta = 1$ poses no problems.

[49] In this example I assume wages fluctuate exogenously over time. However, the proof in this Appendix does not rely on time-variation in wages and the analysis could be done for a stationary economy. In particular, imagine we have many different two-sector economies that have the same underlying structural parameters κ and ν, but differ in their wages (due to differences in labor demand, for example). Then, the regression would have many observations using the cross-section of stationary equilibria in each of these economies.

where $\pi^A_{t+1} = \frac{L^B_{t+1}}{(1-\psi)\, L^A_{t+1} + \psi\, L^B_{t+1}}$, and $\pi^B_{t+1} = \frac{L^A_{t+1}}{(1-\psi)\, L^B_{t+1} + \psi\, L^A_{t+1}}$, and ψ is the misclassification probability.

Then,

$$\tilde{w}^B_{t+1} - \tilde{w}^A_{t+1} = (w^B_{t+1} - w^A_{t+1}) - \psi(w^B_{t+1} - w^A_{t+1})\left[\pi^B_{t+1} + \pi^A_{t+1}\right],$$

and the difference in observed wages, $\tilde{w}^B_{t+1} - \tilde{w}^A_{t+1}$, is smaller than the difference in actual wages, $w^B_{t+1} - w^A_{t+1}$. We can define $\zeta_{t+1} = -\psi\,(w^B_{t+1} - w^A_{t+1})[\pi^B_{t+1} + \pi^A_{t+1}]$ as the measurement error in the regressor. Changes in wages over time induce changes in the employment shares L^A_{t+1} and L^B_{t+1}, thus, characterizing the variance in closed-form is not straightforward. However, note that an increase in w^B_{t+1} will increase π^A_{t+1} *and reduce* π^B_{t+1}, making the total change in $[\pi^B_{t+1} + \pi^A_{t+1}]$ small.[50] Then, to a first order, the term in squared brackets does not change much and we can approximate,

$$Var(\tilde{w}^B_{t+1} - \tilde{w}^A_{t+1}) \approx \left(1 - \psi[\pi^B_{t+1} + \pi^A_{t+1}]\right)^2 Var\left(w^B_{t+1} - w^A_{t+1}\right).$$

Thus, the variance of the difference in observed wages is *smaller* than the variance in actual wages.[51]

I now analyze the covariance term in the numerator of the regression coefficient. Using equation (3), we have that,

$$\tilde{\mu}^{BB}_{t+1} = \frac{L^B_{t+1}\,(1-\psi)^2\,\mu^{BB}_{t+1}}{\tilde{L}^B_{t+1}} + \frac{L^A_{t+1}\,(1-\psi)\psi\,\mu^{AB}_{t+1}}{\tilde{L}^B_{t+1}} + \frac{L^A_{t+1}\,\psi^2\,\mu^{AA}_{t+1}}{\tilde{L}^B_{t+1}} + \frac{L^B_{t+1}\,(1-\psi)\psi\,\mu^{BA}_{t+1}}{\tilde{L}^B_{t+1}}.$$

Since ψ is not very large, the first term in the expression is quantitatively the most important,

[50] Note that $[\pi^B_{t+1} + \pi^A_{t+1}] = \frac{\tilde{L}^A_{t+1} L^A_{t+1} + L^B_{t+1}\tilde{L}^B_{t+1}}{\tilde{L}^A_{t+1}\tilde{L}^B_{t+1}}$, and in a two sector economy, $L^B_{t+1} = (1 - L^A_{t+1})$ and $\tilde{L}^B_{t+1} = (1 - \tilde{L}^A_{t+1})$. Thus, an increase in, say, w^B_{t+1} will increase employment shares L^B_{t+1} and $\tilde{L}^B_{t+1}$, and reduce them in the other industry, leaving $[\pi^B_{t+1} + \pi^A_{t+1}]$ relatively unchanged.

[51] Contrary to the "classical" case of measurement error in the regressor, here the lower variance does not lead to an attenuation bias, but to a magnification as the measurement error is strongly negatively correlated with the regressor.

thus,

$$\log\left(\tilde{\mu}_{t+1}^{BB}\right) - \log\left(\tilde{\mu}_{t}^{AA}\right) \approx \log\left(\mu_{t+1}^{BB}\right) - \log\left(\mu_{t}^{AA}\right) + \log\left(\frac{L_{t+1}^{B}}{\tilde{L}_{t+1}^{B}}\right) - \log\left(\frac{L_{t}^{A}}{\tilde{L}_{t}^{A}}\right).$$

Using this expression, the relationship between covariance of observed variables and actual is,

$$\begin{aligned} Cov\left(\log\left(\tilde{\mu}_{t+1}^{BB}\right) - \log\left(\tilde{\mu}_{t}^{AA}\right), \left(\tilde{w}_{t+1}^{B} - \tilde{w}_{t+1}^{A}\right)\right) \quad \approx \quad & Cov\left(\log\left(\mu_{t+1}^{BB}\right) - \log\left(\mu_{t}^{AA}\right), \left(w_{t+1}^{B} - w_{t+1}^{A}\right)\left[1 - \psi[\pi_{t+1}^{B} + \pi_{t+1}^{A}]\right]\right) + \\ & Cov\left(\log\left(\frac{L_{t+1}^{B}}{\tilde{L}_{t+1}^{B}}\right) - \log\left(\frac{L_{t}^{A}}{\tilde{L}_{t}^{A}}\right), \left(w_{t+1}^{B} - w_{t+1}^{A}\right)\left[1 - \psi[\pi_{t+1}^{B} + \pi_{t+1}^{A}]\right]\right). \end{aligned}$$

Therefore, assuming that $[\pi_{t+1}^{B} + \pi_{t+1}^{A}]$ is constant, as discussed before, then,

$$\tilde{\mathbb{C}}_2 \approx \frac{1}{(1 - \psi[\pi^{B} + \pi^{A}])}\mathbb{C}_2 + \frac{1}{(1 - \psi[\pi^{B} + \pi^{A}])}\frac{Cov\left(\log\left(\frac{L_{t+1}^{B}}{\tilde{L}_{t+1}^{B}}\right) - \log\left(\frac{L_{t}^{A}}{\tilde{L}_{t}^{A}}\right), \left(w_{t+1}^{B} - w_{t+1}^{A}\right)\right)}{Var\left(w_{t+1}^{B} - w_{t+1}^{A}\right)}. \tag{16}$$

The first term in the right is larger than $\mathbb{C}_2$ and the second term in the right is positive since an increase in w_{t+1}^{B} increases L_{t+1}^{B} more than $\tilde{L}_{t+1}^{B}$, and, in future periods, reduces L_{t}^{A} more than $\tilde{L}_{t}^{A}$. Thus, the estimate of $\tilde{\mathbb{C}}_2$ using uncorrected data is larger than $\mathbb{C}_2$ and estimates are biased upwards, leading to a downward bias for the estimate of ν. Note that, with no measurement error, $\psi = 0$, $L_{t}^{A} = \tilde{L}_{t}^{A}$ and $L_{t+1}^{B} = \tilde{L}_{t+1}^{B}$, thus there is no bias and $\tilde{\mathbb{C}}_2 = \mathbb{C}_2$.

Using the previous expressions and manipulating, we can write,

$$\tilde{\mathbb{C}}_1 \approx \mathbb{C}_1 + E\left[\log\left(\frac{L_{t+1}^{B}}{\tilde{L}_{t+1}^{B}}\right) - \log\left(\frac{L_{t}^{A}}{\tilde{L}_{t}^{A}}\right)\right] - \frac{Cov\left(\log\left(\frac{L_{t+1}^{B}}{\tilde{L}_{t+1}^{B}}\right) - \log\left(\frac{L_{t}^{A}}{\tilde{L}_{t}^{A}}\right), \left(w_{t+1}^{B} - w_{t+1}^{A}\right)\right)}{Var\left(w_{t+1}^{B} - w_{t+1}^{A}\right)} E\left(w_{t+1}^{B} - w_{t+1}^{A}\right),$$

where the last two terms form the expression of the constant term in a regression between $\log\left(\frac{L_{t+1}^{B}}{\tilde{L}_{t+1}^{B}}\right) - \log\left(\frac{L_{t}^{A}}{\tilde{L}_{t}^{A}}\right)$ and $(w_{t+1}^{B} - w_{t+1}^{A})$, which is positive given that $E[w_{t+1}^{B} - w_{t+1}^{A}] > 0$ implies that $E[L_{t+1}^{B} - L_{t+1}^{A}] > 0$.[52] Thus, the estimate for $\tilde{\mathbb{C}}_1$ is biased upwards. Then, the

[52] Otherwise, if the intercept was negative, there would be values for which $E[w_{t+1}^{B} - w_{t+1}^{A}] > 0$ and $E[L_{t+1}^{B} - L_{t+1}^{A}] < 0$, which is a contradiction given the structure of the model.

estimated value for the switching cost $\tilde{\kappa} = \frac{(-\tilde{\mathbb{C}}_1)\tilde{\nu}}{1-\beta}$, depends on $(-\tilde{\mathbb{C}}_1)$ and $\tilde{\nu}$, both biased downwards relative to their actual value. Therefore, the estimate for the switching costs κ is biased towards zero as well. Note that with no measurement error there is no bias, as expected.[53]

In sum, in this Appendix I show that in a simple version of the model and under some reasonable parameter restrictions, the estimates of the structural parameters ν and κ using uncorrected data are downward biased relative to their true value. The estimation in Section 5 using a richer model points to biases in the same direction as the ones detailed here. In a model with a richer structure and more sectors, parameters will still be biased due to misclassification, but the direction of the bias would be harder to characterize.

These result highlights why it is important not only to correct mobility measures, but also wages, as uncorrected wage data suffers from measurement error which would induce a bias in estimates, even if using corrected mobility data.

B Estimation equation

In this Appendix I derive equation (5) used to estimate structural parameters in a model with no misclassification following Artuç et al. (2010). Combining equations (1) and (2), we get,

$$v_t^i(\tau) = w_t^i(\tau) + \beta E_t\left[v_{t+1}^j(\tau)\right] - \kappa^{ij}(\tau) - \nu \log\left[\mu_t^{ij}(\tau)\right], \tag{17}$$

which holds for all i and j. As in Artuç et al. (2010), this does not assume perfect foresight and individual form expectations about future variables. Taking the difference of values when industry j is chosen and when industry i is chosen, we get,

$$0 = \beta E_t\left[v_{t+1}^j(\tau) - v_{t+1}^i(\tau)\right] - \kappa^{ij}(\tau) - \nu\left(\log\left[\mu_t^{ij}(\tau)\right] - \log\left[\mu_t^{ii}(\tau)\right]\right).$$

[53] Note that, even under the assumption that misclassification does not affect wages, it is straightforward to adjust equation (16) and show that parameters will still be biased in the same direction.

Use (17) one period forward for $v^j_{t+1}(\tau)$ and $v^i_{t+1}(\tau)$, where the future industry choice in both cases is industry j and taking differences, we get,

$$v^j_{t+1}(\tau) - v^i_{t+1}(\tau) = w^j_{t+1}(\tau) - w^i_{t+1}(\tau) + \kappa^{ij}(\tau) - \nu\left(\log\left[\mu^{jj}_{t+1}(\tau)\right] - \log\left[\mu^{ij}_{t+1}(\tau)\right]\right).$$

Combining the last two equations, we obtain,

$$E_t\left[\frac{\beta}{\nu}\left(w^j_{t+1}(\tau) - w^i_{t+1}(\tau)\right) - \frac{1-\beta}{\nu}\kappa^{ij}(\tau) + \beta\left(\log\left[\mu^{ij}_{t+1}(\tau)\right] - \log\left[\mu^{jj}_{t+1}(\tau)\right]\right) - \left(\log\left[\mu^{ij}_t(\tau)\right] - \log\left[\mu^{ii}_t(\tau)\right]\right)\right] = 0, \tag{18}$$

As argued by Artuç et al. (2010), rational expectations imply that this is a conditional moment that holds when conditioning on any variable known at time t. In particular, parameters can be estimated via the following regression,

$$\left(\log\left[\mu^{ij}_t(\tau)\right] - \log\left[\mu^{ii}_t(\tau)\right] - \beta\left(\log\left[\mu^{ij}_{t+1}(\tau)\right] - \log\left[\mu^{jj}_{t+1}(\tau)\right]\right)\right) = \frac{\beta}{\nu}\left(w^j_{t+1}(\tau) - w^i_{t+1}(\tau)\right) - \frac{1-\beta}{\nu}\kappa^{ij}(\tau) + e^{ij}_{t+1}, \tag{19}$$

where e^{ij}_{t+1} is a rational expectations error, orthogonal to any variable know at time t. This expression can be estimated using an instrumental variables regression using, for example, lagged values of wages and mobility rates as instruments.

