Supplementary Appendix A: Not for Publication

This appendix complements Section 2 of the paper and expands Online Appendix A. Here we quantify the extent to which coding errors affect the probability of an occupational change after a non-employment spell. This is done in two ways. (1) We exploit the change from independent to dependent interviewing that occurred across the 1985 and 1986 SIPP panels, which overlap between February 1986 to April 1987. The change of interviewing technique allows us to estimate a garbling matrix $\mathbf{\Gamma}$ whose elements denote the probability that a worker's true occupation i gets miscoded as another occupation j. (2) We then take advantage of the retrospective coding exercise done to occupational codes in the PSID. Retrospective coding improved the assignment of the occupational codes obtained during the 1970s, but did not affect the codes obtained during later years (see also Kambourov and Manovskii, 2008). We use probabilistic models to estimate the average effect retrospective coding had on the probability of an occupational change.

Across these data sets we obtain a very similar conclusion. Our correction method implies that on average 82% of the observed wave-to-wave occupational transitions after re-employment in the SIPP are genuine, were each wave covers a 4 months interval. In stark contrast, when pooling together workers who changed and did not change employers, we find that 40% of the observed changes among all workers are genuine. Similarly, retrospective coding in the PSID implies that 84% of the observed year-to-year occupational transitions after re-employment are genuine. When pooling together workers who changed and did not change employers, retrospective coding implies that only 44% are genuine. A key insight is that the same propensities to miscode occupations affect very differently the measured occupational change of employer movers and employer stayers. This is important as it implies one cannot use the error correction estimates obtained from pooled samples to adjust the occupational mobility statistics of those who changed employers through spells of unemployment.

We use the estimated $\mathbf{\Gamma}$-correction matrix to eliminate the impact of coding errors on the occupational mobility statistics of the unemployed obtained from the SIPP. We show that occupational miscoding increases observed excess mobility and reduces the importance of net mobility (see also Kambourov and Manovskii, 2013). It also makes occupational mobility appear less responsive to unemployment duration and the business cycle. In Supplementary Appendix B we use the SIPP to evaluate robustness of the occupational mobility patterns derived from our Γ-correction method. To do so, we follow two alternative approaches of measuring occupational mobility: (i) simultaneously mobility of major occupational and major industrial groups at re-employment and (ii) self-reported duration of occupational tenure obtained from the topical modules. The first measure is considered less sensitive to miscoding as it typically requires errors to be made simultaneously along two dimensions. The second captures the worker's own perception of occupational mobility and is not based on occupational coding. We find that the occupational mobility patterns obtained through our classification error model applied to the SIPP, the retrospective coding in the PSID and these alternative measures provide a very consistent picture.

1. A classification error model

In Section 2.1 of the main text we defined the matrix $\mathbf{\Gamma}$, where its elements are the probabilities that an occupation i is miscoded as an occupation j, for all $i, j = 1, 2, ..., O$. There is an important literature that investigates classification error models. Magnac and Visser (1999) identify two main branches. The first one uses assumptions on the measurement error process and auxiliary data on the error rates, where it is assumed that the error rates can be directly observe from the auxiliary data (see Abowd and Zellner, 1985, Poterba and Summers, 1986, and Magnac and Visser, 1999, among others). The second one does not rely on auxiliary data on error rates, but estimates parametric models in the presence of misclassified data, where these models are either reduced form statistical models (see Hausman et al. 1998, among others) or structural economic models (see Kean and Wolpin, 2001, Sullivan, 2009, Roys and Taber, 2017, among others).

Our classification error model builds on Abowd and Zellner (1985) and Poterba and Summer (1986) in that it uses a garbling matrix that captures the errors made in classifying workers. These authors investigate the misclassification of individuals' employment status in the CPS and use reported employment status at the original interview date, at the re-interviewing date (which occurred one week after the original interview) and the reconciliation information provided by the CPS to directly observe the garbling matrix. Their key assumption is that the reconcile information provides the true individuals' labor market status. In contrast, we do not have auxiliary information that allows us to directly recover the garbling matrix. Our challenge is to estimate the garbling matrix.

Our classification error model also relates to Sullivan (2009) in that both approaches provide estimates of an occupation garbling matrix. In contrast, our approach does not rely on the observed optimal choices of economic agents for identification. Rather, miscoding can be estimated using the occupational transitions alone, without requiring further information on wages which themselves can be subject to measurement error. Therefore, it is much simpler to implement relative to Sullivan's (2009) approach which relies on time costly simulation maximum likelihood methods that end up restricting the size of the estimated garbling matrix. This is important as it reduces the overall computational burden when estimating our economic model.

To identify and estimate $\mathbf{\Gamma}$ we make three assumptions:

(*A1*) *Independent classification errors*: conditional on the true occupation, the realization of an occupational code does not depend on workers' labor market histories, demographic characteristics or the time it occurred in our sample. This assumption is also present in Poterba and Summers (1986) and Abowd and Zellner (1985) and is consistent with independent interviewing. In the standard practise of independent interviewing, professional coders base their coding on the verbatim description of the reported work activities without taking into account the respondents' demographic characteristics or earlier work history.[1] Errors introduced by the respondents, however, could be correlated with their characteristics. Assumption *A1* implies that errors in the individuals' verbatim responses are fully captured by the nature of the job these individuals are performing and hence only depend on their *true* occupation. Another implication of assumption *A1* is that $\mathbf{\Gamma}$ is time-invariant. This is important as we will apply our correction method across all years in our sample and one could be concern whether the coding errors estimated in the 1980's are similar to those founds 20 years later. In Section 4 (below) we investigate further these implications.

(*A2*) *"Detailed balance" in miscoding*: $\mathbf{diag}(\mathbf{c})\mathbf{\Gamma}$ is symmetric, where $\mathbf{c}$ is a Ox1 vector that describes the distribution of workers across occupations and $\mathbf{diag}(\mathbf{c})$ is the diagonal matrix of $\mathbf{c}$. This assumption is known as "detailed balance". It implies that the number of workers whose true occupation i gets mistakenly coded as j is the same as the number of workers whose true occupation j gets mistakenly coded as i, such that the overall size of occupations do not change with coding error. This assumption allow us to invert Γ and hence aids our identification arguments. Although undeniably strong, this assumption is a weaker version of the one proposed by Keane and Wolpin (2001) and subsequently used Roys and Taber (2017) to correct of occupation miscoding. We return to discuss *A2* in Section 4.

(*A3*) *Strict diagonal dominance*: $\mathbf{\Gamma}$ is strictly diagonally dominant in that $\gamma_{ii} > 0.5$ for all $i = 1, 2, ..., O$.

[1]For example, during the 1980s and 1990s independent occupational coding in the PSID was done without reference to respondents' characteristics or their work history. However, this information was used in the retrospective coding exercise done to the 1970s occupational codes.

This assumption is also present in Hausman et al., (1998) and implies that it is more likely to correctly code a given occupation i than to miscoded it. The converse would imply occupational mobility rates that are of a magnitude inconsistent with our data. Indeed, we derive an upper bound on code error directly from the data and find that *A3* is verified in our data.

To estimate $\mathbf{\Gamma}$ we exploit the change of survey design between the 1985 and 1986 SIPP panels. Until the 1985 panel the SIPP used independent interviewing for all workers: in each wave all workers were asked to describe their job anew, without reference to answers given at an earlier date. Subsequently, a coder would consider that wave's verbatim descriptions and allocate occupational codes. This practise is known to generate occupational coding errors. In the 1986 panel, instead, the practise changed to one of dependent interviewing (see Lynn and Sala, 2007, Jäckle, 2008, and Jäckle and Eckman, 2020). Respondents were only asked "independently" to describe their occupation if they reported a change in employer or if they reported a change in their main activities without an employer change within the last 8 months. If respondents declared no change in employer *and* in their main activities, the occupational code assigned to the respondent in the previous wave is carried forward.

It is important to note that during February 1986 to April 1987, the 1985 and 1986 panels overlap, representing the *same* population under different survey designs. The identification theory we develop in the next section refers to this population. We then show how to consistently estimate $\mathbf{\Gamma}$ using the population samples.

1.1 Identification of $\mathbf{\Gamma}$

Consider the population represented by 1985/86 panels during the overlapping period and divide it into two groups of individuals across consecutive interviews by whether or not they changed employer or activity. Label those workers who stayed with their employers in both interviews and did not change activity as "employer/activity stayers". By design this group *only* contains true occupational stayers. Similarly, label those workers who changed employers or changed activity within their employers as "employer/activity changers". By design this group contains all true occupational movers and the set of true occupational stayers who changed employers.

Suppose that we were to subject the employer/activity stayers in this population to dependent interviewing as applied in the 1986 panel. Let $\mathbf{c_s}$ denote the Ox1 vector that describes their *true* distribution across occupations and let $\mathbf{M_s} = \mathbf{diag}(\mathbf{c_s})$. In what follows we will use the convention that the (i, j)'th element of an M matrix indicates the flow from occupation i to j. Let $\mathbf{c_s^D}$ denote the Ox1 vector that describes their *observed* distribution across occupations under dependent interviewing and let $\mathbf{M_s^D} = \mathbf{diag}(\mathbf{c_s^D})$. Note that $\mathbf{c_s^D} = \mathbf{\Gamma}'\mathbf{M_s}\overrightarrow{\mathbf{1}}$, where $\overrightarrow{\mathbf{1}}$ describes a vector of ones. $\mathbf{M_s}$ is pre-multiplied by $\mathbf{\Gamma}'$ as true occupations would have been miscoded in the first of the two consecutive interviews. "Overall balance", a weaker version of *A2*, implies that $\mathbf{c_s^D} = \mathbf{diag}(\mathbf{c_s})\mathbf{\Gamma}\overrightarrow{\mathbf{1}} = \mathbf{c_s}$ and hence $\mathbf{M_s^D} = \mathbf{M_s}$.[2]

Next suppose that instead we were to subject the employer/activity stayers in this population to independent interviewing as applied in the 1985 panel. Let $\mathbf{M_s^I}$ denote the matrix that contains these workers' *observed* occupational transition *flows* under independent interviewing. In this case $\mathbf{M_s^I} = \mathbf{\Gamma}'\mathbf{M_s}\mathbf{\Gamma}$. Here $\mathbf{M_s}$ is pre-multiplied by $\mathbf{\Gamma}'$ and post-multiplied by $\mathbf{\Gamma}$ to take into account that the observed occupations of origin and destination would be subject to coding error.

Let $\mathbf{M_m}$ denote the matrix that contains the *true* occupational transition *flows* of employer/activity changers in this population. The diagonal of $\mathbf{M_m}$ describes the distribution of true occupational stayers across occupa-

[2]Overall balance only requires that classification errors do not change the overall occupational distribution rather than the bilateral flows between occupations as also required by detailed balance in *A2*.

tions among employer/activity changers. The off-diagonal elements contain the flows of all true occupational movers. Under independent interviewing we observe $\mathbf{M^I_m} = \mathbf{\Gamma' M_m \Gamma}$. Once again $\mathbf{M_m}$ is pre-multiplied by $\mathbf{\Gamma'}$ and post-multiplied by $\mathbf{\Gamma}$ as the observed occupations of origin and destination would be subject to coding error.

Letting $\mathbf{M^I} = \mathbf{M^I_m} + \mathbf{M^I_s}$ denote the matrix that contains the aggregate occupational transition flows across two interview dates under independent interviewing, it follows that $\mathbf{M^I_s} = \mathbf{M^I} - \mathbf{M^I_m} = \mathbf{\Gamma' M_s \Gamma}$. By virtue of the symmetry of $\mathbf{M_s}$ and "detailed balance" (*A2*), $\mathbf{M_s\Gamma} = \mathbf{\Gamma' M_s}' = \mathbf{\Gamma' M_s}$. Substituting back yields $\mathbf{M^I_s} = \mathbf{M_s \Gamma\Gamma}$. Next note that $\mathbf{M^I_s} = \mathbf{M_s T^I_s}$, where $\mathbf{T^I_s}$ is the occupational transition probability matrix of the employer/activity stayers in this population *observed* under independent interviewing. Substitution yields $\mathbf{M_s T^I_s} = \mathbf{M_s \Gamma\Gamma}$. Multiply both sides by $\mathbf{M_s^{-1}}$, which exists as long as all the diagonal elements of $\mathbf{M_s}$ are non-zero, yields the key relationship we exploit to estimate $\mathbf{\Gamma}$,

$$\mathbf{T^I_s} = \mathbf{\Gamma\Gamma}. \tag{1}$$

To use this equation we first need to show that it implies a unique solution for $\mathbf{\Gamma}$. Towards this result, we now establish that $\mathbf{\Gamma}$ and $\mathbf{T^I_s}$ are diagonalizable. For the latter it is useful to interpret the coding error process described above as a Markov chain such that $\mathbf{\Gamma}$ is the one-step probability matrix associated with this process.

Lemma A.1: *Assumptions A2 and A3 imply that* $\mathbf{\Gamma}$ *and* $\mathbf{T^I_s}$ *are diagonalizable.*

Proof. First note that without loss of generality we can consider the one-step probability matrix $\mathbf{\Gamma}$ to be irreducible. To show this suppose that $\mathbf{\Gamma}$ was not irreducible, we can (without loss of generality) apply a permutation matrix to re-order occupations in $\mathbf{\Gamma}$ and create a block-diagonal $\mathbf{\Gamma'}$, where each block is irreducible and can be considered in isolation. Given *A3*, it follows directly that $\mathbf{\Gamma}$ is aperiodic. Further, assumption *A2* implies that $\mathbf{c_s}$ is a stationary distribution of $\mathbf{\Gamma}$. The fundamental theorem of Markov chains then implies that $\mathbf{c_s}$ is the *unique* stationary distribution of $\mathbf{\Gamma}$. Assumption *A2* also implies that the Markov chain characterised by $\mathbf{\Gamma}$ is reversible with respect to $\mathbf{c_s}$. This means that $\mathbf{\Gamma}$ is similar (in matrix sense) to a symmetric matrix $\mathbf{G}$ such that $\mathbf{G} = \mathbf{diag}(\sqrt{\mathbf{c_s}})\, \mathbf{\Gamma}\, \mathbf{diag}(\sqrt{\mathbf{c_s}})^{-1}$. By symmetry, $\mathbf{G}$ is orthogonally diagonalizable by $\mathbf{Q\Delta Q^{-1}}$, where the diagonal matrix $\mathbf{\Delta}$ contains the associated (real) eigenvalues and $\mathbf{Q}$ is the orthogonal matrix of associated (normalized) eigenvectors. It then follows that $\mathbf{\Gamma}$ is diagonalizable as well. Further, $\mathbf{T^I_s} = \mathbf{diag}(\sqrt{\mathbf{c_s}})^{-1}\, \mathbf{G\, G}\, \mathbf{diag}(\sqrt{\mathbf{c_s}}) = \mathbf{diag}(\sqrt{\mathbf{c_s}})^{-1}\, \mathbf{Q}\, \mathbf{\Delta^2}\, \mathbf{Q^{-1}}\, \mathbf{diag}(\sqrt{\mathbf{c_s}})$, and hence $\mathbf{T^I_s}$ is also orthogonally diagonalizable, with a root of $\mathbf{P\Lambda^{0.5}P^{-1}}$, where $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues of $\mathbf{T^I_s}$, and $\mathbf{P}$ the associated orthogonal matrix with eigenvectors of $\mathbf{T^I_s}$. □

In general one cannot guarantee the uniqueness, or even existence, of a transition matrix that is the (nth) root of another transition matrix (see Higham and Lin, 2011). Here, however, existence is obtained by construction: $\mathbf{T_s}$ is constructed from $\mathbf{\Gamma}$, and in reverse, we can find its roots. The next result shows that $\mathbf{T_s}$ has a unique root satisfying assumptions *A2* and *A3*.

Proposition A.1: $\mathbf{\Gamma}$ *is the unique solution to* $\mathbf{T^I_s} = \mathbf{\Gamma\, \Gamma}$ *that satisfies assumptions A2 and A3. It is given by* $\mathbf{P\Lambda^{0.5}P^{-1}}$*, where* $\mathbf{\Lambda}$ *is the diagonal matrix with eigenvalues of* $\mathbf{T^I_s}$, $0 < \lambda_i \leq 1$, *and* $\mathbf{P}$ *is the orthogonal matrix with the associated (normalized) eigenvectors.*

Proof. Following from the proof of Lemma A.1, a root of $\mathbf{T^I_s}$ is given by $\mathbf{P\Lambda^{0.5}P^{-1}}$, where $\mathbf{\Lambda}$ is the diagonal matrix with eigenvalues of $\mathbf{T^I_s}$ and $\mathbf{P}$ is the orthogonal matrix with the associated (normalized) eigenvectors. Since *A3* implies $\mathbf{\Gamma}$ is strictly diagonally dominant, it follows that the determinant of all its leading principal minors are positive. Moreover, under the similarity transform by pre-/post-multiplication with the diagonal matrices $\mathbf{diag}(\sqrt{\mathbf{c_s}}), \mathbf{diag}(\sqrt{\mathbf{c_s}})^{-1}$, the determinant of all principals minors of the symmetric matrix $\mathbf{G} = \mathbf{diag}(\sqrt{\mathbf{c_s}})\ \mathbf{\Gamma}\ \mathbf{diag}(\sqrt{\mathbf{c_s}})^{-1}$ are positive as well. Hence $\mathbf{G}$ is a symmetric positive definite matrix (with all eigenvalues between 0 and 1, as has $\mathbf{\Gamma}$). It follows that $\mathbf{G}\,\mathbf{G} = \mathbf{S}$ is also positive definite, and $\mathbf{T_s^I} = \mathbf{diag}(\sqrt{\mathbf{c_s}})^{-1}\ \mathbf{S}\ \mathbf{diag}(\sqrt{\mathbf{c_s}})$ is positive definite in the sense that $\mathbf{v}'\ \mathbf{T_s^I}\ \mathbf{v} > 0$ for all $\mathbf{v} \neq \mathbf{0}$, while also all eigenvalues of $\mathbf{T_s^I}$ will be between 0 and 1.

To show the uniqueness of the root of $\mathbf{T_s^I}$ suppose (towards a contradiction) that there exists two different roots $\mathbf{\Gamma}$ and $\mathbf{\Upsilon}$ such that each are similar (in matrix sense), with the same transform involving $\mathbf{diag}(\sqrt{\mathbf{c_s}})$, to different symmetric positive definite matrices $\mathbf{G}$ and $\mathbf{Y}$, where $\mathbf{G}\,\mathbf{G} = \mathbf{S}$ and $\mathbf{Y}\,\mathbf{Y} = \mathbf{S}$. Both $\mathbf{G}$ and $\mathbf{Y}$ are diagonalizable, and have the square roots of the eigenvalues of $\mathbf{S}$ on the diagonal. Given that the squares of the eigenvalues need to coincide with the eigenvalues of $\mathbf{S}$ and assumptions *A2* and *A3* imply that all eigenvalues must be between 0 and 1, without loss of generality we can consider both diagonalizations to have the same diagonal matrix $\mathbf{\Delta}$, where $\mathbf{\Delta}$ is the diagonal matrix of eigenvalues of $\mathbf{T_s^I}$ and these eigenvalues are ordered using a permutation-similarity transform with the appropriate permutation matrices. Let $\mathbf{G} = \mathbf{H\Delta H^{-1}}$ and $\mathbf{Y} = \mathbf{K\Delta K^{-1}}$. Then, it follows that $\mathbf{K^{-1}H\Delta^2H^{-1}K} = \mathbf{\Delta^2}$ and since $\mathbf{K^{-1}H}$ and $\mathbf{\Delta^2}$ commute, implies that $\mathbf{K^{-1}H}$ is a block-diagonal matrix with the size of the blocks corresponding to the multiplicity of squared eigenvalues. Again, since all eigenvalues of $\mathbf{\Delta}$ are positive, this equals the multiplicity of the eigenvalues δ_i itself. But then it must be true that $\mathbf{K^{-1}H\Delta H^{-1}K} = \mathbf{\Delta}$. Then, $\mathbf{G} = \mathbf{H}\ \mathbf{\Delta}\ \mathbf{H^{-1}} = \mathbf{KK^{-1}H}\ \mathbf{\Delta}\ \mathbf{H^{-1}KK^{-1}} = \mathbf{K}\ \mathbf{\Delta}\ \mathbf{K^{-1}} = \mathbf{Y}$ which leads to a contradiction. □

The above results imply that under assumptions *A2* and *A3*, $\mathbf{\Gamma}$ is uniquely identified from the transition matrix of true occupational stayers under independent interviewing, $\mathbf{T_s^I}$.

1.2 Estimation of $\mathbf{\Gamma}$

The next lemma provides an intermediate step towards estimating $\mathbf{\Gamma}$. For this purpose let $PDT(.)$ denote the space of transition matrices that are similar, in the matrix sense, to positive definite matrices.

Lemma A.2: *The function* $f : PDT(\mathbb{R}^{O\times O}) \to PDT(\mathbb{R}^{O\times O})$ *given by* $f(\mathbf{T}) = \mathbf{T^{0.5}}$ *exists and is continuous with* $f(\mathbf{T_s^I}) = \mathbf{\Gamma}$ in the spectral matrix norm.

Proof. Existence follows from Lemma A.1 and Proposition A.1. To establish continuity of the mapping, we follow Horn and Johnson (1990). Let $\mathbf{T_1}$ and $\mathbf{T_2}$ be any two transition matrices in PDT and let $\mathbf{U_1}$ and $\mathbf{U_2}$ be two symmetric positive definite matrices constructed as $\mathbf{U_1} = \mathbf{diag}(\sqrt{\mathbf{c_1}})\ \mathbf{T_1}\ \mathbf{diag}(\sqrt{\mathbf{c_1}})^{-1}$ and $\mathbf{U_2} = \mathbf{diag}(\sqrt{\mathbf{c_2}})\ \mathbf{T_2}\ \mathbf{diag}(\sqrt{\mathbf{c_2}})^{-1}$, where $\mathbf{c_1}$ and $\mathbf{c_2}$ are the unique stationary distributions associated with $\mathbf{T_1}$ and $\mathbf{T_2}$, respectively. We want to show that if $\mathbf{U_1} \to \mathbf{U_2}$, then $\mathbf{U_1^{0.5}} \to \mathbf{U_2^{0.5}}$. First, note that $\|\mathbf{U_1} - \mathbf{U_2}\|_2 = \|\mathbf{U_1^{0.5}}(\mathbf{U_1^{0.5}} - \mathbf{U_2^{0.5}}) + (\mathbf{U_1^{0.5}} - \mathbf{U_2^{0.5}})\mathbf{U_2^{0.5}}\|_2 \geq |\mathbf{x}'\mathbf{U_1^{0.5}}(\mathbf{U_1^{0.5}} - \mathbf{U_2^{0.5}})\mathbf{x} + \mathbf{x}'(\mathbf{U_1^{0.5}} - \mathbf{U_2^{0.5}})\mathbf{U_2^{0.5}}\mathbf{x}|$, where $\mathbf{x}$ is any normalised vector. Assumptions *A2* and *A3* imply $\mathbf{U_1^{0.5}} - \mathbf{U_2^{0.5}}$ exists and is a symmetric matrix. Let $|\lambda| = \rho(\mathbf{U_1^{0.5}} - \mathbf{U_2^{0.5}})$ be the absolute value of the largest eigenvalue of $\mathbf{U_1^{0.5}} - \mathbf{U_2^{0.5}}$ and let $\mathbf{z}$ be the normalized eigenvector associated with λ. Note that $\|\mathbf{U_1} - \mathbf{U_2}\|_2 = |\lambda|$ and $(\mathbf{U_1^{0.5}} - \mathbf{U_2^{0.5}})\mathbf{z} = \lambda\mathbf{z}$. Then $|\mathbf{z}'\mathbf{U_1^{0.5}}(\mathbf{U_1^{0.5}} - \mathbf{U_2^{0.5}})\mathbf{z} + \mathbf{z}'(\mathbf{U_1^{0.5}} - \mathbf{U_2^{0.5}})\mathbf{U_2^{0.5}}\mathbf{z}| \geq |\lambda|\ |\lambda_{min}^{0.5}(\mathbf{U_1}) + \lambda_{min}^{0.5}(\mathbf{U_2})| = \|\mathbf{U_1^{0.5}} - \mathbf{U_2^{0.5}}\|_2\ (\lambda_{min}^{0.5}(\mathbf{U_1}) + \lambda_{min}^{0.5}(\mathbf{U_2}))$, where $\lambda_{min}(\mathbf{U_1})$ denotes the smallest eigenvalue of $\mathbf{U_2}$, which is positive by virtue of assumptions *A2* and *A3*. Then choose a $\delta = \varepsilon\lambda_{min}^{0.5}(\mathbf{U_1})$. It follows that if $\|\mathbf{U_1} - \mathbf{U_2}\|_2 < \delta$, then $\|\mathbf{U_1^{0.5}} - \mathbf{U_2^{0.5}}\|_2 \times \frac{(\lambda_{min}^{0.5}(\mathbf{U_1}) + \lambda_{min}^{0.5}(\mathbf{U_2}))}{\lambda_{min}^{0.5}(\mathbf{U_1})} < \varepsilon$, and therefore $\|\mathbf{U_1^{0.5}} - \mathbf{U_2^{0.5}}\|_2 < \varepsilon$, which establishes the desired continuity. From the fact that $\mathbf{U_1} \to \mathbf{U_2}$ implies $\mathbf{U_1^{0.5}} \to \mathbf{U_2^{0.5}}$, it then also follows that $f(\mathbf{T})$ is continuous. □