Amenities.– If different industries provide different amenities to workers, as the type of jobs that workers perform may be more or less stressful or more or less prestigious, for example, workers will demand a compensating differential to work in industries with a lower level of amenities, all else equal. I assume amenities of industry i, $A^i(\tau)$, are time-invariant and enter additively in the workers period utility. Then, the problem of the worker becomes,

$$v^i_t(\tau) = w^i_t(\tau) + A^i(\tau) + \beta E_t\left[v^j_{t+1}(\tau)\right] - \kappa^{ij}(\tau) - \nu\log\left[\mu^{ij}_t(\tau)\right].$$

Following the same steps as before, we can write the estimation equation as,

$$\left(\log\left[\mu_t^{ij}(\tau)\right] - \log\left[\mu_t^{ii}(\tau)\right] - \beta\left(\log\left[\mu_{t+1}^{ij}(\tau)\right] - \log\left[\mu_{t+1}^{jj}(\tau)\right]\right)\right) = \frac{\beta}{\nu}\left(w_{t+1}^{j}(\tau) - w_{t+1}^{i}(\tau)\right) + \\ + \left[\frac{\beta}{\nu}\left(A^{j}(\tau) - A^{i}(\tau)\right) - \frac{1-\beta}{\nu}\kappa^{ij}(\tau)\right] + e_{t+1}^{ij}.$$

Unfortunately, using this estimating equation and information only on mobility and wages, as in Artuç et al. (2010), it is not possible to identify separately amenities and the switching costs that vary by origin and destination. If we impose that $\kappa^{ij}(\tau) = \kappa(\tau)$ for $i \neq j$, and normalize amenity values for one industry, then it is possible to estimate amenities. However, note that fixed effects in the regression would not only be recovering amenities but also the average level of wages in the industry. Let $\bar{w}^{j}(\tau)$ be average wages of industry j over time, and $\hat{w}_{t+1}^{j}(\tau) = w_{t+1}^{j}(\tau) - \bar{w}^{j}(\tau)$, we can re-write the equation as,

$$\left(\log\left[\mu_t^{ij}(\tau)\right] - \log\left[\mu_t^{ii}(\tau)\right] - \beta\left(\log\left[\mu_{t+1}^{ij}(\tau)\right] - \log\left[\mu_{t+1}^{jj}(\tau)\right]\right)\right) = \frac{\beta}{\nu}\left(\hat{w}_{t+1}^{j}(\tau) - \hat{w}_{t+1}^{i}(\tau)\right) + \\ + \left[\frac{\beta}{\nu}\left(A^{j}(\tau) - \bar{w}^{j}(\tau) - (A^{i}(\tau) - \bar{w}^{i}(\tau))\right) - \frac{1-\beta}{\nu}\kappa^{ij}(\tau)\right] + e_{t+1}^{ij}.$$

Two points are important. First, note that the first stage of the estimation procedure that obtains measures of wages, mobility rates and share of labor by industry corrected from misclassification is not affected by this type of model misspecification. Second, whether the structural model is misspecified, in this case by potentially omitting amenities, does not overturn the main message of the paper, that misclassification errors, which are a prevalent feature of the data, introduce errors in observed wages, mobility rates and the share of labor by industry, leading to biases structural parameters and the estimated effects of international trade.

Time variation.– Now, assume that the cost of switching industries varies over time in the following way. Parameter κ is constant over several years and suddenly it changes in an unanticipated way. For example, at the beginning of each decade, workers are "surprised" by change in this parameter and expect to remain at that new level for all periods in the future. Under this assumption, I can use equation (15) but allowing for different constant terms for each decade. At the end of each decade, the expected value of next year's wages and mobility would differ from the actual value due to this change in κ, but since the change is unanticipated, it will be captured in the expectation error term.

Evolving workers' characteristics: aging.– Moreover, it is easy to incorporate differences in age. Artuç et al. (2010) extend the model for workers that age probabilistically from young to old. But the expression for deterministic aging is simpler. Let some of the characteristics of workers evolve deterministically over time and denote these characteristics by a, like age. Then, we can write the estimating equation as,

$$\begin{aligned}\left(\log\left[\mu_t^{ij}(\tau,a)\right] - \log\left[\mu_t^{ii}(\tau,a)\right] - \beta\left(\log\left[\mu_{t+1}^{ij}(\tau,a+1)\right] - \log\left[\mu_{t+1}^{jj}(\tau,a+1)\right]\right)\right) &= \\ = \frac{\beta}{\nu}\left(w_{t+1}^{j}(\tau,a+1) - w_{t+1}^{i}(\tau,a+1)\right) - \frac{\left(\kappa^{ij}(\tau,a) - \beta\,\kappa^{ij}(\tau,a+1)\right)}{\nu} + e_{t+1}^{ij}.\end{aligned}$$

Clearly, the value function for the worker may be different the last period of her (work)life, but to the extent that the estimation does not use that period, the equations holds.

Heterogeneous switching costs.– Finally, I discuss how to identify and estimate parameters for the case of heterogeneous switching costs, κ, that vary by origin and destination. It is clear from equation (19) that observables in the left-hand side and the right-hand side vary by origin-destination pairs, ij, and by time. Much of the discussion in Section 3 assumed a single time period in a stationary environment. In this case, if κ varies by origin and destination, parameters ν and κ cannot be separately identified. However, with at least two periods of data and under the assumption of constant parameters (sometimes referred to as constant fundamentals), we can identify costs that vary by origin and destination using a stacked-version of equation (19) over two periods, with origin-destination fixed effects.

I now extend the result of Section 3.1 and show how misclassification errors affect the estimates of the switching costs. As in that section, I use two set of parameters: $\nu = 1.8$ and $\nu = 2.2$. The two set of parameters for the matrix of switching costs are randomly drawn from a uniform distribution with mean 7 and 10.0, respectively, and an upper and lower limit that is ±12.5% relative to the mean. Wages are the same as those assumed in Section 3.1, except that in the second period wages in sector 1 increase by 0.1 and the sector 6 decrease by 0.1.[54]

In the absence of misclassification error, the regression with origin-destination fixed effects recovers the parameters perfectly. Table 6 shows what happen with the estimates with a moderate level of misclassification error ($\psi = 0.005$). As in the case of homogeneous switching costs, in this example the estimates of ν and κ are biased downwards due to the errors, with estimated values for ν around 50% lower than the true values and κ between 50% to 67% lower, depending on the case. In this way, the main conclusions on the direction of the bias of Section 3.1 extend to the case of heterogeneous switching costs, κ, that vary by origin and destination.

C Identification

I now show how parameters $\alpha_{j|s}(\tau)$, $p_s(\tau, t)$, and $p_{s'|s}(\tau, t)$ are identified from moments in the data. Recall that $p_s(\tau, t)$ is the share of individuals with characteristics τ actually employed in industry s in period t, and that $p_{s'|s}(\tau, t)$ is the share of workers that switch to industry s' in period t conditional on being employed in industry s the period before. By Assumption 2, these variables are differentiable functions of time, and by Assumption 1, misclassification probability does not vary with time and depends on the workers characteristics and the actual industry of employment in the same period as the observed industry.

[54] Note that all that is needed is some change in wages over time to estimate cost parameters by origin and destination. Given the parametric assumptions specified in the model and the assumption that parameters are constant over time, parameters are identified.

Table 6: The effect of misclassification on parameter estimates for heterogeneous switching costs

Panel A

	True parameter values - true value of $\nu = 1.8$						estimates under misclassification - estimated $\nu = 1.2$					
	Ag./Min.	Const.	Manuf.	Trans./Util	W & Retail	Other	Ag./Min.	Const.	Manuf.	Trans./Util	W & Retail	Other
Agric./Mining	-	6.2	7.0	7.4	7.3	6.8	-	4.0	4.1	4.5	4.6	4.2
Construction	7.1	-	7.6	6.7	6.5	7.8	3.6	-	3.8	3.8	3.9	4.2
Manufacturing	6.3	6.5	-	7.4	7.1	7.0	3.6	4.0	-	4.4	4.3	4.3
Transport/Util	6.9	7.7	6.6	-	6.5	7.1	3.6	4.3	3.6	-	3.9	4.1
Wholesale/Retail	6.7	7.6	6.4	7.2	-	7.1	3.4	4.1	3.3	3.9	-	3.8
Other serv	7.1	7.7	6.6	7.5	7.8	-	3.6	4.1	3.5	4.1	4.2	-

Panel B

	True parameter values - true value of $\nu = 2.2$						estimates under misclassification - estimated $\nu = 1.1$					
	Ag./Min.	Const.	Manuf.	Trans./Util	W & Retail	Other	Ag./Min.	Const.	Manuf.	Trans./Util	W & Retail	Other
Agric./Mining	-	8.8	10.1	10.6	10.4	9.7	-	4.2	4.2	4.7	4.9	4.5
Construction	10.1	-	10.8	9.6	9.3	11.1	3.7	-	3.9	4.0	4.3	4.3
Manufacturing	9.0	9.3	-	10.6	10.2	9.9	3.6	4.1	-	4.6	4.6	4.5
Transport/Util	9.8	11.1	9.5	-	9.3	10.1	3.6	4.3	3.8	-	4.2	4.3
Wholesale/Retail	9.5	10.9	9.1	10.3	-	10.1	3.2	4.0	3.2	3.8	-	3.7
Other serv	10.1	11.0	9.4	10.7	11.1	-	3.4	4.0	3.6	4.1	4.3	-

Note: Parameters estimated by OLS according to equation (19). Each observation is a sector of origin-destination, where origin is different from destination. The example uses 6 sectors, two different time periods with different wages. There are 60 observations in each regression. Panel A shows results for structural parameter values $\nu = 1.8$ and parameters κ as listed on the left matrix in the panel. Misclassification error computed as in Section 3 with $\psi = 0.005$. Panel B shows results for $\nu = 2.2$ and parameters κ as listed on the left matrix in the panel.

C.1 Identification using original and retrospective industry coding in the PSID.

Identification of the misclassification probabilities, α, relies on the availability of original and retrospective coding before 1980 in the PSID data, as discussed in Section 2. Following Kambourov and Manovskii (2008), the assumption is that retrospective coding is free of misclassification errors. Intuitively, the sample before 1980 contains all the information needed to identify not only misclassification probabilities, but also mobility rates and employment shares as, for each individual, we have information on the originally coded industry and the retrospective codes.