Let $\hat{\mathbf{T}}^{\mathbf{I}}_{\mathbf{s}}$ denote the sample estimate of $\mathbf{T}^{\mathbf{I}}_{\mathbf{s}}$ and let $\hat{\mathbf{\Gamma}}$ be estimated by the root $(\hat{\mathbf{T}}^{\mathbf{I}}_{\mathbf{s}})^{0.5} \in PDT(\mathbb{R}^{O\times O})$ such that $\hat{\mathbf{\Gamma}} = (\hat{\mathbf{T}}^{\mathbf{I}}_{\mathbf{s}})^{\mathbf{0.5}} = \hat{\mathbf{P}}\hat{\mathbf{\Lambda}}^{\mathbf{0.5}}\hat{\mathbf{P}}^{-\mathbf{1}}$, where $\hat{\mathbf{\Lambda}}$ is the diagonal matrix with eigenvalues of $\hat{\mathbf{T}}^{\mathbf{I}}_{\mathbf{s}}$, $0 < \hat{\lambda}_i^{0.5} \leq 1$ and $\hat{\mathbf{P}}$ the orthogonal matrix with the associated (normalized) eigenvectors. We then have the following result.

Proposition A.2: $\mathbf{\Gamma}$ *is consistently estimated from* $(\hat{\mathbf{T}}^{\mathbf{I}}_{\mathbf{s}})^{0.5} \in PDT(\mathbb{R}^{O\times O})$ *such that* $\hat{\mathbf{\Gamma}} = (\hat{\mathbf{T}}^{\mathbf{I}}_{\mathbf{s}})^{\mathbf{0.5}} = \hat{\mathbf{P}}\hat{\mathbf{\Lambda}}^{\mathbf{0.5}}\hat{\mathbf{P}}^{-\mathbf{1}}$. *That is,* $plim_{n\to\infty}\hat{\mathbf{\Gamma}} = \mathbf{\Gamma}$.

Proof. From Lemma A.1 and Proposition A.1 it follows that if we know $\mathbf{T}^{\mathbf{I}}_{\mathbf{s}}$, then we can find the unique $\mathbf{\Gamma}$ that underlies it, constructing it from the eigenvalues and eigenvectors of $\mathbf{T}^{\mathbf{I}}_{\mathbf{s}}$. To estimate $\mathbf{T}^{\mathbf{I}}_{\mathbf{s}}$ one can use the sample proportion $\hat{\mathbf{M}}_{\mathbf{ij}}/\sum_{k=1}^{O}\hat{\mathbf{M}}_{\mathbf{ik}}$ and note that this converges in probability to $(\mathbf{T}^{\mathbf{I}}_{\mathbf{s}})_{\mathbf{ij}}$ (see Anderson and Goodman, 1957; Billingsley 1961, thm 1.1-3.) for all occupations, given assumptions *A2* and *A3*. Hence, $plim_{n\to\infty}\hat{\mathbf{T}}^{\mathbf{I}}_{\mathbf{s}} = \mathbf{T}^{\mathbf{I}}_{\mathbf{s}}$. Then, we derive $\hat{\mathbf{\Gamma}}$ from $\hat{\mathbf{T}}^{\mathbf{I}}_{\mathbf{s}}$ according to $\hat{\mathbf{P}}\hat{\mathbf{\Lambda}}^{\mathbf{0.5}}\hat{\mathbf{P}}^{-\mathbf{1}}$, per Proposition A.1. By continuity of the mapping in Lemma A.2, it follows that $plim_{n\to\infty}\hat{\mathbf{\Gamma}} = \mathbf{\Gamma}$, and our estimator is consistent. □

Note that to identify and estimate $\mathbf{\Gamma}$ in the SIPP it is not sufficient to directly compare the aggregate occupational transition flows under independent interviewing with the aggregate occupational transition flows under dependent interviewing. To show this let $\mathbf{M}^{\mathbf{D}} = \mathbf{M}^{\mathbf{I}}_{\mathbf{m}} + \mathbf{M}^{\mathbf{D}}_{\mathbf{s}}$ denote the matrix that contains the aggregate occupational transition flows across two interview dates under dependent interviewing for employer/activity stayers and under independent interviewing for employer/activity movers. Subtracting $\mathbf{M}^{\mathbf{I}} = \mathbf{M}^{\mathbf{I}}_{\mathbf{m}} + \mathbf{M}^{\mathbf{I}}_{\mathbf{s}}$ from $\mathbf{M}^{\mathbf{D}}$ yields $\mathbf{M}^{\mathbf{D}}_{\mathbf{s}} - \mathbf{M}^{\mathbf{I}}_{\mathbf{s}} = \mathbf{M}_{\mathbf{s}} - \mathbf{\Gamma}'\mathbf{M}_{\mathbf{s}}\mathbf{\Gamma}$. Given the symmetry assumed in *A2*, the latter expression has $0.5n(n-1)$ exogenous variables on the LHS and $0.5n(n+1)$ unknowns (endogenous variables) on the RHS, leaving $\mathbf{\Gamma}$ (and $\mathbf{M}_{\mathbf{s}}$) unidentified.

In addition to $\mathbf{M}^{\mathbf{D}} - \mathbf{M}^{\mathbf{I}} = \mathbf{M}_{\mathbf{s}} - \mathbf{\Gamma}'\mathbf{M}_{\mathbf{s}}\mathbf{\Gamma}$ one can use $\mathbf{M}^{\mathbf{D}} = \mathbf{\Gamma}'\mathbf{M}_{\mathbf{m}}\mathbf{\Gamma} + \mathbf{M}_{\mathbf{s}}$, which contains the remainder information. When $\mathbf{M}_{\mathbf{m}}$ has mass on its diagonal, however, this additional system of equations has n^2 exogenous variables on the LHS and n^2 unknowns (arising from $\mathbf{M}_{\mathbf{m}}$) on the RHS. This implies that with the n unknowns remaining from $\mathbf{M}^{\mathbf{D}} - \mathbf{M}^{\mathbf{I}} = \mathbf{M}_{\mathbf{s}} - \mathbf{\Gamma}'\mathbf{M}_{\mathbf{s}}\mathbf{\Gamma}$, one is still unable to identify $\mathbf{\Gamma}$ and $\mathbf{M}_{\mathbf{s}}$.

Corollary A.1: *If* $\mathbf{M}_{\mathbf{m}}$ *has mass on its diagonal,* $\mathbf{\Gamma}$ *cannot be identified from* $\mathbf{M}^{\mathbf{I}}$ *and* $\mathbf{M}^{\mathbf{D}}$ *alone.*

The intuition behind this result is that by comparing aggregate occupational transition flows under dependent and independent interviewing, it is unclear how many workers are 'responsible' for the change in occupational mobility between $\mathbf{M}^{\mathbf{D}}$ and $\mathbf{M}^{\mathbf{I}}$. Only when the diagonal of $\mathbf{M}_{\mathbf{m}}$ contains exclusively zeros, identification could be resolved and one can recover $\mathbf{M}_{\mathbf{s}}$, $\mathbf{\Gamma}$ and $\mathbf{M}_{\mathbf{m}}$ as the number of equations equals the number of unknowns.[3] An implication of the above corollary is that interrupted time-series analysis that is based on the difference in occupational mobility at the time of a switch from independent to dependent interviewing, does not identify the precise extent of the average coding error, but provides a downwards biased estimate.

To identify $\mathbf{\Gamma}$, however, Proposition A.2 implies that one can use the observed occupational transition flows of a sample of *true* occupational stayers that are subject to two rounds of independent interviewing. Some of these workers will appear as occupational stayers and some of them as occupational movers. Ideally, such a sample of workers should be isolated directly from the 1985 panel. Unfortunately, the questions on whether the individual changed activity or employer were only introduced in the 1986 panel, as a part of the switch to dependent interviewing. As a result, the 1985 panel by itself does not provide sufficient information to separate employer/activity stayers from employer/activity movers. Instead we use 1986 panel to estimate $\hat{\mathbf{M}}^{\mathbf{I}}_{\mathbf{m}}$. We

[3] However, in the SIPP this case is empirically unreasonable as it requires that all employer/activity changers be true occupational movers.

can infer $\mathbf{M_s^I}$ indirectly by subtracting the observed occupational transition flow matrix $\hat{\mathbf{M}}_\mathbf{m}^\mathbf{I}$ in the 1986 panel from the observed occupational transition flow matrix $\hat{\mathbf{M}}^\mathbf{I}$ in the 1985 panel. This is possible as the 1986 panel refers to the same underlying population as the 1985 panel and separates the employer/activity changers, who are independently interviewed.

Corollary A.2: $\hat{\mathbf{\Gamma}}$ *is consistently estimated from* $\hat{\mathbf{T}}_\mathbf{s}^\mathbf{I}$ *when the latter is estimated from* $\hat{\mathbf{M}}^\mathbf{I} - \hat{\mathbf{M}}_\mathbf{m}^\mathbf{I}$

This result is important to implement our approach. It follows as the population proportions underlying each cell of $\hat{\mathbf{M}}_\mathbf{s}$, the sample estimate of $\mathbf{M_s}$, are consistently estimated. In turn, the latter follows from the standard central limit theorem for estimating proportions, which applies to $\hat{\mathbf{M}}^\mathbf{I}$, $\hat{\mathbf{M}}_\mathbf{m}^\mathbf{I}$ and its difference. Proposition A.2 then implies that $\hat{\mathbf{\Gamma}}$ is consistently estimated.

1.3 Implementation

To implement our correction method we take the overlapping period of the 1985/86 panels. To increase the sample size we also use observations from the 1987 panel for the period between February 1987 and April 1987.[4] This panel has an identical setup to that of the 1986 panel (dependent interviewing and other relevant aspects) and is likewise representative of the population during the period of study.

Interviews throughout the SIPP are conducted every four months and collect information pertaining to the last four months, where these four months are considered to be a wave. We compare the reported occupational code of a worker in a given interview with the reported occupational code of that worker in the subsequent interview. An observation is therefore a pair of occupational codes, a reported 'source' and (potentially identical) 'destination' occupation. To keep comparability across interviews as clean as possible, and to focus on measuring occupations in the primary job, we only consider those workers who throughout the two waves stayed in full-time employment and who reported having only one employer at any point in time. These restrictions imply that in the estimation of $\mathbf{\Gamma}$ we do not include non-temporary laid off workers who experienced a short unemployment episodes and returned to their same jobs and employers. We also restrict attention to those workers who do not have imputed occupations, were not enrolled in school and were between 19-66 years old. These restrictions yield 28,302 wave/individual observations for the 1985 panel, 27,801 wave/individual observations for the 1986 panel and 5,922 wave/individual observations for the 1987 panel.

Tables 1 and 2 show the demographic and occupational characteristics (based on the major occupations of the 1990 SOC), respectively, of the samples across the three panels. In the last column of each table we test characteristic-by-characteristic whether the proportion of workers with a given characteristic in the 1985 sample is statistically indistinguishable from the proportion of workers with the same characteristic in the pooled 1986/87 sample. Across all the characteristics analysed, we cannot reject at a 5% level that the proportions in the 1985 sample and the corresponding proportions in the 1986-87 sample are the same. With the exception of the proportion of Asian Americans and the proportion of workers whose source occupation is management, similarity cannot be rejected even at a 10% level. Although not shown here, we also cannot rejected at a 10% level that the proportion of workers across source and destination industries are the same when comparing the 1985 and 1986/87 samples. This analysis thus confirm that the observations used for our exercise are taken from the same underlying population.[5]

[4] To avoid seasonality effects we re-weight all observations such that each observation in a given month has the same weight as another observation in any other month.

[5] To further rule out any meaningful impact from the observed differences in occupational distributions, we re-calculated all statistics after re-balancing the weights on the source/destination occupations to create identical occupational distributions. This exercise yields minimal effects on our statistics. For example, the observed occupational mobility changes by 0.01 percentage point at most. The

Table 1: Demographic characteristics - February 1986 to April 1987

	SIPP 1985	SIPP 1986	SIPP 1987	p-value (no difference)
Education				
less than high school	14.72	15.27	14.55	0.386
high school grad	38.10	37.36	36.80	0.361
some college	24.51	24.94	24.71	0.546
college degree	22.67	22.43	23.95	0.746
Age category				
19-24	12.29	12.62	12.90	0.458
25-29	16.72	16.15	16.36	0.357
30-34	15.84	15.40	16.00	0.512
35-39	15.02	15.30	13.99	0.806
40-44	11.65	11.50	11.80	0.804
45-49	9.10	8.87	9.22	0.667
50-54	7.72	7.90	8.25	0.600
55-59	7.07	7.31	7.01	0.600
60-64	4.20	4.64	4.05	0.221
Ethicity				
white	86.29	86.57	86.25	0.729
black	10.62	10.81	10.73	0.782
american indian, eskimo	0.49	0.62	0.42	0.401
asian or pacific islander	2.60	2.01	2.60	0.090
Other				
men	54.20	55.27	53.93	0.112
married	65.51	66.15	64.54	0.550
living in metro area	76.33	76.24	75.97	0.885

Workers aged 19-66, not enrolled in school, in two adjacent waves, measured in the first month of the current wave, with employment in one firm only in the previous wave, and employment in one (but possibly different) employer only in the wave that follows, without any self-employment, with un-imputed occupations reported in both waves. Person weights are used to scale observations per month within panel group (1985 versus 1986+1987).

We estimate the occupational flows of employer/activity stayers by $\hat{\mathbf{M}}^{\mathbf{I,85}} - \hat{\mathbf{M}}_{\mathbf{m}}^{\mathbf{I,86/87}} = \hat{\mathbf{M}}_{\mathbf{s}}^{\mathbf{I}}$, where $x = 85$ ($x = 86/87$) in $\hat{\mathbf{M}}_{\mathbf{i}}^{\mathbf{I,x}}$ refers to the 1985 sample (1986/87 sample). In the 1986/87 sample, where we can observe employer/activity changers directly, we find that in 2.31% of (weighted) observations workers changed employers and in 4.65% workers report an activity change within their employers. This implies that more than 93% of the 1985 sample should be made up of employer/activity stayers.

A potential concern from using this survey design change could be the reliability in the 1986 implementation of dependent interviewing and its comparison with the data collected in the overlapping period of the 1985 panel. For example, any trail/learning period in the implementation of dependent interviewing in the 1986 panel could affect our results. To evaluate whether any trial/learning period in the implementation of dependent interviewing affected the occupational mobility rates we compare the occupational mobility rates obtained from the 1986 panel to those obtained from subsequent ones. Assuming improvements made after 1986 affected the measurement of occupational mobility, we should observe a meaningful change in these rates when using subsequent panels. Figure 1 (see below for a detail explanation of the graph) suggests that any trail/learning period in the implementation of dependent interviewing did not have a major effect on average occupational mobility.[6]

reason for this small change is because there is the large proportion of the implied true stayers in the sample, which means that $\hat{\mathbf{T}}_{\mathbf{s}}^{\mathbf{I}}$ is not very sensitive to proportional changes in $\hat{\mathbf{M}}_{\mathbf{m}}^{\mathbf{I}}$.

[6]This is consistent with Hill (1994), who instead points to a potential higher level of attrition in the "older" 1985 panel relative to the "newer" 1986 panel as a concern. He argues that using the "1986 final panel weights, taken from the 1985 and 1986 Full Panel Longitudinal Research Files", while not perfect, seem to be the best available solution to tackle this problem. We follow the same practice in our analysis.

Table 2: Distribution of workers across occupations - February 1986 to April 1987

Distribution across source occupations (occupation code) (%)				
	SIPP 1985	SIPP 1986	SIPP 1987	p-value (no difference)
managing occupations	12.31	13.11	13.99	0.064
professional speciality	13.40	12.92	13.18	0.380
technicians and related support	3.85	3.89	4.08	0.829
sales occ.	9.74	9.98	9.86	0.599
admin support	18.67	18.17	18.26	0.366
services	10.96	11.66	11.19	0.166
farming/fish/logging	1.09	1.08	1.03	0.940
mechanics and repairers	4.87	4.47	4.69	0.231
construction and extractive	3.47	3.65	3.60	0.503
precision production	4.01	4.19	3.75	0.643
machine operators/assemblers	9.27	8.89	8.63	0.355
transportation and materials moving	4.63	4.34	4.31	0.340
laborers	3.73	3.66	3.42	0.714
Distribution across destination occupations (occupation code) (%)				
managing occupations	12.58	13.27	14.18	0.103
professional speciality	13.31	12.90	13.10	0.451
technicians and related support	3.82	3.92	4.01	0.693
sales occ.	9.76	9.89	9.77	0.797
admin support	18.53	18.20	18.20	0.546
services	10.96	11.57	11.19	0.229
farming/fish/logging	1.08	1.07	1.03	0.924
mechanics and repairers	4.87	4.46	4.73	0.222
construction and extractive	3.59	3.60	3.56	0.954
precision production	4.01	4.26	3.77	0.480
machine operators/assemblers	9.24	8.93	8.77	0.455
transportation and materials moving	4.62	4.31	4.33	0.310
laborers	3.62	3.62	3.36	0.906

Workers aged 19-66, not enrolled in school, in two adjacent waves, measured in the first month of the current wave, with employment in one firm only in the previous wave, and employment in one (but possibly different) firm only in the wave that follows, without any self-employment, with un-imputed occupations reported in both waves. Person weights are used to scale observations per month within panel group (1985 versus 1986+1987).

2. Results and Discussion

Table 3 shows the occupational transition matrix $\hat{\mathbf{T}}^{\mathbf{I}}_{\mathrm{s}}$ for true occupational stayers derived from $\hat{\mathbf{M}}^{\mathbf{I}}_{\mathrm{s}}$ using the 1985 sample based on the major occupations of the 1990 SOC. The (i, j)'th element of $\hat{\mathbf{T}}^{\mathbf{I}}_{\mathrm{s}}$ indicates the transition probability from occupation i to j. This matrix implies that 18.46% of true occupational stayers get classified as occupational movers.[7] Although not shown here, a similar conclusion arises when we calculate the same matrix based on the 2000 SOC, which we use for our results in the main text.

2.1 The estimate of $\mathbf{\Gamma}$

Following Proposition *A.1* we can then recover the garbling matrix $\mathbf{\Gamma}$ by using $\hat{\mathbf{T}}^{\mathbf{I}}_{\mathrm{s}}$ and equation (1). Table 4 shows the estimated $\hat{\mathbf{\Gamma}}$ based on the major occupations of the 1990 SOC, while Table 5 shows the estimated $\hat{\mathbf{\Gamma}}$ based on the major occupations of the 2000 SOC. These estimates imply that on average the incorrect occupational code is assigned in around 10% of the cases. Since a spurious transition is likely to be created when

[7]This value lies within the expected bounds. To construct an upper bound consider the 1985 sample and assume that all observed occupational transitions are spurious. In this case we can expect that at most 19.71% of the observations would be miscoded. To construct a lower bound consider the 1986/87 sample and calculate the number of observations in which an occupational move is reported among the employer/activity changers. These observations account for 2.60% of all observations in the 1986/87 sample. Assuming that the effect of miscoding is to generate a net increase in the number of occupational changes, we can expect that at least 19.71%-2.60%=17.11% of the observations would be miscoded.

Table 3: Observed occupational transition matrix of true occupational stayers, SOC 1990, $\hat{\mathbf{T}}^{\mathbf{I}}_{\mathrm{s}}$, (%)

OCCUPATIONS	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
(1) Managing Occ.	**75.7**	3.7	1.0	5.1	8.8	1.5	0.1	0.7	0.7	1.5	0.5	0.4	0.3
(2) Professional Spec.	3.3	**87.9**	2.9	0.4	2.2	1.5	0.1	0.5	0.2	0.6	0.2	0.1	0.1
(3) Technicians	3.1	10.1	**69.5**	0.5	5.9	3.8	0.1	2.7	0.6	1.3	2.1	0.2	0.3
(4) Sales Occ.	6.5	0.5	0.2	**83.3**	3.8	1.5	0.1	0.7	0.2	0.5	0.2	0.8	1.8
(5) Admin. Support	5.8	1.6	1.2	2.0	**84.5**	1.0	0.1	0.3	0.1	0.5	0.7	0.6	1.6
(6) Services	1.7	1.8	1.3	1.3	1.8	**87.2**	0.2	1.3	0.7	0.5	0.6	0.6	0.9
(7) Farm/Fish/Logging	1.3	1.7	0.5	1.3	1.1	2.0	**84.7**	0.6	0.6	0.2	1.0	3.1	2.0
(8) Mechanics	1.8	1.3	2.1	1.3	1.1	2.9	0.1	**79.2**	2.1	2.6	3.0	0.8	1.6
(9) Construction	2.4	0.7	0.6	0.5	0.4	2.2	0.2	3.0	**77.7**	1.7	1.9	1.2	7.6
(10) Precision Prod.	4.5	2.0	1.2	1.1	2.4	1.4	0.0	3.2	1.5	**69.1**	11.1	0.2	2.3
(11) Mach. Operators	0.7	0.3	0.9	0.2	1.4	0.7	0.1	1.6	0.7	4.8	**84.1**	0.7	3.8
(12) Transport	1.1	0.4	0.2	1.6	2.2	1.4	0.7	0.8	0.9	0.2	1.5	**85.1**	3.9
(13) Laborers	1.0	0.4	0.3	4.6	8.2	2.7	0.6	2.1	7.3	2.5	9.7	5.0	**55.7**

either the source or destination occupation is miscoded, the probability of observing a spurious transition for a true occupational stayer is nearly twice as large. Our methodology then suggests that coding error is indeed substantial under independent interviewing. Its magnitude is of similar order as found in other studies analysing the extent of errors in occupational coding (see Campanelli et al., 1997, Sullivan, 2009, Roys and Taber, 2017, and vom Lehn et al., 2021).

Table 4: Estimate of the garbling matrix, SOC 1990, $\hat{\mathbf{\Gamma}}$, (%)

OCCUPATIONS	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)
(1) Managing Occ.	**86.8**	2.0	0.5	2.8	4.9	0.8	0.1	0.4	0.4	0.8	0.2	0.2	0.1
(2) Professional Spec.	1.8	**93.7**	1.6	0.2	1.1	0.8	0.1	0.2	0.1	0.3	0.1	0.1	0.0
(3) Technicians	1.7	5.6	**83.3**	0.2	3.3	2.1	0.1	1.5	0.3	0.7	1.1	0.1	0.1
(4) Sales Occ.	3.6	0.2	0.1	**91.2**	1.9	0.8	0.1	0.3	0.1	0.2	0.0	0.4	1.0
(5) Admin. Support	3.2	0.8	0.7	1.0	**91.8**	0.5	0.0	0.1	0.0	0.3	0.3	0.3	1.0
(6) Services	0.9	0.9	0.7	0.7	0.9	**93.3**	0.1	0.7	0.4	0.3	0.3	0.3	0.5
(7) Farm/Fish/Logging	0.7	0.9	0.3	0.7	0.5	1.0	**92.0**	0.3	0.3	0.1	0.5	1.7	1.1
(8) Mechanics	1.0	0.6	1.2	0.7	0.5	1.5	0.1	**88.9**	1.2	1.5	1.6	0.4	0.9
(9) Construction	1.3	0.3	0.3	0.2	0.0	1.1	0.1	1.6	**88.0**	0.9	0.8	0.6	4.7
(10) Precision Prod.	2.6	1.0	0.7	0.6	1.2	0.7	0.0	1.8	0.8	**83.0**	6.3	0.1	1.3
(11) Mach. Operators	0.3	0.1	0.5	0.0	0.7	0.4	0.1	0.8	0.3	2.7	**91.5**	0.4	2.3
(12) Transport	0.5	0.2	0.1	0.8	1.1	0.7	0.4	0.4	0.4	0.0	0.7	**92.2**	2.3
(13) Laborers	0.3	0.1	0.1	2.7	4.8	1.5	0.3	1.2	4.5	1.4	5.7	3.0	**74.3**

Two additional messages come out of Tables 4 and 5. (i) Different occupations have very different propensities to be assigned a wrong code. For example, when using the 1990 SOC we find that individuals whose true occupation is "laborers" have a 74% probability of being coded correctly, while individuals whose true occupation is "professional speciality" have a 94% probability of being coded correctly. (ii) Given a true occupation, some coding mistakes are much more likely than others. For example, workers whose true occupation is "laborers" have a much larger probability to be miscoded as "machine operators" (5.7%), "construction" (4.5%) or "admin. support" (4.8%) than as "managers" (0.3%) or "professionals" (0.1%). Our methodology enable us to take these differences into account by correcting observed occupational flows by source-destination occupation pair. This provides cleaner net mobility estimates, where the identity of the origin and destination occupation matters.