Using the PSID sample up to the year 1980 and data on retrospective coding, we can construct the log-likelihood of the sample as follows,

$$\begin{aligned}
\tilde{\mathcal{L}}(\tau^i, d^i, \omega^i | p_{s,t}(\tau), p_{s'|s,t}(\tau), w_{s,t}(\tau)) \quad = \quad & \sum_{i=1}^{N}\sum_{j_1=1}^{J}\sum_{j_2=1}^{J}\cdots\sum_{j_T=1}^{J}\sum_{s_1=1}^{J}\sum_{s_2=1}^{J}\cdots\sum_{s_T=1}^{J} d^{i,*}_{j_1}\times d^{i,*}_{j_2|j_1}\times\ldots\times d^{i,*}_{j_T|j_{T-1}}\times \\
& d^{i}_{s_1}\times d^{i}_{s_2|s_1}\times\ldots\times d^{i}_{s_T|s_{T-1}}\times \Bigg[\log\left[p_{s_1}(\tau^i_1) f(\log(\omega^n_{s_1}/w_{s_1}(\tau^i_1)))\right] + \\
& \log\left[p_{s_2|s_1}(\tau^i_2) f(\log(\omega^n_{s_2}/w_{s_2}(\theta,\tau^i_2)))\right] + \ldots + \\
& \log\left[p_{s_T|s_{T-1}}(\tau^i_T) f(\log(\omega^n_{s_T}/w_{s_T}(\theta,\tau^i_T)))\right] \Bigg]
\end{aligned}$$

where $d^i_{s'|s,t+1}$ and $d^{i,*}_{j'|j,t+1}$ are dummy variables that are equal to one if individual i is retrospectively recorded as employed in industry s' at $t+1$ and s at t and originally coded in industry j and j', respectively, and zero otherwise. Then, maximum likelihood estimators are simply,

$$\begin{aligned}
p_s(\tau,t) \quad &= \quad \frac{\sum_{i=1}^{N}\sum_{j_1=1}^{J}\sum_{j_2=1}^{J}\cdots\sum_{j_T=1}^{J} d^{i,*}_{j_1}\times d^{i,*}_{j_2|j_1}\times\ldots\times d^{i,*}_{j_T|j_{T-1}}\times d^i_{s,t}\,\mathbb{I}(\tau^i=\tau)}{\sum_{i=1}^{N}\sum_{j_1=1}^{J}\sum_{j_2=1}^{J}\cdots\sum_{j_T=1}^{J}\sum_{s=1}^{J} d^{i,*}_{j_1}\times d^{i,*}_{j_2|j_1}\times\ldots\times d^{i,*}_{j_T|j_{T-1}}\times d^i_{s,t}\,\mathbb{I}(\tau^i=\tau)}, \\
p_{s'|s}(\tau,t+1) \quad &= \quad = \frac{\sum_{i=1}^{N}\sum_{j_1=1}^{J}\sum_{j_2=1}^{J}\cdots\sum_{j_T=1}^{J} d^{i,*}_{j_1}\times d^{i,*}_{j_2|j_1}\times\ldots\times d^{i,*}_{j_T|j_{T-1}}\times d^i_{s'|s,t+1}\,\mathbb{I}(\tau^i=\tau)}{\sum_{i=1}^{N}\sum_{j_1=1}^{J}\sum_{j_2=1}^{J}\cdots\sum_{j_T=1}^{J}\sum_{s'=1}^{J} d^{i,*}_{j_1}\times d^{i,*}_{j_2|j_1}\times\ldots\times d^{i,*}_{j_T|j_{T-1}}\times d^i_{s'|s,t+1}\,\mathbb{I}(\tau^i=\tau)},
\end{aligned}$$

where $\mathbb{I}(\tau^i = \tau)$ is an indicator equal to one if worker i's characteristics are equal to τ. These expressions show that parameters $p_{s'|s,t+1}$ and $p_{s,t}$ are identified from the population moments, $\mathbb{E}\left[d^i_{s,t}|\tau, t\right]$, $\mathbb{E}\left[d^i_{s'|s,t+1}|\tau, t\right]$ in the data. Given assumptions for f, identification of parameters related to that density and wages is straightforward. For example, in this work I assume that f is the normal density, and I parametrize $w_{s,t}(\tau)$ as linearly in the elements of τ and a fixed effect for s, in which case, the maximum likelihood estimator coincides with OLS.

It is possible to identify misclassification probabilities, $\alpha_{j|s}$, using retrospective and originally coded data directly. The maximum likelihood estimator leads to the moment condition $\alpha_{j|s} = \frac{\mathbb{E}\left[d^{i,*}_{j,t} d^i_{s,t}|\tau\right]}{\mathbb{E}\left[d^i_{s,t}|\tau\right]}$. However, the PSID sample for the years 1968-1980 is small and this can affect the estimates of $\alpha_{j|s}$, particularly since employment in some industries is very small, leading to very few (or no) observations with $d^{i,*}_{j,t} d^i_{s,t} = 1$ for some j and s.

An alternative way to estimate $\alpha_{j|s}$ uses only originally coded data and the parameters previously estimated.[55] Define the likelihood as,

$$\begin{aligned}\mathcal{L}(\tau^i, d^i, \omega^i | p_{s,t}(\tau), p_{s'|s,t}(\tau), w_{s,t}(\tau), \alpha_{j|s}) \quad = \\ \sum_{s_1=1}^{J} \sum_{s_2=1}^{J} \cdots \sum_{s_T=1}^{J} \prod_{j_1=1}^{J} \prod_{j_2=1}^{J} \cdots \prod_{j_T=1}^{J} \Bigg[\left[p_{s_1}(\tau^n_1, \theta)\, \alpha_{j_1|s_1}(\tau^n_1) f(\log(\omega^n_{j_1}/w_{s_1}(\theta, \tau^n_1)))\right] \times \\ \left[p_{s_2|s_1}(\tau^n_2, \theta)\, \alpha_{j_2|s_2}(\tau^n_2) f(\log(\omega^n_{j_2}/w_{s_2}(\theta, \tau^n_2)))\right] \times \ldots \times \\ \left[p_{s_T|s_{T-1}}(\tau^n_T, \theta)\, \alpha_{j_T|s_T}(\tau^n_T) f(\log(\omega^n_{j_T}/w_{s_T}(\theta, \tau^n_T)))\right] \Bigg]^{d^n_{j_1} \times d^n_{j_2} \times \ldots \times d^n_{j_T}} .\end{aligned}$$

Note that this likelihood does not use data on retrospective coding. Let $L^i(s^i_t = s)$ be the joint likelihood of the observed, originally coded, data and individual i being in true industry s at time t, given parameters. Then, the following equality holds when the estimated

[55] The discussion here follows closely Arcidiacono and Miller (2011), page 1843.

parameters are equal to the true population parameters,[56]

$$
\begin{aligned}
\sum_{i=1}^{N} \log \mathcal{L}(\tau^i, d^i, \omega^i | p_{s,t}(\tau), p_{s'|s,t}(\tau), w_{s,t}(\tau), \alpha_{j|s}) \quad = & \\
& \sum_{i=1}^{N} \sum_{s_1=1}^{J} \sum_{s_2=1}^{J} \cdots \sum_{s_{t-1}=1}^{J} \sum_{s_{t+1}=1}^{J} \cdots \sum_{s_T=1}^{J} \prod_{j_1=1}^{J} \prod_{j_2=1}^{J} \cdots \prod_{j_T=1}^{J} \frac{L^i(s_t^i = s)}{\sum_{\ell=1}^{J} L^i(s_t^i = \ell)} \times \\
& \log \Bigg[\Big[p_{s_1}(\tau_1^n, \theta)\, \alpha_{j_1|s_1}(\tau_1^n) f(\log(\omega_{j_1}^n / w_{s_1}(\theta, \tau_1^n))) \Big] \times \\
& \Big[p_{s_2|s_1}(\tau_2^n, \theta)\, \alpha_{j_2|s_2}(\tau_2^n) f(\log(\omega_{j_2}^n / w_{s_2}(\theta, \tau_2^n))) \Big] \times \ldots \times \\
& \Big[p_{s_T|s_{T-1}}(\tau_T^n, \theta)\, \alpha_{j_T|s_T}(\tau_T^n) f(\log(\omega_{j_T}^n / w_{s_T}(\theta, \tau_T^n))) \Big] \Bigg]^{d_{j_1}^n \times d_{j_2}^n \times \ldots \times d_{j_T}^n}
\end{aligned}
$$

Then, the maximum likelihood estimator for $\alpha_{j|s}$ satisfies,

$$
\begin{aligned}
\alpha_{j|s}(\tau) \quad &= \quad \underset{\alpha_{j|s}}{\operatorname{argmax}} \sum_{i=1}^{N} \log \mathcal{L}(\tau^i, d^i, \omega^i | \hat{p}_{s,t}(\tau), \hat{p}_{s'|s,t}(\tau), \hat{w}_{s,t}(\tau), \alpha_{j|s}) \\
&= \quad \frac{\sum_{i=1}^{N} \sum_{t=1}^{T} d_{j_t=j}^n \, \frac{\hat{L}^i(s_t^i = s)}{\sum_{\ell=1}^{J} \hat{L}^i(s_t^i = \ell)} \, \mathbb{I}(\tau_t^n = \tau)}{\sum_{i=1}^{N} \sum_{t=1}^{T} \sum_{\tilde{j}=1}^{J} \sum_{\tilde{s}=1}^{J} d_{j_t=\tilde{j}}^n \, \frac{\hat{L}^i(s_t^i = s)}{\sum_{\ell=1}^{J} \hat{L}^i(s_t^i = \ell)} \, \mathbb{I}(\tau_t^n = \tau)},
\end{aligned}
$$

where the likelihood is conditional on parameter estimates for $\hat{p}_{s,t}(\tau), \hat{p}_{s'|s,t}(\tau)$, and $\hat{w}_{s,t}(\tau)$ obtained before using retrospective coding information.

Identification of parameters $\alpha_{j|s}(\tau)$ can be assessed by studying the following moments,

$$
\begin{aligned}
p_j^*(\tau, t) \quad &= \quad \sum_{s=1}^{J} \alpha_{j|s}(\tau)\, p_s(\tau, t), \\
p_j^*(\tau, t)\, p_{j'|j}^*(\tau, t+1) \quad &= \quad \sum_{s=1}^{J} \sum_{s'=1}^{J} \alpha_{j|s}(\tau)\, \alpha_{j'|s'}(\tau)\, p_s(\tau, t)\, p_{s'|s}(\tau, t+1),
\end{aligned}
$$

where, $p_j^*(\tau, t) = \mathbb{E}\left[d_t^{i,*} | \tau, t\right]$ and $p_{j'|j}^*(\tau, t+1), p_s(\tau, t) = \frac{\mathbb{E}\left[d_{j',t+1}^{i,*} d_{j,t}^{i,*} | \tau, t\right]}{\mathbb{E}\left[d_{j,t}^{i,*} | \tau, t\right]}$ are moments of originally coded data, and, as discussed before,$p_s(\tau, t)$ and $p_{s'|s}(\tau, t+1)$ are moments of retro-

[56] This result uses Bayes rule and is at the heart of the theory behind the EM algorithm.

spective data. These equations form a linear-quadratic system in $\alpha_{j|s}(\tau)$ which can lead to more that one solution, although not all of them may be admissible as $\alpha_{j|s}(\tau)$ has to be a probability.

Since the EM algorithm finds a local maximum, the starting conditions for the algorithm may define which solution is obtained.[57] In a related context, but with two choices, Hausman et al. (1998) define a monotone condition for identification, which requires that the elements outside the diagonal of the matrix of $\alpha_{j|s}(\tau)$ are not too large, i.e. that misclassification errors are not overly excessive. Thus, using the EM algorithm, a reasonable starting condition is a matrix for $\alpha_{j|s}(\tau)$ with this characteristic, which would lead to the maximum likelihood estimate that satisfies this condition.[58]

C.2 Identification conditional on misclassification probabilities

The PSID post-1980 and the CPS do not have retrospective industry information. However, I now show that, conditional on the estimated values for α, we can identify mobility rates and employment shares using the maximum likelihood estimators,

$$\big(\hat{p}_{s,t}(\tau), \hat{p}_{s'|s,t}(\tau), \hat{w}_{s,t}(\tau)\big) = \underset{p_{s,t}(\tau), p_{s'|s,t}(\tau), w_{s,t}(\tau)}{\text{argmax}} \mathcal{L}(\tau^i, d^i, \omega^i | p_{s,t}(\tau), p_{s'|s,t}(\tau), w_{s,t}(\tau), \hat{\alpha}_{j|s}).$$

Using similar arguments as in the previous subsection, the maximum likelihood estimators satisfy,

$$\begin{aligned}
\hat{p}_{s_1=s}(\tau,\theta) &= \frac{\sum_{i=1}^{N} \sum_{j_1=1}^{J} d_{j_1}^n \, q_{s_1=s}^n \mathbb{I}(\tau_1^n = \tau)}{\sum_{i=1}^{N} \sum_{\tilde{s}=1}^{J} \sum_{j_1=1}^{J} d_{j_1}^n \, q_{s_1=\tilde{s}}^n \, \mathbb{I}(\tau_1^n = \tau)}, \\
\hat{p}_{s_{t+1}=s'|s_t=s}(\tau,\theta) &= \frac{\sum_{i=1}^{N} \sum_{t=1}^{T-1} \sum_{j_{t+1}=1}^{J} \sum_{j_t=1}^{J} d_{j_t}^n d_{j_{t+1}}^n \frac{\hat{L}^i(s_{t+1}^i = s', s_t^i = s)}{\sum_{\ell=1}^{J} \hat{L}^i(s_{t+1}'^i = \ell, s_t^i = s)} \mathbb{I}(\tau_{t+1}^n = \tau)}{\sum_{i=1}^{N} \sum_{t=1}^{T-1} \sum_{j_{t+1}=1}^{J} \sum_{j_t=1}^{J} \sum_{\tilde{s}'=1}^{J} d_{j_t}^n d_{j_{t+1}}^n \frac{\hat{L}^i(s_{t+1}^i = s', s_t^i = s)}{\sum_{\ell=1}^{J} \hat{L}^i(s_{t+1}'^i = \ell, s_t^i = s)} \mathbb{I}(\tau_{t+1}^n = \tau)}
\end{aligned}$$

[57] Clearly, the maximum likelihood estimator will be the one that delivers the highest value of the likelihood. The discussion here is about two possible set of parameters that deliver this maximum value for the likelihood.