Table 5: Estimate of the garbling matrix, SOC 2000, $\hat{\mathbf{\Gamma}}$, (%)

OCCUPATIONS	(11)	(13)	(15)	(17)	(19)	(21)	(23)	(25)	(27)	(29)	(31)	(33)	(35)	(37)	(39)	(41)	(43)	(45)	(47)	(49)	(51)	(53)
(11) Management Occ.	**84.2**	2.3	0.3	0.8	0.3	0.3	0.1	0.3	0.1	0.3	0.1	0.1	0.4	0.1	0.4	3.3	3.5	0.5	0.7	0.5	1.2	0.3
(13) Business & Finance Oper.	4.7	**82.8**	0.9	0.5	0.2	0.2	0.2	0.3	0.0	0.0	0.0	0.1	0.0	0.0	0.0	1.6	7.6	0.0	0.1	0.1	0.5	0.1
(15) Computer & Math. Occ.	2.6	3.7	**85.7**	1.6	0.4	0.0	0.0	0.4	0.5	0.0	0.1	0.1	0.0	0.0	0.0	0.0	4.4	0.0	0.1	0.7	0.2	0.2
(17) Architect & Eng. Occ.	2.2	0.7	0.5	**87.0**	1.3	0.0	0.0	0.0	1.2	0.3	0.0	0.0	0.0	0.1	0.0	0.1	1.1	0.0	0.9	2.4	2.2	0.1
(19) Life, Phys, and Soc. Sci. Occ.	2.4	0.6	0.4	3.3	**82.9**	0.0	0.2	1.1	0.4	2.2	0.5	0.0	0.0	0.5	0.0	0.3	1.7	0.7	0.5	1.1	1.0	0.3
(21) Comm & Soc. Service Occ.	2.8	0.7	0.0	0.0	0.0	**89.8**	0.0	1.0	0.4	0.5	0.7	0.3	0.2	0.1	0.4	0.0	2.6	0.0	0.2	0.1	0.1	0.3
(23) Legal	1.3	1.6	0.0	0.0	0.5	0.0	**93.0**	0.0	0.0	0.0	0.0	0.5	0.0	0.0	0.1	0.1	3.5	0.0	0.0	0.0	0.0	0.0
(25) Educ., Training & Library	0.5	0.3	0.1	0.0	0.3	0.2	0.0	**97.3**	0.1	0.2	0.1	0.0	0.0	0.0	0.3	0.1	0.6	0.0	0.0	0.0	0.0	0.0
(27) Arts, Dsgn, Ent., Sports & Media	1.0	0.2	0.6	3.8	0.5	0.4	0.0	0.3	**89.2**	0.0	0.0	0.0	0.0	0.0	0.0	0.7	1.3	0.0	0.2	0.2	1.7	0.4
(29) Healthcare Pract. & Tech. Occ.	0.7	0.0	0.0	0.2	0.7	0.1	0.0	0.3	0.0	**92.7**	3.3	0.0	0.1	0.1	0.1	0.1	1.0	0.0	0.0	0.1	0.2	0.1
(31) Healthcare Support	0.5	0.0	0.0	0.0	0.3	0.5	0.0	0.2	0.0	6.8	**88.5**	0.0	0.6	0.4	0.3	0.0	1.6	0.0	0.0	0.1	0.3	0.0
(33) Protective Service	0.7	0.3	0.1	0.0	0.0	0.2	0.2	0.1	0.1	0.1	0.0	**95.0**	0.5	0.1	0.2	0.1	1.8	0.3	0.0	0.1	0.4	0.5
(35) Food Prep/Serving & Rel.	0.9	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.3	0.2	**95.3**	0.5	0.0	1.5	0.4	0.0	0.0	0.0	0.4	0.2
(37) Building/grounds Clean. & Maint.	0.3	0.0	0.0	0.1	0.2	0.1	0.0	0.0	0.0	0.1	0.2	0.0	0.6	**90.3**	0.1	0.4	0.2	0.3	1.7	2.5	1.1	1.8
(39) Personal Care & Service Occ.	3.3	0.0	0.0	0.0	0.0	0.4	0.1	1.6	0.0	0.5	0.6	0.4	0.0	0.3	**91.4**	0.9	0.6	0.0	0.0	0.1	0.0	0.2
(41) Sales & Rel. Occ.	3.0	0.7	0.0	0.0	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	0.5	0.1	0.1	**91.7**	1.7	0.1	0.1	0.4	0.3	1.3
(43) Office & Admin. Support	1.7	1.8	0.3	0.2	0.1	0.2	0.1	0.2	0.1	0.2	0.2	0.2	0.1	0.0	0.0	0.9	**91.6**	0.0	0.0	0.1	0.6	1.3
(45) Farm, Fish. & Forestry	4.5	0.2	0.0	0.0	1.0	0.1	0.0	0.1	0.0	0.0	0.0	0.5	0.1	0.9	0.1	0.7	0.3	**87.5**	0.1	0.3	0.8	3.4
(47) Construction & Extraction	1.3	0.1	0.0	0.6	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.1	0.0	0.2	0.1	0.0	**90.8**	1.6	1.9	2.1
(49) Install., Maint. & Repair Occ.	0.8	0.1	0.1	1.4	0.2	0.0	0.0	0.0	0.0	0.1	0.0	0.0	0.0	1.5	0.0	0.6	0.4	0.1	1.4	**89.0**	3.0	1.1
(51) Production Occ.	0.8	0.2	0.0	0.5	0.1	0.0	0.0	0.0	0.1	0.1	0.0	0.0	0.1	0.3	0.0	0.2	0.8	0.0	0.6	1.1	**93.1**	1.9
(53) Transportation & Mater. Moving	0.3	0.1	0.0	0.1	0.0	0.0	0.0	0.0	0.1	0.1	0.0	0.1	0.1	0.7	0.0	1.5	2.9	0.4	1.2	0.7	3.4	**88.2**

2.2 The effect of $\hat{\boldsymbol{\Gamma}}$ on the gross occupational mobility rate

Figure 1 depicts the effect of $\hat{\boldsymbol{\Gamma}}$ when computing *wave-to-wave* occupational mobility rates using the major occupational groups of the 2000 SOC. For this exercise we augment the 1985/86/87 samples with workers from the 1984 and 1988 panels, which satisfy the same sample restrictions as before. Figure 1 depicts the average wave-to-wave occupational mobility rate obtained during each year. For the year 1986 we only use the observations that cover the February to December period obtained from the 1986 sample. We label these observations "1986s" as they are the ones we use in our original sample to estimate $\hat{\boldsymbol{\Gamma}}$. In the case of the year 1987, we present the average occupational mobility rate obtained for the January to April period (labelled "1987s") separately from the average occupational mobility rate obtained from the remaining months (labelled "1987r"). The two vertical lines mark the time period in which dependent and independent interviewing overlap.

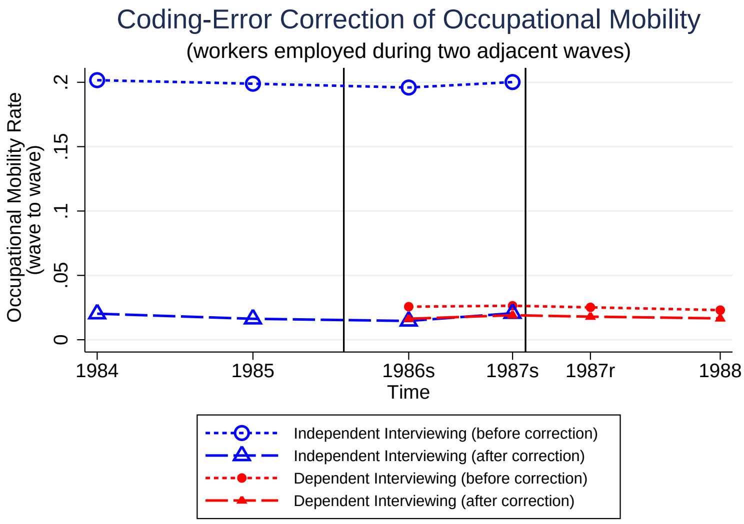


Figure 1: Correcting occupational mobility rates for workers employed in two subsequent waves

The short-dashed line with hollow circular markers depicts the observed occupational mobility rates obtained from pooling together employer/activity stayers and changers using the 1984/85 samples, where all respondents were subject to independent interviewing. This pooled sample yields occupational mobility rates that lie between 19.6%-20% and average 19.7% between the vertical lines. Under independent interviewing, $\mathbf{M_m}$ is garbled both at the source and destination occupations and hence is observed as $\boldsymbol{\Gamma}'\mathbf{M_m}\boldsymbol{\Gamma}$. Pre- and post-multiplying the latter by the respective inverses $(\boldsymbol{\Gamma}')^{-1}, \boldsymbol{\Gamma}^{-1}$ recovers $(\boldsymbol{\Gamma}')^{-1}\boldsymbol{\Gamma}'\mathbf{M_m}\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{-1} = \mathbf{M_m}$. Applying this procedure to the 1984/85 samples yields the long-dashed line with hollow triangular markers. The result is a drop in the occupational mobility rate to about 1.6%.

Next consider the 1986/88 samples. Here independent interviewing is only applied to employer/activity changers. Therefore, the observed overall occupational mobility rate for these samples is based on $\mathbf{M_s^{D,86/87}} + \mathbf{M_m^{I,86/87}}$. The matrix $\mathbf{M_s^{D,86/87}}$ does not contribute to any occupational transitions while matrix $\mathbf{M_m^{I,86/87}}$ does. This difference implies a much lower observed aggregate mobility rate. The latter is depicted by the short-dashed line with filled circular markers in Figure 1 and averages 2.6%. Correcting the occupational flows of the 1986-88 samples using $\hat{\mathbf{M}}_\mathbf{s}^{\mathbf{D},86/87} + (\hat{\boldsymbol{\Gamma}}')^{-1}\hat{\mathbf{M}}_\mathbf{m}^{\mathbf{I},86/87}\hat{\boldsymbol{\Gamma}}^{-1}$ yields the series depicted by the long-dashed line with solid triangular markers. The result is a drop in the occupational mobility rate to about 1.7%, which is very close to the $\boldsymbol{\Gamma}$-corrected occupational mobility rate from the 1984/85 samples. Indeed, in Figure 1 the blue (independent-interviewing) and red (dependent-interviewing) long-dashed lines of the $\boldsymbol{\Gamma}$-corrected measures nearly coincide between the two vertical lines.[8]

The fact that we obtain very similar corrected mobility rates after using the same underlying $\mathbf{\Gamma}$-correction matrix in two different survey designs, suggests that our methodology captures the extent of coding error quite well. Our methodology also seems to work well in other dimensions. We find that the occupational mobility rates in the years before 1986 are adjusted downwards to numbers that are very similar to the numbers obtain during the 1986s-1987s window. For example, the $\mathbf{\Gamma}$-corrected occupational mobility in 1985 is 1.63%, while during the 1986s-1987s window we obtain 1.62%. For the years after the 1986s-1987s window, we find that the $\mathbf{\Gamma}$-corrected occupational mobility rate is within 0.02% of the one obtained during this window. Further, Figure 1 shows that any changes in the level of the $\mathbf{\Gamma}$-corrected occupational mobility rate series appear to track changes in the uncorrected series. This also suggests that our correction method does not seem to introduce additional randomness into the occupational mobility process.

We apply our $\mathbf{\Gamma}$-correction method to those who changed employers with an intervening spell of unemployment or non-employment. We then compute the host of statistics shown in the main text and Supplementary Appendix B. We show that occupational miscoding increases observed gross mobility. In the raw data we compute an occupational mobility rate at re-employment of 53.1% based on the 2000 SOC. After applying our $\mathbf{\Gamma}$-correction, we obtain an occupational mobility rate at re-employment of 44.4%. We also find that coding errors makes occupational mobility appear less responsive to unemployment duration. Since in short unemployment spells true occupational staying is more common, miscoding creates relatively more spurious mobility and therefore our method corrects more the short spells, leading to a steeper relationship between occupational mobility and unemployment duration (see Table 1 and Figure 1 in Supplementary Appendix B). Miscoding also reduces the degree of procyclicality of gross mobility. This arises as coding errors will generate more spurious mobility in times were there are more true stayers (see Table 6 and Figure 11 in Supplementary Appendix B).

In contrast, we find that miscoding reduces the contribution of net occupational mobility (see also Kambourov and Manovskii, 2013). The average net mobility rate (as defined in the main text) increases from 3.6% (uncorrected) to 4.2% (corrected), a nearly 15% increase. To understand why this arises, consider the true net mobility transition flow matrix, $\mathbf{M_{net}}$, and note that this matrix does not have mass on its diagonal. Under independent interviewing coding errors imply that the true net mobility matrix would be observed as $\mathbf{M^I_{net}} = \mathbf{\Gamma}'\mathbf{M_{net}}\mathbf{\Gamma}$, which could have mass on the diagonal and hence biasing downwards net mobility flows. That is, coding errors mistakenly convert some true mobility flows into occupational stays, while miscoding for stayers is completely symmetric with respect to origin and destination occupations, and therefore should not give rise to spurious net mobility.

2.3 The differential impact of coding error on employer/activity stayers and movers

Note that our $\mathbf{\Gamma}$-corrected occupational mobility rate for those workers who changed employers through a spell of unemployment is 16.4% lower than the raw one (44.4% vs 53.1%, 2000 SOC). This adjustment is much smaller in relative terms than the one suggested by Kambourov and Manovskii (2008). They argue that on average around 50% of *all* year-to-year observed occupational mobility in the raw PSID data is due to coding error. Indeed when constructing the year-to-year occupational mobility rate for our pooled sample of employer/activity stayers and movers in the SIPP, we find that only 39.3% of observed occupational changes are genuine (an uncorrected rate of 26.7% vs a $\mathbf{\Gamma}$-corrected rate of 10.5%).

[8] Note that in the estimation of $\mathbf{\Gamma}$ we used $\hat{\mathbf{M}}^{\mathbf{I},86/87}_{\mathbf{m}}$, but we did not impose the additional restriction on $(\hat{\mathbf{\Gamma}}')^{-1}\hat{\mathbf{M}}^{\mathbf{I},86/87}_{\mathbf{m}}\hat{\mathbf{\Gamma}}^{-1}$ to equal $(\hat{\mathbf{\Gamma}}')^{-1}\hat{\mathbf{M}}^{\mathbf{I},85}\hat{\mathbf{\Gamma}}^{-1}$.

These findings are not mutually inconsistent. The key to this difference is that the *relative* importance of coding error varies greatly with the true propensity of an occupational change among employer stayers and among employer changers. True occupational changes are more likely to be accompanied by changes in employers (for examples of this argument see Hill, 1994, Moscarini and Thomsson, 2007, and Kambourov and Manovskii, 2009) and, vice versa, employer changes are more likely to be accompanied by occupational changes. As such, the *relative* adjustment that Kambourov and Manovskii (2008) find does not automatically carry over to different subsets of workers or flows measured at a different frequency.

As an example, consider an individual who is observed moving from "managers" to "laborers". Suppose "managers" was this individual's true source occupation, but the observed destination occupation was the result of coding error. If this individual is a true occupational stayer, the observed transition will wrongly tell us that the individual *stopped* being a manager and will generate a false occupational move. Instead, if this individual is a true occupational mover, the observed transition, although wrongly coded, will still capture the fact that the individual stopped being a manager and hence will capture a true occupational move. Given that true occupational changes are more likely to occur along side employer changes, there will be more workers among employer changers (relative to employer stayers) whose categorization as an observed occupational mover will not change after using the $\mathbf{\Gamma}$-correction. This implies that the measured occupational mobility rate of employer changers would have a relative smaller adjustment than the measured occupational mobility rate of employer stayers. Kambourov and Manovskii (2008) pooled together employer changers and stayers. Since the latter group represent the vast majority of workers in their sample (as well as in our sample), the relative adjustment proposed by these authors is naturally much larger.

Consider the following iterative back-of-the-envelope approximation to understand why there must be a larger proportion of true occupational movers among those who changed employers than among those who did not change employers. Recall that the $\mathbf{\Gamma}$-correction method implies that true occupational stayers will be coded as movers in about 20% of the times. This happens irrespectively of whether the worker changed employers or not. If we were to suppose that all of the unemployed who regain employment were true occupational stayers, the difference between their observed mobility rate (53%) and coding error (20%) would immediately imply that among the unemployed there must be true occupational movers and these true movers would represent at least 33% of the unemployed. This result then shrinks the population of occupational stayers (the "population at risk") among the unemployed to at most 67%, which (proceeding iteratively) implies the maximum extent of spurious flows produced under the same miscoding propensity is reduced to $0.2 \times 67\%$. In turn, the latter implies an updated lower bound on the percentage of true occupational movers among the unemployed of 39.6%. Proceeding iteratively, one arrives to a lower bound on the percentage of true occupational movers among the unemployed of 41.25%, which is close to the gross mobility we obtained after applying the $\mathbf{\Gamma}$-correction. A similar procedure but applied to employer stayers shows a much lower 'true' mobility rate among this group.

Table 6 shows that the differential impact of coding error on employer changers and stayers is present when considering several alternative occupational classifications as well as mobility across industries. In all these cases we use wave-to-wave mobility rates. The first column, $\mathbb{P}(\tilde{M}|S)$, presents the probability that a true employer/activity stayer in the 1985 sample is assigned the wrong occupational code and hence is observed as an occupational mover $\tilde{M}$. This probability is 17.8% when using the 2000 SOC. This implies that under independent interviewing we will observe an occupational mobility rate of 17.8% among the employer/activity stayers. This rate increases slightly when using the 1990 SOC, 19.7%, and remains high even when we aggregate oc-

Table 6: Inferred Coding Error Probabilities and Observed vs. Underlying Occupational Mobility

Classification	$\mathbb{P}(\tilde{M}\|S)$	$\mathbb{P}(\tilde{M}\|U)$	$\mathbb{P}(M\|U)$	$\frac{\mathbb{P}(M\|U)}{\mathbb{P}(\tilde{M}\|U)}-1$	$\mathbb{P}(\tilde{o}\neq o\|U)$
2000 SOC (22 cat)	0.178	0.531	0.444	-0.164	0.095
1990 SOC (13 cat)	0.197	0.507	0.401	-0.209	0.105
1990 SOC (6 cat)	0.148	0.402	0.317	-0.213	0.077
NR/R Cognitive, NR/R Manual (4 cat)	0.110	0.332	0.263	-0.208	0.058
Cognitve, R Manual / NR Manual (3 cat)	0.083	0.273	0.218	-0.199	0.043
Major industry groups (15 cat)	0.101	0.523	0.477	-0.088	0.055

Sample: unemployed between 1983-2013, in 1984-2008 SIPP panels, subject to conditions explained in data construction appendix (most importantly: unimputed occupations (resp. industries), with restrictions to avoid right and left censoring issues.) $\mathbb{P}(\tilde{M}|S)$: probability that the wrong code is assigned to a true stayer; $\mathbb{P}(\tilde{o}\neq o|U)$: probability that the wrong code is assigned to an unemployed worker; $\mathbb{P}(\tilde{M})$: observed occupational mobility among the unemployed; $\mathbb{P}(M)$: inferred underlying true mobility (proportion of unemployed). *NR/R* refers to routine vs. non-routine. Further details on the classifications are explained in the data construction appendix.

cupations into six categories, 14.8%.[9] Aggregating occupations into four tasked based categories (routine vs. non-routine and manual vs. cognitive) only brings down the observed mobility rate of true occupational stayers to 11%. As many other studies, we also find the probability that a true stayer is observed as a mover is lower when considering industries instead of occupations.

The second and third columns show the observed, $\mathbb{P}(\tilde{M}|U)$, and $\mathbf{\Gamma}$-corrected, $\mathbb{P}(M|U)$, occupational mobility rate of those workers who changed employers through unemployment. These are obtained using the probability that an unemployed worker in the 1984-2008 SIPP panels is assigned the wrong occupational code at re-employment, $\mathbb{P}(\tilde{o}\neq o|U)$. We observe that across all classifications the (relative) difference between the observed and the $\mathbf{\Gamma}$-corrected occupational mobility rates of the unemployed, $\mathbb{P}(\tilde{M}|U)-\mathbb{P}(M|U)$, is about half the size of $\mathbb{P}(\tilde{M}|S)$. The fourth column shows this difference in relative terms. It is clear that across all classifications the same coding error generates a much larger difference (in absolute and relative terms) between the observed and $\mathbf{\Gamma}$-corrected occupational mobility rates of employer stayers than among employer changers.

3. Measuring coding error in the PSID

We now broaden the above analysis and use probabilistic models based on the PSID (as in Kambourov and Manovskii, 2008) to assess the impact of coding errors on the probability of an occupational change. The advantage of the $\mathbf{\Gamma}$-correction method is that it captures all sources of error in assigning occupation codes. It delivers an identification procedure that is not subject to the issue highlighted in Corollary A.1 and that recovers the extent to which coding errors arise at the level of each occupation. Further, the SIPP provides a large sample size in which we can apply our correction method. The advantage of the PSID is that retrospective coding affected directly the way occupational transitions among employer changers were measured. We exploit this feature and assess the impact of coding error on employer movers and employer stayers separately and compare the results with the ones obtained from the SIPP using our $\mathbf{\Gamma}$-correction method. We find a very consistent picture across the two data sets.

To assess the impact of retrospective coding in reducing coding errors, we use the PSID retrospective occupation-industry supplementary data files, which contain the re-coding the PSID staff performed on the occupational mobility records obtained during the 1968-1980 period. Since the 1981-1997 records were not re-coded and collected under independent interviewing, the earlier period can be used to construct "clean" occupational mobility rates and to analyse the effect of measurement error at the coding stage. In constructing

[9] The six groups are: (1) managers/professional speciality; (2) tech support/admin support/sales; (3) services; (4) farm/forest/fisheries; (5) precision production/craft/repair; and (6) operators, fabricators and laborers. These correspond to the summary occupational group of the 1990 SOC.

our sample we closely follow Kambourov and Manovskii (2008, 2009). The details of this sample are described in the Supplementary Appendix B.7.

3.1 Gross occupational mobility rates

As in Kambourov and Manovskii (2008) we define the *overall* occupational mobility rate as the fraction of employed workers whose occupational code differs between years t and $t+1$ divided by the number of workers who were employed in year t. As these authors we also consider those workers who were employed at the time of the interview in year $t-1$, unemployed in t and employed at the time of the interview in year $t+1$. Further, we define the *within-employer* occupational mobility rate as the fraction of workers employed who did not change employers but exhibit a different occupational code between years t and $t+1$, divided by the number of employed workers who did not change employers between years t and $t+1$. Similarly, we define the *across-employer* occupational mobility rate be the fraction of employed workers who's occupational code differs between years t and $t+1$ and reported an employer change between these years, divided by the number of employed workers in year t who have reported an employer change between years t and $t+1$. To identify employer changes we follow the procedure detailed in Kambourov and Manovskii (2009), Appendix A1.

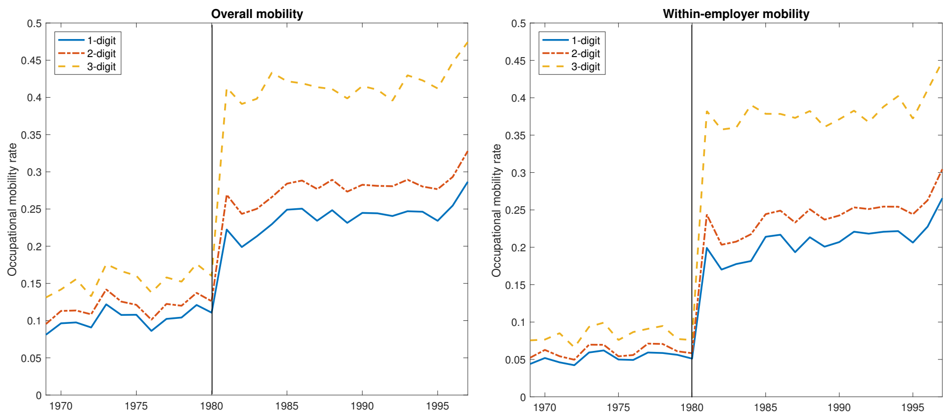


Figure 2: Overall and within-employer occupational mobility rates

The left panel of Figure 2 depicts the yearly overall occupational mobility rate at a one-, two- and three-digit level of aggregation, using the 1970 SOC. When retrospective re-coding was used, the overall occupational mobility rate experienced a large downward shift, ranging between 10 to 25 percentage points, depending on the level of aggregation of the occupational codes. These drops suggest that only between 38% to 45% of all occupational moves are genuine. This is very similar to the conclusion reached by Kambourov and Manovskii (2008).