[58] Given the small sample size in the PSID, this is the approach I follow in Section 5.

where $\hat{L}^i(s^i_{t+1} = s', s^i_t = s)$ is the joint likelihood of the data and individual i having true industries s and s' at t and $t+1$ respectively.

Under some conditions, which can be verified given the data and the values of parameters α, these estimators are unique. To see this, note that the equation

$$p^*(\tau, t) = \mathbb{E}\left[d^{i,*}_t | \tau, t\right] = \alpha^T_{j|s}(\tau)\, p(\tau, t).$$

This equation links observed moments and parameters α, defining a linear system for $p(\tau, t)$.[59] Since employment shares and each row of matrix α add to one, the system has order $(J-1)$ and can easily be solved for if the rank of matrix α is $(J-1)$.[60] Thus, given estimated values for α and the moments $\mathbb{E}\left[d^{i,*}_t | \tau, t\right]$, the actual employment shares, $p(\tau, t)$ are identified. In addition, the equations $p^*_j(\tau, t)\, p^*_{j'|j}(\tau, t+1) = \sum_{s=1}^{J} \sum_{s'=1}^{J} \alpha_{j|s}(\tau)\, \alpha_{j'|s'}(\tau)\, p_s(\tau, t)\, p_{s'|s}(\tau, t+1)$, link observed moments and parameters α with unobserved $p_{s'|s}$, forming a linear system of rank $J \times (J-1)$. To illustrate this, take the case of $J = 3$, and omit the dependence on τ and t to simplify the notation. Then, vectorizing the system we get,

$$\begin{bmatrix} p^*_1\, p^*_{1|1} \\ p^*_1\, p^*_{2|1} \\ p^*_1\, p^*_{3|1} \\ p^*_2\, p^*_{1|2} \\ p^*_2\, p^*_{2|2} \\ p^*_2\, p^*_{3|2} \\ p^*_3\, p^*_{1|3} \\ p^*_3\, p^*_{2|3} \\ p^*_3\, p^*_{3|3} \end{bmatrix} = \begin{bmatrix} \alpha_{1|1}\,\alpha_{1|1}\,p_1 & \alpha_{1|1}\,\alpha_{1|2}\,p_1 & \alpha_{1|1}\,\alpha_{1|3}\,p_1 & \cdots & \alpha_{1|3}\,\alpha_{1|1}\,p_3 & \alpha_{1|3}\,\alpha_{1|1}\,p_3 & \alpha_{1|3}\,\alpha_{1|1}\,p_3 \\ \alpha_{1|1}\,\alpha_{2|2}\,p_1 & \alpha_{1|1}\,\alpha_{2|2}\,p_1 & \alpha_{1|1}\,\alpha_{2|3}\,p_1 & \cdots & \alpha_{1|3}\,\alpha_{2|2}\,p_3 & \alpha_{1|3}\,\alpha_{2|2}\,p_3 & \alpha_{1|3}\,\alpha_{2|2}\,p_3 \\ \alpha_{1|1}\,\alpha_{3|3}\,p_1 & \alpha_{1|1}\,\alpha_{3|2}\,p_1 & \alpha_{1|1}\,\alpha_{3|3}\,p_1 & \cdots & \alpha_{1|3}\,\alpha_{3|3}\,p_3 & \alpha_{1|3}\,\alpha_{3|3}\,p_3 & \alpha_{1|3}\,\alpha_{3|3}\,p_3 \\ \alpha_{2|1}\,\alpha_{1|1}\,p_1 & \alpha_{2|1}\,\alpha_{1|2}\,p_1 & \alpha_{2|1}\,\alpha_{1|3}\,p_1 & \cdots & \alpha_{2|3}\,\alpha_{1|1}\,p_3 & \alpha_{2|3}\,\alpha_{1|1}\,p_3 & \alpha_{2|3}\,\alpha_{1|1}\,p_3 \\ \alpha_{2|1}\,\alpha_{2|2}\,p_1 & \alpha_{2|1}\,\alpha_{2|2}\,p_1 & \alpha_{2|1}\,\alpha_{2|3}\,p_1 & \cdots & \alpha_{2|3}\,\alpha_{2|2}\,p_3 & \alpha_{2|3}\,\alpha_{2|2}\,p_3 & \alpha_{2|3}\,\alpha_{2|2}\,p_3 \\ \alpha_{2|1}\,\alpha_{3|3}\,p_1 & \alpha_{2|1}\,\alpha_{3|2}\,p_1 & \alpha_{2|1}\,\alpha_{3|3}\,p_1 & \cdots & \alpha_{2|3}\,\alpha_{3|3}\,p_3 & \alpha_{2|3}\,\alpha_{3|3}\,p_3 & \alpha_{2|3}\,\alpha_{3|3}\,p_3 \\ \alpha_{3|1}\,\alpha_{1|1}\,p_1 & \alpha_{3|1}\,\alpha_{1|2}\,p_1 & \alpha_{3|1}\,\alpha_{1|3}\,p_1 & \cdots & \alpha_{3|3}\,\alpha_{1|1}\,p_3 & \alpha_{3|3}\,\alpha_{1|1}\,p_3 & \alpha_{3|3}\,\alpha_{1|1}\,p_3 \\ \alpha_{3|1}\,\alpha_{2|2}\,p_1 & \alpha_{3|1}\,\alpha_{2|2}\,p_1 & \alpha_{3|1}\,\alpha_{2|3}\,p_1 & \cdots & \alpha_{3|3}\,\alpha_{2|2}\,p_3 & \alpha_{3|3}\,\alpha_{2|2}\,p_3 & \alpha_{3|3}\,\alpha_{2|2}\,p_3 \\ \alpha_{3|1}\,\alpha_{3|3}\,p_1 & \alpha_{3|1}\,\alpha_{3|2}\,p_1 & \alpha_{3|1}\,\alpha_{3|3}\,p_1 & \cdots & \alpha_{3|3}\,\alpha_{3|3}\,p_3 & \alpha_{3|3}\,\alpha_{3|3}\,p_3 & \alpha_{3|3}\,\alpha_{3|3}\,p_3 \end{bmatrix} \times \begin{bmatrix} p_{1|1} \\ p_{2|1} \\ p_{3|1} \\ p_{1|2} \\ p_{2|2} \\ p_{3|2} \\ p_{1|3} \\ p_{2|3} \\ p_{3|3} \end{bmatrix}.$$

Then, provided that the matrix in the right hand side is of rank $J \times (J-1)$,[61] the moments in the left hand side and those that define the matrix in the right side identify the unique unobserved mobility matrix, $p_{s'|s}(\tau, t+1)$, after imposing that the solution has all elements

[59] The superindex T denotes the transpose of the matrix.

[60] Since the matrix α is known, it is easy to check if the rank condition is satisfied.

[61] As before, since this matrix is observed, it is easy to check if the rank condition is satisfied.

positive and they add to one by rows.

Importantly, note that while the likelihood function uses observations on wages, the moment conditions specified before do not rely on information on wages for identification. In principle, one could use these moment conditions to estimate parameters $p_s(\tau, t)$ and $p_{s'|s}(\tau, t+1)$ directly. However, the advantage of using maximum likelihood with the EM algorithm is that we can obtain corrected measures of wages, which, as discussed in Appendix A, are critical to obtain unbiased estimates of structural parameters ν and κ.

Last, structural parameters ν and κ are identified by manipulating the estimates of the coefficients of the regression equation (19), provided these estimates are obtained using corrected measures of mobility and wages. Then, inverting the mapping between parameters and regression coefficients, leads to a unique expression for ν and κ.

D Data on mobility and wages

D.1 PSID

The PSID is a longitudinal survey with a nationally representative sample of over 18,000 individuals in the United States. The survey contains information about these individuals and their descendants, and has been collected continuously since 1968. The PSID contains information about demographic characteristics and socio-economic conditions, including labor earnings and characteristics of individuals' jobs at the time of the interview, such as information about industry and occupation. The data has an annual frequency from 1968 to 1997 and biennially since then.

In this paper I use only the annual information. The main advantage of the PSID relative to other available data sources, is the long panel dimension. An important limitation of the PSID is that the sample size is not very large and, since industry switches are infrequent, estimates of mobility and their statistical significance may be affected by sampling error.

My sample includes males between 25 and 64 years of age, which is the same sample selection criteria on Artuç et al. (2010). Moreover, I use the core Survey Research Center (SRC) sample of the PSID, which is nationally representative, and exclude the Survey of Economic Opportunity (SEO) sample from the analysis, which oversamples low-income households.[62] Unless otherwise noted, all of the moments are computed using individuals that are employed at the time of the survey. Moments on industry mobility are computed over two consecutive years for individuals employed in each of those years.

I use the original and retrospective codes in the PSID sample up to 1980 to obtain estimates of the misclassification probability. Estimation of other parameters uses the CPS data.

Industries and mobility: Up to 2003, the PSID reports information about the industry of individuals using the codes in the 1970 Census of Population.[63] Between 1971 and 1980, industry is coded to two digit-level, and three digit from 1981 onwards for the original coding. Retrospective coding starts in 1968 and ends in 1980, and is coded using three-digits codes also following the industry classification from the 1970 Census of Population. Table 7 shows how the original PSID industry codes map into one of the six broad industries used in the analysis.

Moments on industry mobility are computed over two consecutive years for individuals employed in each of those years.

Real wages: I measure real earnings using PSID data on nominal wages, salaries, and hours, and deflate wages and salaries by the Consumer Price Index for all urban consumers (CPI). For hourly workers, nominal hourly wages are taken directly from the PSID data, and are available from 1968 to 1997 for heads and wives. For salary workers, the PSID constructs a measure of hourly wages for salary workers up until 1992 for head and wives. From 1993 to

[62] In later years the PSID also added immigrant and Latino samples, which I also exclude from my analysis.

[63] The industry codes in the 1970 Census of Population follow closely the 1967 Standard Industrial Classification codes.

Table 7: Broad industry groups

	Original industry coding (1971-1980) 2-digit PSID industry codes	Original industry coding (1981-1997), and Retrospective coding (1968-1997) 3-digit PSID industry codes
Agriculture/Mining	11; 21	17 - 28; 47 - 57
Construction	51	67 - 77
Manufacturing	30 - 49; 85	107 - 398
Transport/Utilities	55-57;	407 - 479
Wholesale/Retail	61; 62; 69	507 -699
Other services	71 - 84; 86 - 92	707 - 937

Note: Columns 2 and 3 show the mapping between PSID industry codes, which follow the 1970 Population Census, and the broad groups I use in the analysis.

1997, the PSID reports data on salary and payment period separately. I use these variables to construct an hourly wage following the methodology from before 1993. In particular, to create hourly wage data for salary workers, the salary variable is divided by the typical number of hours worked in an employee's payment period. For example, an employee with an annual salary would have their salary divided by 2000, which is the approximate number of hours worked in a year. Annual earnings is just the product of wages and hours, which is the measure of real wages I use.[64] In the estimation I use all individuals, but for those workers with values of annual earnings lower than $5,000 or larger than $1,000,000, I do not use information on earnings in the estimation.

Evidence of misclassification

One way to evaluate the effects of misclassification errors is comparing the original and retrospective industry information of individuals in the sample. Table 8 shows the proportion of individuals for which industry codes are different. The table shows employment shares in columns 2 and 3, and proportions of disagreement for all individuals by industry in columns 4 to 8, which distinguish by gender and age groups. At this level of aggregation, differences in industry codes assigned to individuals are, on average, 8%, but there is important het-

[64] In some cases, hourly wage, salary wage, and salaries are top coded. I do not use these observations in my analysis.

erogeneity across different industries and demographic groups.[65] Wholesale and retail trade and construction are on the higher end of misclassification errors, while all other services are at the lower end. Across age groups, discrepancy levels seem similar, but they tend to be lower for women.

Table 8: Differences in original and retrospective industry and occupation coding (fraction of cases with differences, PSID 1968-1980)

Industries	Employment shares (%)		Fraction with differences in coding				
	Original	Retrospective	All	Men	Women	24-35 yrs.	36-64 yrs.
Agric./Mining	4.8	5.0	0.101	0.094	-	0.088	0.109
			(0.010)	(0.010)	-	(0.016)	(0.013)
Construction	7.2	7.6	0.127	0.126	-	0.130	0.125
			(0.010)	(0.010)	-	(0.016)	(0.012)
Manufacturing	28.4	29.4	0.074	0.077	0.053	0.077	0.072
			(0.004)	(0.004)	(0.009)	(0.007)	(0.005)
Transport/Util	8.5	8.1	0.070	0.073	0.050	0.086	0.059
			(0.007)	(0.008)	(0.019)	(0.012)	(0.008)
Wholesale/Retail	13.9	14.4	0.153	0.178	0.073	0.147	0.157
			(0.007)	(0.009)	(0.011)	(0.012)	(0.010)
Other serv.	37.1	35.5	0.049	0.066	0.019	0.049	0.049
			(0.003)	(0.004)	(0.003)	(0.005)	(0.004)
All	100.0	100.0	0.081	0.093	0.038	0.083	0.080
			(0.002)	(0.003)	(0.003)	(0.004)	(0.003)

Note: Author's calculation using PSID data between 1971 and 1980 for industries. Sample restricted to SRC individuals between 25 and 64 years old employed at the time of the survey in two consecutive years. Standard errors in parenthesis. Cells with missing values are cases with a small number of observations.

D.2 Monthly CPS

The CPS collects information on individuals for a sampled household at the same address for four consecutive months, stops for eight months, and then surveys them again for another four months.[66] I use the monthly CPS and match individuals with their survey conducted a year before and compute their employment or non-employment status and work industry, accounting for any change between interviews as a transition. Since I need information on

[65] The level of disaggregation in industry classification matters. Differences between original and retrospective codes are even larger for more disaggregated industry groups.

[66] I access the CPS data from https://cps.ipums.org/cps/. See Flood et al. (2021) for a detailed reference.

wages, I use only the Outgoing Rotation Groups, which are the individuals at month 4 and 8 of the survey.[67]

I use information on weekly earnings deflated by CPI. For observations with top coded earnings, I multiply the value of the top code by 1.5 and drop observation with small values of earnings (the bottom 1% of the earnings distribution by year). Weekly earnings information is available from 1982 onwards.

Table 9 shows mobility rates for different years using the CPS data as originally coded. The mobility rates in the CPS are similar to those in the PSID. For example, mobility rates in 1983 are very close to those in the first panel of Table 1.

Table generated by Excel2LaTeX from sheet 'table append'

The procedure to obtain industry and occupation information in the monthly CPS is very similar to the PSID as individuals provide a verbal description on the type of business of their employer and the tasks they perform, which are then independently coded by a professional coder (Bureau, 2015).[68] As discussed in Murphy and Topel (1987), Moscarini and Thomsson (2007), and Kambourov and Manovskii (2008), industry and occupation information is subject to misclassification errors, which, as in the case of the CPS, greatly inflates mobility rates.[69]

[67] The main limitation with the CPS is that individuals who move to a different address cannot be matched to previous surveys. The match rate over a year is close to 75%. Mortality, residence change, and nonresponse rates are the main drivers of the 25% mismatch rate. I link individuals over time using IPUMS variable CPSIDP, and additionally drop observation for which gender or race is different over time, or age increases after 12 months by more than two years. I weight observation using variable LNKFW1YWT. In some months in years 1984, 1985, 1994, 1995 and 2004, changes in individual identifiers in the CPS make linking individuals over time impossible, leading to a lower number of individuals linked over time in those years.

[68] According to the CPS interviewing manual, Census employees in Jeffersonville, Indiana assign industry codes to individuals based on the business or industry description individuals provide. Similarly, coders assign occupation codes on the basis of the kind of work the specified person usually does and on a description of his/her most important activities or duties.

[69] In 1994 the CPS introduced dependent coding as a way to reduce the magnitude of misclassification errors. Basically, under dependent coding, individuals are asked to report information on industry and occupation the first time they participate in the sample, while in subsequent months they are asked whether they are still employed with the same firm and performing the same tasks as in the previous month. If that is the case, then codes for the second month are simply a copy of those from the first month, reducing spurious mobility. However, note that dependent coding only helps for individuals that do not change their employer or main tasks from one month to the next, and is only available at higher frequencies, as information on

Table 9: Mobility rates using originally coded data in the CPS

			1983 - original			
	Agric./Mining	Construction	Manufacturing	Transport/Util	Wholesale/Retail	Other serv.
Agric./Mining	0.861	0.021	0.028	0.011	0.030	0.049
Construction	0.013	0.835	0.036	0.027	0.030	0.059
Manufacturing	0.006	0.013	0.893	0.011	0.045	0.032
Transport/Util	0.007	0.018	0.023	0.890	0.022	0.040
Wholesale/Retail	0.010	0.022	0.081	0.017	0.809	0.061
Other serv.	0.006	0.019	0.027	0.011	0.031	0.906

			1989 - original			
	Agric./Mining	Construction	Manufacturing	Transport/Util	Wholesale/Retail	Other serv.
Agric./Mining	0.847	0.029	0.036	0.016	0.035	0.036
Construction	0.013	0.825	0.040	0.019	0.037	0.065
Manufacturing	0.007	0.018	0.866	0.013	0.053	0.044
Transport/Util	0.008	0.020	0.027	0.869	0.029	0.047
Wholesale/Retail	0.010	0.021	0.078	0.019	0.801	0.071
Other serv.	0.006	0.020	0.032	0.014	0.034	0.894

			2000 - original			
	Agric./Mining	Construction	Manufacturing	Transport/Util	Wholesale/Retail	Other serv.
Agric./Mining	0.794	0.040	0.035	0.024	0.049	0.059
Construction	0.012	0.810	0.042	0.026	0.043	0.067
Manufacturing	0.008	0.025	0.810	0.018	0.074	0.064
Transport/Util	0.008	0.026	0.033	0.821	0.041	0.072
Wholesale/Retail	0.012	0.030	0.078	0.029	0.760	0.090
Other serv.	0.006	0.023	0.032	0.021	0.041	0.877

			2015 - original			
	Agric./Mining	Construction	Manufacturing	Transport/Util	Wholesale/Retail	Other serv.
Agric./Mining	0.777	0.045	0.042	0.030	0.045	0.062
Construction	0.017	0.802	0.035	0.032	0.037	0.077
Manufacturing	0.013	0.035	0.767	0.025	0.071	0.089
Transport/Util	0.012	0.032	0.034	0.776	0.049	0.096
Wholesale/Retail	0.013	0.033	0.067	0.035	0.743	0.109
Other serv.	0.007	0.021	0.030	0.025	0.039	0.878

Note: Author's calculation using CPS outgoing rotation group data between 1982 and 2016 for industries. Sample restricted to individuals between 25 and 64 years old employed at the time of the survey in two consecutive years. Mobility rates for 1983 is an average of mobility rates between 1982 and 1984, and similarly for the other periods.

Under the assumption that misclassification probabilities are similar between the PSID and the CPS, which given the similarities in the way individuals are asked to report this information and the way the description is then coded, seems reasonable,[70] then I can use estimates of α from Table 3 to correct mobility in the CPS data. Thus, I implement this using CPS data and the EM algorithm detailed in Section 4, but taking the values of α as

industry and occupation is coded independently in month 5, which will affect all measured mobilities one-year apart.

[70] The similar level and evolution of mobility rates computed with originally coded data in the PSID and CPS in Figure 1 provides support to this assumption.

given from Table 3.[71] I adapt equation (11) as follows,

$$\hat{\mathscr{L}}(\mathbf{d},\boldsymbol{\tau},\boldsymbol{\omega}|\hat{p}_j,\hat{p}_{j'|j},\hat{w}_s,\alpha) \;=\; \sum_{n=N}^{N}\sum_{s=1}^{J}\sum_{s'=1}^{J}\sum_{j=1}^{J}\sum_{j'=1}^{J} d_j^n\, d_{j'}^n q_{ss'}^n \times \\ \log\left[\hat{p}_s(\tau^n)\,\alpha_{j|s}(\tau^n)\,\hat{p}_{s'|s}(\tau^n)\alpha_{j'|s'}(\tau^n)\,f(\log(\omega^n/\hat{w}_s(\tau^n)))f(\log(\omega'^n/\hat{w}_{s'}(\tau^n)))\right],$$

as the time dimension in the CPS is only two consecutive periods at the yearly frequency, and use equations (10), (12), and (13) to obtain estimates of industry mobility corrected of misclassification.

D.3 March CPS

Here I highlight different problems with another usual source of industry and occupational mobility, which is used by Artuç et al. (2010).

The Annual Social and Economic Supplement of the CPS, which is referred to as the March CPS since it is conducted in that month, asks individuals not only about industry and occupation of their current job in March, but also about the characteristics of the longest job they held in the previous calendar year. While other studies have associated the information in the March CPS to a yearly mobility rate, it is not clear that the frequency of mobility computed using this measure delivers a yearly mobility. In fact, as Kambourov and Manovskii (2013) argue, "one might expect that on average it refers to the occupation held in the middle of the previous year, so that by comparing the occupation in the March CPS to the one in the Basic March CPS, mobility over only a nine-month period is identified. Even this does not seem to be the case, however. Instead, we suggest that the March CPS measures mobility between March of this year and the very end of the previous year." (page 182) In this way, Kambourov and Manovskii (2013) argue that the March CPS suffers from a severe time aggregation problem and measures mobility at the two or three month frequency.

[71] I allow for a polynomial of order three in time in the multinomial logits used in the EM algorithm. As argued by Moscarini and Thomsson (2007), the patterns of mobility over time are non-monotone, thus a higher order polynomial allows for a more flexible evolution of mobility over time.

Since the frequency is effectively much shorter, this explains why measured out-mobility is smaller relative to other surveys. There are many reasons for this severe time aggregation. For example, individuals with many different jobs may find it hard to recall correctly their longest one and report the most recent one. This recall bias is a well-known problem of panel data created using retrospective questions, as in the March CPS.

Artuç et al. (2010)'s correction of March CPS mobility flows:

In Section IV.B, Artuç et al. (2010) discuss this problem with the March CPS data and propose a simple way to correct the March CPS mobility moments. In particular, they use data on gross industry flow rates computed using the National Longitudinal Survey of Youth (NLSY79), which "do not suffer from the timing problems just described for the March CPS" (page 1023). While this statement is true, the NLSY suffers from other problems. First, the NLSY79 follows a representative cohort of individuals with age between 14 and 22 in the year 1979. Since young individuals have higher mobility rates and employment changes than older individuals, it is possible that recall bias affects younger workers more than older workers and the proposed correction should vary by age. Similarly, they use a common procedure to correct all mobility flows, but in industries with high employment turnover and higher levels of out-mobility, it is reasonable to expect workers to have more jobs and be more severely affected by recall bias. Then it is not clear that the correction they propose extends to workers with different characteristics or in a similar way to all industries.

Second, the NLSY79 introduced dependent coding for industry and occupation in 1994. Before that, industry information was coded independently and likely subject to misclassification errors similar to the ones I study in my paper. In fact, Neal (1999) argues that "the occupation and industry codes in the NLSY79 contain many errors that imply false changes in industry or occupation." (page 245) This is not surprising since the questions on industry and occupations in the NLSY79 questionnaire are worded almost identically to the questions in the PSID and CPS, which are then coded by the BLS staff independently each year, similarly to how it is done in the CPS and PSID. In this way, the target data that Artuç et al. (2010) use to correct mobility measures suffers from misclassification errors similar to those in the PSID and CPS. Therefore, Artuç et al. (2010) correction increases the amount of mobility in the March CPS to the level of mobility one would observe in the data with misclassification errors. This is likely the reason why ACM estimate that the frequency of moves in the raw March CPS data is around five months, while Kambourov and Manovskii (2013) estimate it to be two or three months.