The right panel of Figure 2 shows that the *within-employer* occupational mobility rates experienced even stronger drops than the overall ones under retrospective coding. In contrast, the left panel of Figure 3 shows that the impact of coding error in the *across-employer* occupational mobility rates is much more moderate and hardly visible when aggregating occupations at a one-digit level. These results then suggest that the impact of coding error on the overall mobility rates mainly arises from those workers who did not change employers, where employer stayers account on average for 87.1% of all employed workers in a given year, while those who changed employers account for the remainder 12.9%. We find a similar conclusion based on the SIPP data.

Next consider the effects of coding error on the occupational mobility rate of only those workers who

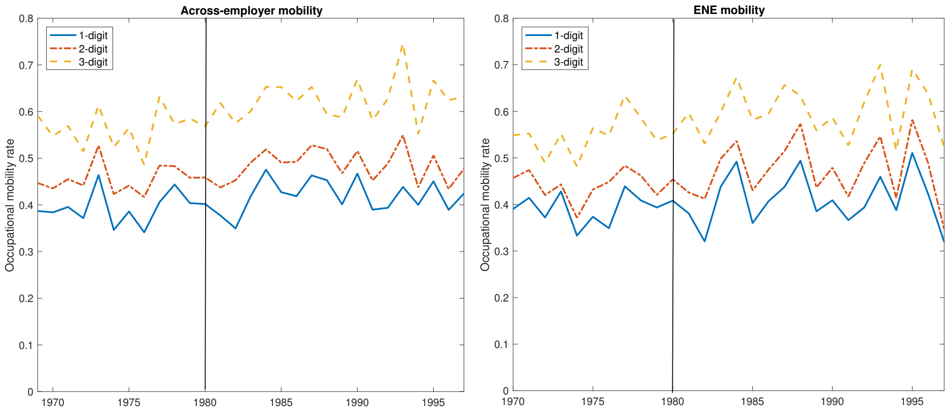


Figure 3: Across-employer occupational mobility rates

changed employers through a spell of non-employment (ENE). To construct the ENE occupational mobility rate we consider (i) those workers who were employed at the interview date in year $t-1$, non-employed at the interview date in year t and once again employed at the interview date in year $t+1$; and (ii) those workers employed at the interview dates in years t and $t+1$, but who declared that they experienced an *involuntary* employer change between these two interviews. An involuntary change is defined as those cases were the worker declared a job separation due to "business or plant closing", due to "being laid off or were fired" or their "temporary job ended" (see Supplementary Appendix B.7 for details). We divide these flows by the number of workers who changed employers through a spell of non-employment during the corresponding years. The right panel of Figure 3 shows that coder error once again seems to have a small effect on the occupational mobility rates of these workers.

3.2 Probabilistic models

The visual impressions given by the above figures on the effects of coding error are confirmed when estimating the effects of retrospective coding using a probit or a linear probability model. In these regressions the dependent variable takes the value of one if the worker changed occupation and zero otherwise. We include the indicator variable "break" which takes the value of one during the years in which the PSID used retrospective coding. In addition we control for age, education, full or part-time work, occupation of origin, region of residence, aggregate and regional unemployment rates, a quadratic time trend and number of children.[10]

Table 7 shows the marginal effects for the probit regressions.[11] It shows that retrospective coding had a large and significant effect on reducing the probability of changing occupations when all workers were included in the sample. Furthermore, the values of the marginal effects of the "break" indicator are very close to the amount by which the overall occupational mobility series shifted when retrospective coding was used, as depicted in Figure 2.

Our estimates also show that the effect of retrospective coding is much more moderate when we condition

[10] As in Kambourov and Manovskii (2008), the education indicator variable takes the value of one when the worker has more than 12 years of education and zero otherwise. This is to avoid small sample problems if we were to divide educational attainment in more categories. The regional unemployment rates are computed using US states unemployment rates.

[11] These estimates are obtained using the personal weights provided by each survey, but similar results are obtained when using the unweighted data. We also obtained very similar results when using the linear probability model on weighted and unweighted data and when using robust standard errors and clustering standard errors at a yearly level.

Table 7: The effect of measurement error on the PSID (probit marginal effects)

	All workers			Across employer			ENE		
	1-digit	2-digits	3-digits	1-digit	2-digits	3-digits	1-digit	2-digits	3-digits
Unemp rate	-0.002	-0.004	0.001	-0.026**	-0.025**	-0.021**	-0.058***	-0.057***	-0.032*
Reg unemp rate	-0.003	-0.003	-0.003	0.015**	0.014*	0.013*	0.046***	0.039***	0.022*
Age	-0.008***	-0.008***	-0.011***	-0.007	-0.008	-0.007	-0.025*	-0.029**	-0.022*
Age squared	0.6 e-4***	0.6 e-4**	0.9 e-4***	0.4 e-4	0.3 e-4	0.2 e-4	0.26 e-4	0.29 e-4*	0.19 e-3
Education	0.015***	0.016***	0.007	0.026*	0.026*	0.030*	0.024	0.034	0.069**
Break	-0.133***	-0.165***	-0.260***	-0.040	-0.069**	-0.065**	-0.029	-0.077	-0.065
Full-time	0.034***	0.017	-0.001	-0.019	-0.055	-0.101***	0.051	0.043	-0.065
N Obs	39,047	39,047	38,841	4,962	4,935	4,656	1,792	1,782	1,576
R^2	0.010	0.119	0.184	0.040	0.067	0.134	0.064	0.081	0.175

Levels of significance: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

the sample on workers who changed employers and when we consider the ENE sample.[12]

3.3 A comparison of coding errors across the PSID and SIPP

The point estimates obtained in Table 7 suggest that the probability of an occupational change for those who changed employers and for those who changed employer through non-employment spells, should be lowered on average by 3 percentage points at a one-digit level, 8 percentage points at a two-digit level and 7 percentage points at a three-digit level to capture the effect of coding error. To compare these estimates to the ones obtained from the SIPP, first note that retrospective coding is done using the same descriptions of the "kind of work" individuals gave in past interviews and hence captures coding errors introduced at the coding stage (see Sullivan, 2010). In addition to this coder error, our $\mathbf{\Gamma}$-correction method also takes into account that the source of code disagreement can originate in different descriptions of the same work (respondent error). Hence we would expect a higher correction when using the $\mathbf{\Gamma}$-correction than when using retrospective coding. Taking this feature into account and noting that coder error is expected to be the most important source error (see Mathiowetz, 1992), the PSID estimates compare very well with the adjustments implied by the $\mathbf{\Gamma}$-correction method.

In particular, when aggregating occupations into major categories (2000 SOC or 1990 SOC) and using the $\mathbf{\Gamma}$-correction, the corrected average occupational mobility rate for the non-employed was approximately 11 percentage points lower than the one obtained using the raw SIPP data. Noting that the occupational aggregations used in the SIPP and the two-digit aggregation used in the PSID lead to very similar ENE occupational mobility rates, the 11 percentage point adjustment obtained from the SIPP is thus close to the 8 percentage points suggested by Table 7.[13] The difference between the adjustments obtained from the SIPP and the PSID (3 percentage points) then provides a rough estimate of the impact of the respondent error. This estimate then implies that the importance of coder error is about 2.6 times larger than the importance of the respondent error. This is remarkably consistent with Mathiowetz (1992), who shows that the importance of coder error is two times larger than the importance of the respondent error when aggregating codes at a one-digit level and five times larger than the importance of the respondent error when aggregating codes at a three-digit level.

Furthermore, both the SIPP and the PSID data sets strongly suggest that the percentage reduction of occu-

[12]We do not include the within-employer occupational mobility in Table 7 because, as suggested by the graphical analysis, the results are very similar to the ones obtained with the full sample.

[13]For the 1985-1995 period, during which the PSID and SIPP overlap, the average "year-to-year" occupational mobility rate of the non-employed in the PSID and the SIPP was both around 53.1% and 47.6% when using the major occupational categories of the 1990 SOC.

pational mobility due to coding error varies substantially between employer stayers and employer movers. In the case of employer stayers, a large percentage of transitions are implied to be spurious. In the PSID we find that at a two-digit level 45% of yearly transitions are spurious, while in the SIPP we find that 40% of the yearly transitions of employer stayers are spurious. In the case of employer movers, occupational mobility is reduced by about 10 percentage points, but high occupational mobility remains (around 40% comparing before and after an employer changers), after applying retrospective coding or after using the $\mathbf{\Gamma}$-correction. As discussed earlier, this difference arises as among employer stayers the proportion of true occupational stayers is high and coding errors translate into a large amount of spurious mobility. Among those who changed employers through non-employment there is a much smaller proportion of true occupational stayers and hence the "population at risk" to be assigned a spurious occupational change is smaller.[14]

4. Discussion of assumptions $A1$ and $A2$

We now turn to discuss the two assumptions that appear the most restrictive in our analysis. As mentioned earlier, assumption $A3$ is verified in our data.

4.1 Assumption $A1$

This assumption requires that the realization of an occupational code does not depend on workers' labor market histories, demographic characteristics or the time it occurred in our sample. Therefore it implies that errors in the individuals' verbatim responses are fully captured by the nature of their job and hence only depend on their *true* occupation. It also implies that $\mathbf{\Gamma}$ is time-invariant. Since we use these implications extensively in the implementation of our correction method, we now investigate them further to help us evaluate the strength of $A1$. Our main conclusion is that our $\mathbf{\Gamma}$ correction method appears to pick up heterogeneity in miscoding across occupations that is robust to estimations on subsamples of the population. Likewise, we find evidence that even though the $\mathbf{\Gamma}$ matrix was estimated using 1985-1986 data, it captures well miscoding observed in recent years.

4.1.1 Worker heterogeneity

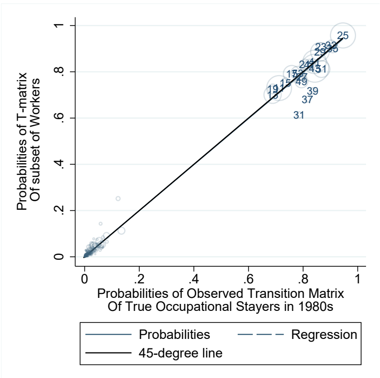
We investigate two aspects of worker heterogeneity that could deliver very different estimates of $\mathbf{\Gamma}$. We first consider whether the accuracy of the answer to the occupation question is affected by differences in workers' education attainment. A concern would be that more educated workers can better explain the type of job they are performing and hence coding errors would be less severe among these workers than in those with lower education. We then consider whether the accuracy of the occupation information is affected by a worker's interview status. In the SIPP either the worker reports his/her occupation him/herself to the interviewer (self-report) or another person of the same household reports his/her occupation (proxy). One would be concern that proxies answers are more prone to coding error.

Our analysis relies on re-estimating $\mathbf{\Gamma}$ using the 1985-1986 sample but on subsamples based on different educational attainment categories and interview status. As discussed above, the estimation process recovers a transition matrix of purely spurious mobility, $\hat{\mathbf{T}}^{\mathbf{I}}_{\mathbf{s}}$, (as shown in Table 3 for the 1990 SOC). We can then compare the estimated matrix of each subsample with the one computed in our main analysis. This comparison allows

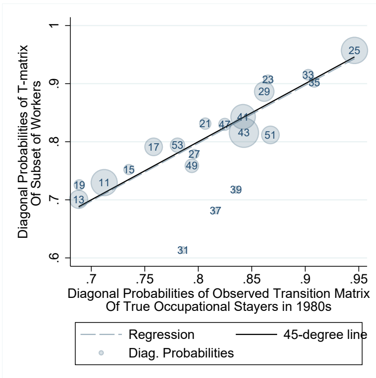
[14] vom Lehm et al. (2021) analyses the impact of coding errors on occupational mobility in the CPS when pooling together movers and stayers and find similar results to the ones we derive in the SIPP for the same pooled sample.

us to gauge whether the aforementioned characteristics lead to different miscoding even when conditioning on occupation.