In summary, while the March CPS offers an alternative measure of industry and occupation mobility at the yearly frequency, it presents a different type of problems relative to those in the PSID and the monthly CPS (and other similar surveys).

D.4 Misclassification in other surveys

Here I review other publicly available datasets with information on U.S. workers with a panel structure: the National Longitudinal Survey of Youth (NLSY79) and the Survey of Income and Program Participation (SIPP). These two surveys, together with the PSID and the CPS, are the most widely used public sources of microdata in labor, trade, and macroeconomics for the United States.

The NLSY79 and the SIPP contain information about the characteristics of the job of individuals over time and the panel dimension is relatively long. The NLSY79 is a nationally representative survey of a cohort of over 12,500 young men and women living in the United States in 1979. Individuals in this cohort were ages 14 to 22 when first interviewed in 1979 and the U.S. Bureau of Labor Statistics interviewed these individuals yearly up to 1994 and every two years after that. This survey contains information on demographic characteristics, education and employment choices, and income, among other information.

The SIPP is composed of a series of panels starting in 1983. In each panel, a nationally representative sample of U.S. households are surveyed between 2.5 to 4 years, depending on the panel, with sample size ranging from approximately 14,000 to 52,000 interviewed households. It provides information on individual's characteristics, employment, income, and government program participation.

Similar to the PSID and the CPS, the NLSY79 and the SIPP suffer from misclassification errors. This has been documented in Carrillo-Tudela and Visschers (2023) for the SIPP, and Neal (1999) for the NLSY79. In all these surveys, the way individuals are asked about the characteristics of their firm (related to industry) and the tasks they perform on their job (related to occupations) are very similar and the way these answers are mapped to an industry and occupation code are also similar. Thus, it is likely that the misclassification probability across these different dataset would be similar as well.

In the mid-1990s, these surveys (and the CPS) introduced dependent industry and occupation coding. Under dependent coding, a worker that reports that he/she did not change employers or, in the case of occupations, the characteristics of the job, the industry and occupation codes are imputed to be the same as those coded for the previous period.

While dependent coding can greatly reduce misclassification errors for some individuals, I would like to highlight three important limitations. First, for individuals that actually change employers, information about industry and occupation are coded independently and thus, subject to misclassification errors. This applies to direct employer-to-employer transitions, which are about 2.5% per month on average and above 5% per quarter (Fujita, Moscarini, & Postel-Vinay, 2020), but also to transitions with short unemployment spells. Second, yearly industry and occupation mobility in the CPS is independently coded, as this information in survey month 5 is no longer connected to previous survey months. Then, for yearly mobility in the CPS, all of the sample is subject to misclassification errors. This also extends to the NLSY and SIPP at other frequencies. Finally, it is worth noting that, while dependent coding alleviates misclassification errors, for workers that actually move, dependent coding does not apply and information is subject to misclassification. Then, the patterns of moving, which are informative about the amount and direction of reallocation, will still be affected by misclassification.

In addition, Fujita et al. (2020) document problems with dependent coding in the CPS. As they put it, "the incidence of missing answers to the question on change of employer sharply increases starting with the introduction of a new software instrument to conduct interviews in January 2007 and of the Respondent Identification Policy in 2008-2009." Note that these new policies by the U.S. Census Bureau and the Bureau of Labor Statistics are not specific to the CPS and also affect other surveys administered by these agencies.

Correcting mobility in the SIPP and NLSY79 It is possible to adapt the method I develop in Section 4 to correct industry and occupation mobility in these surveys. Similar to the way I proceed with the CPS data, it is straightforward to use the estimated value of the misclassification probability matrix α (estimated using the PSID data), and conditional on the value of α use the EM algorithm and equations (7) to (13) to obtain estimates of industry (or occupation) mobility corrected for misclassification.

E Trade shocks: welfare and reallocation

While there is evidence of a substantial bias in parameter estimates, the effect of this bias on the speed of reallocation or on welfare after a shock is not immediately clear. To better understand this, I proceed in two ways. The first exercise is simpler and assumes that wages, and their changes, are exogenous. For this I use the model presented in Section 3, but using the dynamic hat algebra of Caliendo et al. (2019). Since the first exercise is a partial equilibrium model focusing on labor supply only, the second exercise incorporates production and international trade, with a well-specified labor demand. Wages and their changes are the results of trade shocks and general equilibrium across all industries and countries. In both exercises I compare the evolution of labor across industries and the impact on workers' welfare in the United States to a shock that permanently affects wages across sectors in an asymmetric way.

E.1 Exogenous wages

I begin with a simple exercise that focuses solely on labor supply and welfare due to exogenous wage changes. As discussed in the main text, misclassification biases parameter estimates towards zero. The goal of this example is to highlight that different parameter values can lead to important economic differences in the effects of shocks on worker reallocation. In this example, I exogenously increase wages in all sectors except manufacturing by 3%, and decrease manufacturing wages by 3%. I use two sets of parameter values: $\nu = 1.8$ and $\kappa = 7$, which correspond to those in Table 5 using original coded data, and $\nu = 2.2$ and $\kappa = 10$, which correspond to the estimates using data corrected for misclassification.[72]

The left panel of Figure 3 shows the yearly evolution of the change in the manufacturing employment share, relative to the initial value, after the change in wages for these two sets of parameter values. In both cases, relative wages in manufacturing fall and workers reallocate out of manufacturing and into other sectors. However, the total long-run impact on the manufacturing share of employment differ across these parameter values. In particular, the total manufacturing employment drop is larger using parameters estimated with corrected data. By year 30, the differences in the employment share across the two economies is almost one percentage point. Since the drop in the manufacturing employment share by that time is between 6.5 and 7.5 percentage points, the differences are substantial, in the order of 15% . The right panel of the figure shows the change in welfare, measured as consumption

[72] Wages are calibrated to approximate mobility flows and average labor allocation in the data given the values for ν and κ. Since parameters are different for the two economies due to misclassification errors, the vector of wages needed to obtain similar labor allocation are different as well. I calibrate the initial wage vector of the model to be $w = [0.81, 0.87, 1.01, 1.00, 0.82, 1.09]$ for the case of $\nu = 1.8$ and $\kappa = 7$, and use $w = [0.8640.881.0151.00.821.07]$ for the case of $\nu = 1.8$ and $\kappa = 7$. This applies to this exercise only as for the next two exercises initial conditions are taken from the data.

equivalent, for a worker that starts the period in each of the listed industries.[73] Since estimated mobility is higher and parameters are lower when using originally coded data, the option value for workers in manufacturing is better. Thus, despite the same wage change in manufacturing in both economies, welfare is not so negatively impacted in the economy that uses estimates from originally coded data. For a worker attached to manufacturing during the period of the shock, welfare increases, comparable to a permanent rise of 1% of consumption in the economy with estimates from from originally coded data, while the welfare change is negligible and close to an increase of 0.1% of consumption when using parameters obtained with data corrected for misclassification. In the other sectors, welfare increases, as wages increase, but the option value is more negatively affected in the model with parameter estimates from originally coded data.

The estimated parameters κ and ν, together with the calibrated values of wages, approximate reasonably well mobility flows and labor allocation observed in the data, but cannot fit them perfectly. This is driven largely by the assumption of a common cost of switching for all sectors. Other than wage differences, different sectors are very symmetric in this economy and is difficult for a single parameter to generate the rich patterns of mobility observed in the data. Moreover, the initial labor allocation measured in the data may not correspond to a stationary equilibrium. As labor adjusts slowly to shocks, shocks that occurred before the initial observed period will still have an impact on the evolution of labor, mobility and wages.

To overcome these potential issues, I employ the dynamic hat algebra developed in Caliendo et al. (2019), which can accommodate arbitrary costs of switching sectors and does not require the economy to be in a stationary equilibrium initially. Conditional on a value for parameters ν and β, the method uses mobility flows and initial labor shares to compute the dynamics of industry employment along the transition and the changes in welfare

[73] See Appendix G for the details on the consumption equivalent change in welfare.

to an unanticipated wage shock.[74] I assume the same exogenous and unanticipated wage shock as in the previous case. The economy gradually converges to a stationary equilibrium compatible with the new wages.

Figure 4 shows the evolution of the manufacturing employment share in panel (a), and the changes in welfare due to the shock in panel (b), for the two different cases: using estimates under original coding and estimates corrected for misclassification. Both economies are not in a stationary equilibrium initially and the manufacturing labor share under no shocks is gradually falling over time. Thus, in panel (a), I show the evolution of the manufacturing labor share after the wage shock hits and subtract the evolution of the labor share for the economy with no shock. In other words, the figure shows the change in the labor share due to the wage shock. Similar to the previous case, the economy with parameters and mobility estimated with corrected data reaches a lower level of manufacturing employment than the economy with originally coded data. The magnitudes are somewhat smaller than those discussed before, but are still important. By year 15, differences across these two differently measured economies are close to 0.80 percentage points. Since the drop in the manufacturing employment share is close to 3.2 percentage points, misclassification can alter the estimated impact of the shock by 25%. In the long run, differences are slightly bigger, but the discrepancies persist. Panel (b) shows that estimates of the welfare changes of the shock are also different depending on whether original or corrected data is used. In both cases, welfare increases in all sectors but the increase is smaller for workers initially attached to manufacturing. Relative to the previous case, there is a more pronounced heterogeneity in the welfare effects across sectors, a product of the more flexible structure. However, the effects computed using estimates from data with original coding show a consumption equivalent welfare increase of 1.30%, while the effects computed using corrected data are

[74] I use employment shares and mobility matrices estimated with the CPS data for the year 2000. For one economy I use this data as originally coded and for the other I use the corrected values estimated in Section 5. Appendix F details the equations used in the dynamic hat algebra, which are derived from the structural model of Section 3.

0.85%, a difference of almost fifty percent of the magnitude. Wages increase in all other sectors leading to important welfare increases, but the increases are larger when using the estimates from the data corrected for misclassification. This is a reflection of the lower mobility flows and how, in turn, this lower mobility translates into differences in the option value of moving.

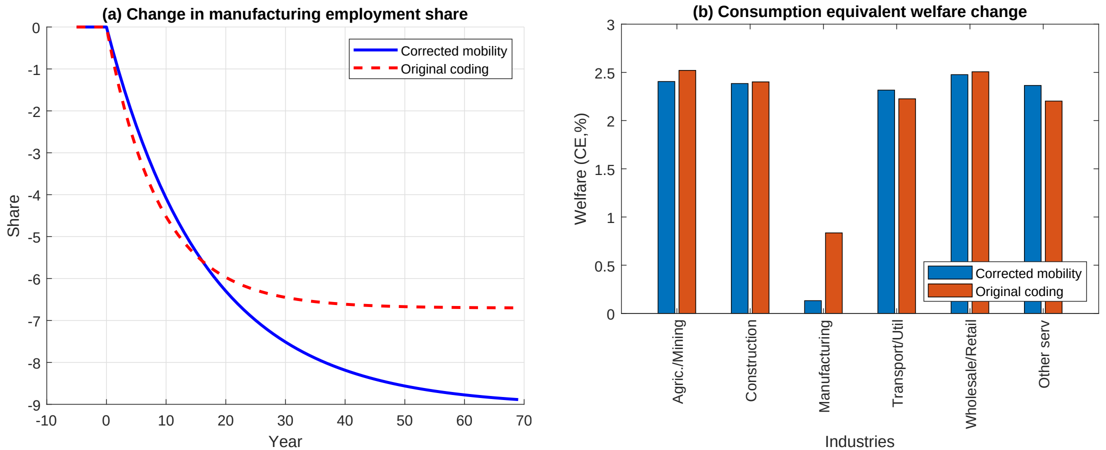
E.2 General equilibrium and trade shock

I now extend the model to include production and characterize labor demand, such that wages endogenously adjust to a trade shock. This exercise is relevant since, unlike the previous cases, the evolution of wages and the total effect of the trade shock on wages will be affected by the speed and total magnitude of labor reallocation across sectors.

The production side of the general equilibrium model is a simpler version of Caliendo et al. (2019). In particular I calibrate the model for three countries in the world and abstract from modeling regions within countries. Moreover, I abstract from input-output relations across sectors and assume the only factors of production are labor and structures (fixed capital). Other than in the United States, I assume that workers do not face any frictions to reallocate across labor markets, so that in each country there is a unique labor market wage. All labor and goods markets are competitive and prices and wages are determined by supply and demand at each time.

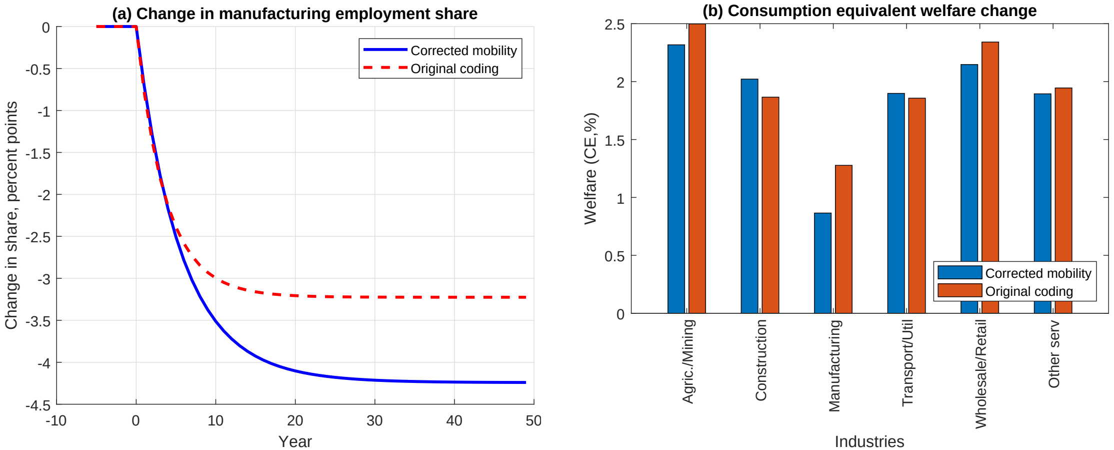
Following Eaton and Kortum (2002) I assume that in each sector there is a continuum of varieties that can potentially be traded across countries. The technology to produce any of these varieties is described by a Cobb-Douglas function with workers and structures as inputs. The total factor productivity of country i and sector j in variety x is $z^{ij}(x)$. I assume that each $z^{ij}(x)$ is the result of an independent Fréchet random draw with dispersion parameter θ and scale parameter A^{ij}. The output for a producer of an intermediate variety

Figure 3: Effects of exogenous wage changes under different parameter estimates



Note: Evolution of manufacturing employment in the model with exogenous wage changes under parameter values $\nu = 2.54$ and $\kappa = 11.3$ for original coding, and $\nu = 3.45$ and $\kappa = 17.03$ for corrected for misclassification. In both cases initial wages are $w = [0.85, 0.95, 1.08, 0.99, 1.02, 1.09]$, and the shock increases wages by 3% in all sectors but manufacturing, and decreases manufacturing wages by 3%.

Figure 4: Effects of exogenous wage changes in dynamic hat algebra model



Note: Evolution of the changes in the manufacturing employment share in the dynamic hat algebra model with exogenous wage changes under parameter $\nu = 2.54$ for original coding, and $\nu = 3.45$ for the data corrected for misclassification. In both cases the shock increases wages by 3% in all sectors but manufacturing, and decreases manufacturing wages by 3%.

with efficiency $z^{ij}(x)$ is,

$$q_t^{ij}(x) = z^{nj}(x)(h_t^{ij}(x))^{\xi^i}(l_t^{ij}(x))^{1-\xi^i},$$

where $l_t^{ij}(x)$, $h_t^{ij}(x)$ are labor and fixed capital (structures) inputs used in variety x, respectively. Parameter ξ^i is the share of structures in production. Structures are in fixed supply in each country and sector, and the total availability is given by H^{ij}. Denote by r_t^{ij} the rental price of structures in country i and sector j, then, by competition and free entry and exit, the price of producing one unit of variety x is equal to its unit costs,

$$\tilde{p}_t^{ij}(x) = \frac{B^{ij}\ (r_t^{ij})^{\xi^i}\ (w_t^{ij})^{1-\xi^i}}{z^{nj}(x)}$$

where B^{nj} is a constant that depends on parameters.

Trade costs are represented by $\lambda_t^{ij,nj}$ and are of the iceberg type. One unit of any variety of intermediate good j shipped from country i to n requires producing $\lambda_t^{ij,nj} \geq 1$ units in country i. If a good is nontradable, then $\lambda = \infty$. Competition across countries implies that the price paid for a particular variety of good j in country i is given by the minimum unit cost of acquiring that variety across countries, taking into account trade costs. That is,

$$p_t^{ij}(x) = \min_n \left\{\tilde{p}_t^{nj}(x)\lambda_t^{nj,ij}\right\}.$$

Intermediate varieties from sector j are needed to produce a final good from sector j, denoted by Q_t^{ij}, according to,

$$Q_t^{ij} = \left(\int (\tilde{q}_t^{ij}(x))^{1-1/\eta^{ij}}\ dx\right)^{\eta^{ij}/(\eta^{ij}-1)},$$

where $\tilde{q}_t^{ij}(x)$ denotes the quantity of variety x sourced from the lowest cost producer. Competitive behavior implies zero profits at all times.

Given the properties of the Fréchet distribution, the price of the sectoral aggregate good j in country n at time t is (Eaton & Kortum, 2002),

$$P_t^{nj} = \Gamma^{ij} \left(\sum_{i=1}^{N} \left(B^{ij} \, (r_t^{ij})^{\xi^i} \, (w_t^{ij})^{1-\xi^i} \right)^{-\theta} (\lambda_t^{ij,nj})^{-\theta} (A^{ij})^{\theta} \right)^{-1/\theta}, \tag{20}$$

where Γ^{ij} is a constant. Moreover, the share of varieties that country i sources from country n is given by,

$$\pi_t^{nj,ij} = \frac{\left(B^{nj} \, (r_t^{nj})^{\xi^n} \, (w_t^{nj})^{1-\xi^n} \right)^{-\theta} (\lambda_t^{nj,ij})^{-\theta} (A^{nj})^{\theta}}{\sum_{m=1}^{N} \left(B^{mj} \, (r_t^{mj})^{\xi^m} \, (w_t^{mj})^{1-\xi^m} \right)^{-\theta} (\lambda_t^{mj,ij})^{-\theta} (A^{mj})^{\theta}}. \tag{21}$$

Trade can be imbalanced due to differences across countries between where income is generated and where it is consumed. I assume there is a unit mass of rentiers in each country. Rentiers cannot relocate to other regions. They manage the local structures, rent them to local firms, and send all the rents to a global portfolio. Rentiers in country n have a total ownership share ι^n of the global portfolio of rents, with $\sum_{n=1}^{N} \iota^n = 1$. The difference in rental income generated and consumed in country i generate trade surplus equal to $\sum_{j=1}^{J} r_t^{ij} \, H^{ij} - \iota^i \chi_t$, where $\chi_t = \sum_{i=1}^{N} \sum_{j=1}^{J} r_t^{ij} H^{ij}$ are the total rents in the world.

Workers and rentiers consume a Cobb-Douglas aggregate of all final goods from all sectors, with parameters γ^j, common for all countries. Let X_t^{nj} be the total expenditure on the final good of sector j good in country i. Then, goods market clearing implies

$$X_t^{ij} = \gamma^j \, \left(\sum_{k=1}^{J} w_t^{ik} L_t^{ik} + \iota^i \chi_t \right),$$

where $\sum_{k=1}^{J} (w_t^{ik} L_t^{ik} + \iota^i \chi_t)$ is total income, and the value of the final demand, in country i.

Labor market clearing in country i and sector j implies,

$$L_t^{ij} = \frac{(1-\xi^i)}{w_t^{ij}} \sum_{n=1}^{N} \pi_t^{ij,nj} \, X_t^{nj},$$

while the market clearing condition for structures is,

$$H^{ij} = \frac{\xi^i}{r_t^{ij}} \sum_{n=1}^{N} \pi_t^{ij,nj} \; X_t^{nj}.$$

These equilibrium conditions can be manipulated and expressed in changes (hat algebra) to simplify the calibration and the computations. Appendix F presents the details.

I calibrate the general equilibrium model to three countries: the United States, other advanced economies, and the rest of the world.[75] I use the same six broad sectors as used throughout the paper. Initial values of trade shares, total expenditures, and value of production are from the World Input-Output Dataset (WIOD), 2013 release, for the year 1997.[76] I calibrate parameter θ to be equal to 4, as estimated by Simonovska and Waugh (2014). I abstract from differences in labor and capital (structures) income share across countries and assume $\xi^i = 0.33$ for all countries, a common value used in macroeconomics. The calibration of parameter ι^n is straightforward using data on trade deficits, total world output and the share of each country in world production.[77] Finally, I use information about final demand consumption by sector to get parameters γ^j of the Cobb-Douglas consumption aggregator, and use information on trade flows by sector to calibrate π.

I calibrate the initial value of the shares of employment by sector for the United States to those in the CPS data for the year 2000. Similarly, I use the matrix of workers' mobility by industry as estimated using CPS data for the year 2000. For one economy I use the data as originally coded and for the other I use the corrected values estimated in Section 5. I

[75] Other advanced economies include: Australia, Austria, Belgium, Canada, Germany, Denmark, Spain, Finland, France, United Kingdom, Greece, Ireland, Italy, Japan, Korea, Luxembourg, Netherlands, Norway, Portugal, and Sweden. The rest of the world include: Bulgaria, Brazil, China, Cyprus, Czech Republic, Estonia, Hungary, Indonesia, India, Lithuania, Latvia, Mexico, Malta, Poland, Romania, Russia, Slovak Republic, Slovenia, Turkey, Taiwan, and those countries classified as rest of the world in the dataset.

[76] See Timmer, Dietzenbacher, Los, Stehrer, and De Vries (2015) for a description of the data.

[77] The formula is, $\iota^n = \frac{\text{deficit}^n}{\xi\,(\text{World Output})} + \frac{\text{country}^n \text{ output}}{\text{World Output}}$. The formula is intuitive. Multiplying and dividing the last term by ξ, it becomes the country's generated rents as a share of world rents. A trade deficit is possible when ι is larger than the share of rents the country generates and sends to the global portfolio.

compute the implicit wages consistent with these labor shares and the total labor income from WIOD.[78] A normalization is needed to pin down nominal variables. I normalize the price of the final consumption aggregate in the United States is equal to one in every period.

The results for this case given a trade shock in the form of an increase in TFP in the rest of the world are the ones presented at the end of Section 5.

F Dynamic hat algebra

Here I manipulate equations (1) and (2) and derive the equations employed in the dynamic hat algebra computations. Take differences over two periods using equation (1) under the assumption of no shocks, then,

$$v_{t+1}^{j}(\tau) - v_{t}^{j}(\tau) = \left[\frac{w_{t+1}^{j}(\tau)}{w_{t}^{j}(\tau)} - 1\right] w_{t}^{j}(\tau) + \nu \log\left[\sum_{\ell=1}^{J} \mu_{t}^{j\ell}(\tau) \exp\left(\beta(v_{t+2}^{\ell}(\tau) - v_{t+1}^{\ell}(\tau))\right)^{1/\nu}\right]. \tag{22}$$

Similarly, take the ratio of the expression in equation (2) over two consecutive periods under the same assumption, then

$$\mu_{t+1}^{j\ell}(\tau) = \frac{\mu_{t}^{j\ell}(\tau) \exp\left(\beta(v_{t+2}^{\ell}(\tau) - v_{t+1}^{\ell}(\tau))\right)^{1/\nu}}{\sum_{m=1}^{J} \mu_{t}^{jm}(\tau) \exp\left(\beta(v_{t+2}^{m}(\tau) - v_{t+1}^{m}(\tau)\right)^{1/\nu}}. \tag{23}$$

Equations (22) and (23) are special cases of those in Proposition 2 in (Caliendo et al., 2019), but allowing for worker heterogeneity in τ. In period 1, workers are "suprised" by a shock. Denote with a " ˆ ", variables after the shock, then, the following expressions must be used in that first period,

$$\hat{v}_{1}^{j}(\tau) - v_{0}^{j}(\tau) = \left[\frac{\hat{w}_{1}^{j}(\tau)}{w_{0}^{j}(\tau)} - 1\right] w_{0}^{j}(\tau) + \nu \log\left[\sum_{\ell=1}^{J} \mu_{0}^{j\ell}(\tau) \psi_{1}^{\ell} \exp\left(\beta(\hat{v}_{2}^{\ell}(\tau) - \hat{v}_{1}^{\ell}(\tau))\right)^{1/\nu}\right], \tag{24}$$

[78] As is usual in the literature, I adjust slightly the first-period value of expenditures/production by sector and country such that, given parameters and initial values of other endogenous variables, the economy is in equilibrium in the first period.

$$\mu_1^{j\ell}(\tau) = \frac{\mu_0^{j\ell}(\tau)\psi_1^{\ell} \exp\left(\beta(\hat{v}_2^{\ell}(\tau) - \hat{v}_1^{\ell}(\tau))\right)^{1/\nu}}{\sum_{m=1}^{J} \mu_0^{jm}(\tau)\psi_1^{m} \exp\left(\beta(\hat{v}_2^{m}(\tau) - \hat{v}_1^{m}(\tau))\right)^{1/\nu}}, \tag{25}$$

where $\psi_1^{\ell} = \exp\left(\beta(\hat{v}_1^{\ell}(\tau) - v_1^{\ell}(\tau))\right)^{1/\nu}$. For periods $t > 1$, the evolution of value function and mobility flows in changes in the economy hit by the shock can be characterized by equations identical to (22) and (23), but with variables with "ˆ". In a similar way, these expressions are a special case of Proposition 3 in Caliendo et al. (2019).

In both these cases, given a sequence of wage changes, a solution is found by imposing the terminal condition $v_T^j(\tau) - v_{T-1}^j(\tau) = 0$, for sufficiently large T, and taking μ_0 as given, for example by measuring flows in the data.

Production

To simplify computations, we can use equilibrium conditions in changes. For this, take ratios of equilibrium conditions (20) and (21) over two consecutive periods, such that,

$$\pi_{t+1}^{nj,ij} = \frac{\pi_t^{nj,ij} \left(\frac{w_{t+1}^{nj}}{w_t^{nj}}\right)^{-\theta(1-\xi^n)} \left(\frac{r_{t+1}^{nj}}{r_t^{nj}}\right)^{-\theta\xi^n} \left(\frac{\lambda_{t+1}^{nj,ij}}{\lambda_t^{nj,ij}}\right)^{-\theta}}{\sum_{m=1}^{N} \pi_t^{mj,ij} \left(\frac{w_{t+1}^{mj}}{w_t^{mj}}\right)^{-\theta(1-\xi^m)} \left(\frac{r_{t+1}^{mj}}{r_t^{mj}}\right)^{-\theta\xi^m} \left(\frac{\lambda_{t+1}^{mj,ij}}{\lambda_t^{mj,ij}}\right)^{-\theta}},$$

and the changes in the price of the Cobb-Douglas aggregate final good,$P_t^{C,n}$, is,

$$\frac{P_{t+1}^{C,n}}{P_t^{C,n}} = \prod_{j=1}^{J} \left(\sum_{m=1}^{N} \pi_t^{mj,ij} \left(\frac{w_{t+1}^{mj}}{w_t^{mj}}\right)^{-\theta(1-\xi^m)} \left(\frac{r_{t+1}^{mj}}{r_t^{mj}}\right)^{-\theta\xi^m} \left(\frac{\lambda_{t+1}^{mj,ij}}{\lambda_t^{mj,ij}}\right)^{-\theta} \right)^{-\frac{\gamma_j}{\theta}}.$$

These two expressions, together with market clearing conditions, characterize equilibrium in each period t for a given value of labor supply in each sector and country, L_t^{ij}.[79]

[79] As mentioned in Section 6, I normalize nominal variables such that $P_t^{C,n} = 1$ in every period in the United States.

G Consumption equivalent welfare changes

I now derive the expressions for the equivalent variation in consumption. Using (1) and (2), we can write the value at period 1 as,

$$v_1^j(\tau) = w_1^j(\tau) + \beta\, v_2^j(\tau) - \nu \log\left(\mu_1^{j\ell}(\tau)\right).$$

Iterating this expression forward

$$v_1^j(\tau) = \sum_{t=1}^{\infty} \beta^{t-1} \left[w_t^j(\tau) - \nu \log\left(\mu_1^{j\ell}(\tau)\right)\right].$$

Define variables with a "ˆ", as those corresponding to the equilibrium after an unanticipated shock hits in period 1. Thus, lifetime utility in period 1 after a shock is $\hat{v}_1^j$. The consumption equivalent change in welfare is the scalar $\delta^j(\tau)$, for each j and τ, such that,

$$\hat{v}_1^j(\tau) = \sum_{t=1}^{\infty} \beta^{t-1} \left[w_t^j(\tau)(1+\delta^j(\tau)) - \nu \log\left(\mu_1^{j\ell}(\tau)\right)\right].$$

In words, this simply says that a change in consumption of δ^j every period in the economy with no shocks, would bring lifetime utility of that economy to the levels of utility of the economy with the shock. Manipulating these equations we get,

$$\delta^j(\tau) = \frac{(1-\beta)\ \left(\hat{v}_1^j(\tau) - v_1^j(\tau)\right)}{w_1^j(\tau)}.$$

H Heterogeneous costs of industry switching

In this appendix I allow for the estimates of structural parameters to depend on workers' characteristics and to vary over time, extending the results of Section 5.

Figure 1 and Table 4 show that industry mobility increased over time. In principle, the structural model may account for this pattern with no additional features. If wages change over time in a way that induces workers to reallocate from some industries into others, the simple model with constant parameters κ and ν may be able to rationalize the data.

In addition, the estimated values of κ and ν in Table 5 are large when compared with average wages. The assumption of homogeneous workers can be biasing upwards the estimates of ν as the model is trying to explain the differences in average mobility rates across industries with differences in average industry wages. Then, controlling for worker heterogeneity can be important for the estimates of structural parameters.

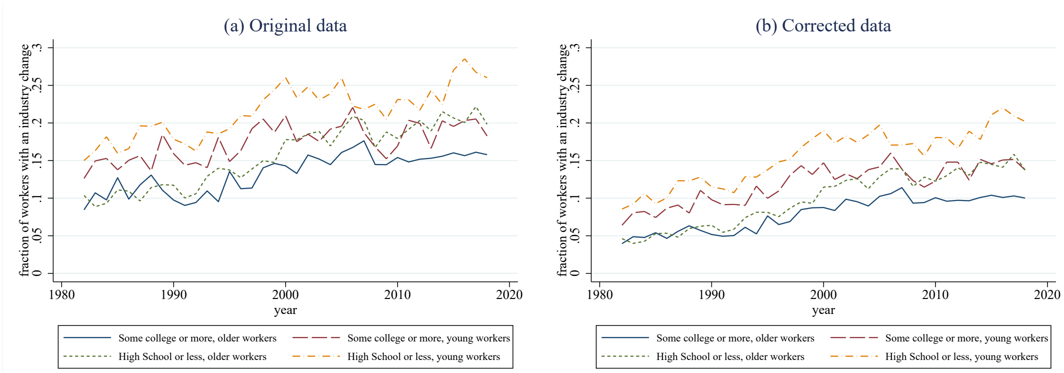
Since misclassification probabilities and mobility patterns differ by industry, the *observed* mobility flows of some groups of workers may be more affected by misclassification than that of others due to their industry choices.

In my analysis, I control for level of education, age, and time period. In the computation of corrected measures of mobility, age and time are numerical variables that vary by one unit every year. I discretize education into two groups, those with an education of high school or less and those with some years of college or more. Figure 5 shows average (gross) mobility rates using original CPS coding (left panel) and the corrected measures (right panel) estimated with the EM algorithm. The different lines in each graph show the evolution over time for workers with different education levels and age.[80] The patterns that emerge across panels are striking and have been covered in the literature, for example in Kambourov and Manovskii (2008). Younger workers switch industries with a higher probability than older workers. Similarly, less educated workers have a higher switching probability than more educated workers. Comparing the left and right panels of the figure, we can gauge the effects of misclassification, where mobility rates tend to be 50 to 100 percent larger when computed using original codes relative to the estimated mobility corrected for misclassification.

Artuç et al. (2010) extends the model to allow workers to age probabilistically from

[80] Note that age is kept constant over time, so the rates are estimated for workers with the same age across different time periods.

Figure 5: Probability of broad industry switching for different workers corrected measure vs. original data



Note: Original data refers to yearly industry mobility computed using CPS industry data as originally coded. Corrected data are estimates of industry switching probability corrected for misclassification, estimated via EM algorithm. See text for details on the estimation method. Young workers are those with age between 25 and 30. Older workers are those with age between 45 and 50. Sample includes male workers only.

young to old. They estimate structural parameters controlling for worker heterogeneity and age groups. However, using a version of the model with deterministic aging is simpler. Let some of the characteristics of workers evolve deterministically over time and denote these characteristics by a, like age. As before, characteristics τ are invariant over time for a worker. For simplicity, assume structural parameters ν and κ are constant over time, but κ can vary with workers' characteristics. Then, we can write the estimating equation as,[81]

$$\left(\log\left[\mu_t^{ij}(\tau,a)\right] - \log\left[\mu_t^{ii}(\tau,a)\right] - \beta\left(\log\left[\mu_{t+1}^{ij}(\tau,a+1)\right] - \log\left[\mu_{t+1}^{jj}(\tau,a+1)\right]\right)\right) =$$

$$= \frac{\beta}{\nu}\left(w_{t+1}^{j}(\tau,a+1) - w_{t+1}^{i}(\tau,a+1)\right) - \frac{\left(\kappa^{ij}(\tau,a) - \beta\,\kappa^{ij}(\tau,a+1)\right)}{\nu} + e_{t+1}^{ij}.$$

Table 10 shows the estimates for different types of workers. Column 3 in the table estimates the model parameters using mobility rates and wages computed using original CPS coding. Column 4 estimates parameters using mobility and wage measures corrected for misclassification. First, comparing the estimates with those of an homogeneous moving

[81] See Appendix B for details.

Table 10: Parameter estimates under worker heterogeneity

		Using original coding	Corrected for misclassification
ν		1.48	1.70
		(0.15)	(0.74)
κ			
HS or less	young	4.27	9.55
		(0.64)	(4.38)
HS or less	old	4.33	11.95
		(0.61)	(5.31)
Some college +	young	4.18	10.61
		(0.58)	(4.67)
Some college +	old	4.24	13.01
		(0.59)	(5.60)

Note: Estimated using equation (15) and yearly information about industry mobility and industry wages, assuming $\beta = 0.97$. Parameter ν is assumed constant across characteristics. Parameter κ varies with worker characteristics as specified in the text. Worker defined as young if between 25 and 35 years old. Old worker defined between 36 and 64 years old. IV regressions instrumented using two-year lagged values of wages and mobility rates. Standard errors obtained via block bootstrap. Bootstrap simulations with top and bottom 2% values of ν trimmed.

cost assumption, estimates for ν and κ are somewhat smaller, but less precisely estimated. Estimates of ν using corrected mobility measures are 23% lower than the ones in Table 5, and differences in the estimates of κ depend on the worker's age group, byt can be 30% larger in some cases.[82] Second, across columns highlights the biases that misclassification produce on parameter estimates. In particular, the estimate of ν is close to 15% larger and estimates of κ can be up to three times larger when using corrected measures. Under this specification, parameters are less precisely estimated. In all cases, standard errors take into account that estimation is conducted in stages, and that later stages use estimated parameter values from previous stages.

[82] The estimated value of parameter ν in Table 10 using originally coded data is 1.47, with parameter κ close to 4.2. As a comparison, the estimate of ν in Artuç et al. (2010) allowing for worker observed heterogeneity is 1.6, with estimated κ between 3.5 and 10 (see their Table 8).