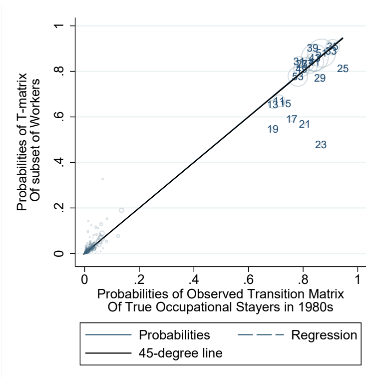
Education differences To study the effect of education differences, we divide the 1985-1986 sample into two groups: (i) high skilled workers which captures all those individuals with at least some college and (ii) low skilled workers which capture all those individuals with at most a high school degree. About one-third of the sample covers high skilled group. Note that occupations and education are highly correlated in the data. For example, very few non-college-educated workers can be found in architecture and engineering occupations, while very few college-educated workers can be found in production and some service occupations. Hence, part of the impact of education would already be captured by conditioning on occupations. However, the aim of this exercise is to evaluate the impact of the within-occupation variation in education on the level of miscoding.



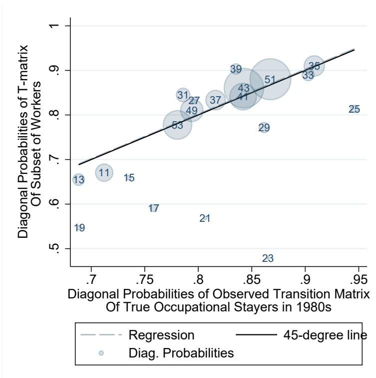
(a) High skilled - All cells



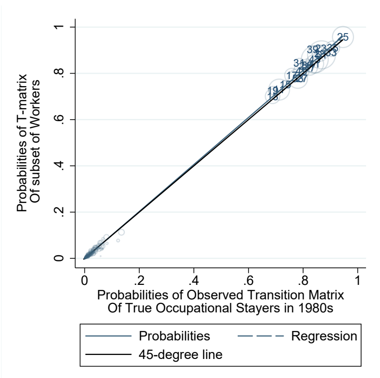
(b) High skilled - Diagonal



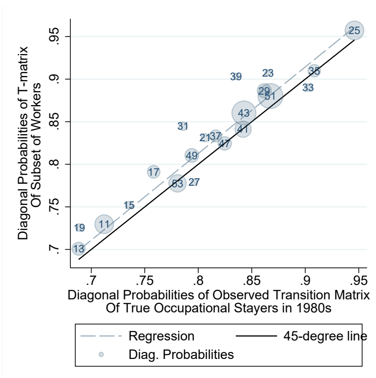
(c) Low skilled - All cells



(d) Low skilled - Diagonal



(e) Dominant education - All cells



(f) Dominant education - Diagonal

Figure 4: Comparing the estimated spurious transition matrices by education

Each panels (a)-(d) of Figure 4 depicts a scatter plot in which on the y-axis are the elements of the transition matrix $\hat{\mathbf{T}}^{\mathbf{I}}_{\mathbf{s}}$ estimated for either the high skilled subsample (panels (a)-(b)) or the low skilled subsample (panels (c)-(d)) and on the x-axis are the elements of $\hat{\mathbf{T}}^{\mathbf{I}}_{\mathbf{s}}$ estimated in our main analysis. We use the major occupations categories of the 2000 SOC and depict the same numbering of occupations as in Table 5 for the diagonal elements. Panels (a) and (c) show all the elements of the corresponding matrices, while panels (b) and (d) focus only on the diagonal elements. It is in the latter where the $\mathbf{\Gamma}$ matrix captures the heterogeneity in the probability of being coded correctly. Although the correction method formally involves applying the inverse of $\mathbf{\Gamma}$, there is an intuitive relation between the diagonal elements and the probability of being miscoded. The higher a diagonal element, the lower miscoding tends to be. Further note that as most of the mass in these transitions matrices lies on the diagonal elements they will be observed on the top right of the graph, while the off-diagonal elements will be closer to the origin. For the former we also show circles around them representing the relative size of the occupations.

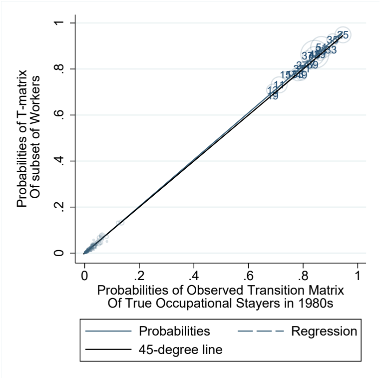
If educational differences across workers created very different coding errors, we would observe large deviations from the 45-degree line. The latter would suggest important differences between the coding errors obtained in our main analysis and the ones obtained when taking into account differences in workers' education. Instead, panels (a)-(d) shows that this is not the case. The correlation of coding errors is very close to one. Using OLS to fit a regression line among the observations, we observe that the regression line lies nearly on top of the 45-degree line.

Panels (e)-(f) of Figure 4 present a different approach to evaluate the effect of education differences on coding errors. In this case we only consider, for each occupation, the observations of those workers with the most dominant education (more than 50%) in that occupation. That is, we counterfactually impose the coding error attributed to the dominant education group on all workers in such an occupation. As before, if educational differences have meaningful effects on coding errors, controlling for occupations, then we would observe the regression line diverge significantly from the 45-degree line. Instead, we once again observe that these lie very close to each other, with only a slight deviation due to the diagonal elements (see also Table 8 below).

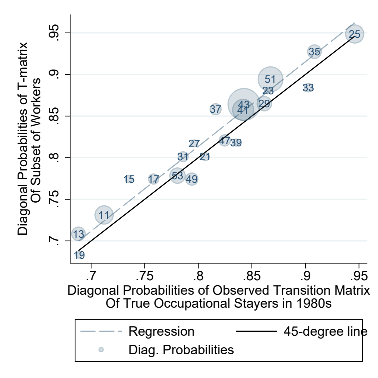
Interview status differences To investigate the impact of differences in the interview status of a worker on miscoding, we divide the 1985-1986 sample into those who were interviewed in person (self-interviewed) in two consecutive waves and those who had their information given by a proxy (proxy interview) at least in one of the waves. This divides the sample roughly in half: 55% were interviewed in person and 45% involve a proxy interview.

Figure 5 presents the same scatter plot exercises as describe above, but this time using the interview status instead of education attainment. We can observe a very high correlation between the elements of the transition matrix $\hat{\mathbf{T}}^{\mathbf{I}}_{\mathbf{s}}$ for each of these subsamples and of the transition matrix $\hat{\mathbf{T}}^{\mathbf{I}}_{\mathbf{s}}$ estimated in our main analysis. In this case the regression line is also nearly on top of the 45-degree line, with a very slight deviation at the diagonal elements. This deviation implies that for those interviewed in person the slope is a little steeper than one, while it is a little lower than one for those interviewed by proxy. This captures that the self-interviewed appear slightly more accurate, with slightly less spurious mobility, than the proxy interviewed. Nevertheless, in both cases we obtained a very similar conclusion. Occupations like managers, business and financial operators, computer and mathematical occupations and physical scientists are more prone to miscoding; while occupations like education, training and library occupations, food services and protective services are less prone to miscoding.

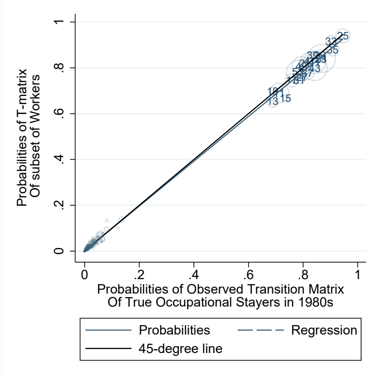
One could further subdivide the analysis using the interaction between education and interview status categories to gain a further insight. However, the above analysis suggests that we would find once again that



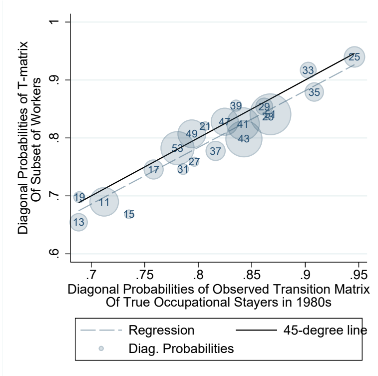
(a) Self-interview - All cells



(b) Self-interview - Diagonal



(c) Proxy-interviewed - All cells



(d) Proxy-interview - Diagonal

Figure 5: Comparing the estimated spurious transition matrices by interview status

subdividing the sample into these categories would not meaningfully change our estimated $\hat{\mathbf{T}}^{\mathbf{I}}_{\mathbf{s}}$ and hence $\mathbf{\Gamma}$ matrix.

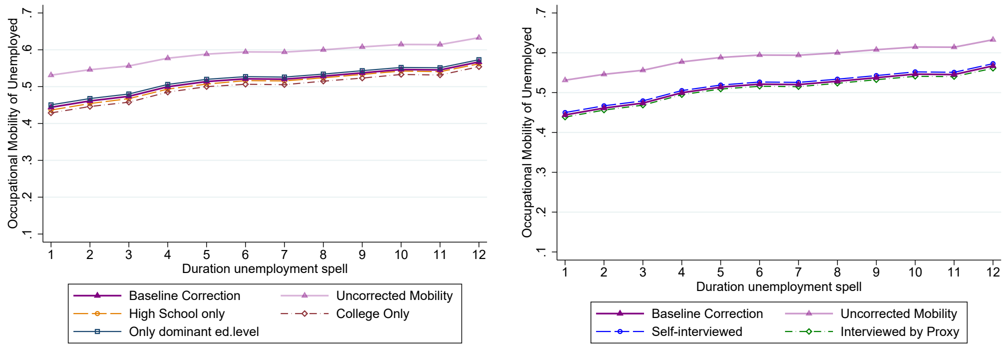
Implications for measured occupational mobility Table 8 summarises the above results, showing that the overall level of occupational staying among true stayers estimated across all subsamples is very similar to the one estimated in our main analysis. The correlations when using all elements of the matrices are nearly one. As mentioned above these correlations drop but still remain very high when only considering the diagonal elements. That is, the $\mathbf{\Gamma}$ matrix capture miscoding differences across occupation that is present (i) whether the individual is the one being interviewed or a proxy provides his/her information and (ii) across education categories.[15]

Next we apply the implied $\mathbf{\Gamma}$ obtained for each of the above subgroup of workers to correct the mobility-duration profile documented in Section 2.2 of the main text. Figure 6a considers the mobility-duration profiles when using the $\mathbf{\Gamma}$ matrices obtained from high skilled and low skilled workers. Figure 6b considers the same mobility-duration profile but this time using the $\mathbf{\Gamma}$ matrices obtained from the self and proxy interviewed. In both graphs we depict the uncorrected mobility duration profile (without any smoothing), and the one corrected with our baseline $\mathbf{\Gamma}$.

[15] Interestingly, high skilled workers are a bit more likely to be miscoded than low skilled workers. This could arise as the occupation typically performed by the high skilled are more specialized than those performed by the low skilled, making miscoding more likely in the former. Indeed the increased miscoding reflects mostly the occupations that typically performed by the high skilled (11 to 29 in the 2000 SOC). These are associated with a 78.1% of occupational stayers in the estimated $\hat{\mathbf{T}}^{\mathbf{I}}_{\mathbf{s}}$, while the proportion of occupational staying in the occupations typically performed by low skilled workers (31 to 53 in the 2000 SOC) is 83.7%. If we were only to consider the high skilled in all occupations we find 79.9% of occupational staying, while if we only consider the low skilled in all occupations we obtain 82.8% of occupational staying.

Table 8: Observed Occupation Staying of True Stayers, Correlations across $\hat{\mathbf{T}}^{\mathbf{I}}_{\mathbf{s}}$ Estimates.

	level	corr. w/ baseline			level	corr. w/ baseline	
	occ stay	all	diag		occ. stay	all	diag
Baseline	0.822	1.000	1.000		**Education**		
	By interview status			Dominant educ.	0.832	1.000	0.982
Self-interviewed	0.838	1.000	0.979	High skilled	0.799	0.998	0.900
Proxy interview	0.799	0.999	0.953	Low skilled	0.828	0.999	0.867



(a) Educational attainment

(b) Interview status

Figure 6: Corrected Mobility-Duration Profile from $\mathbf{\Gamma}$ estimated on subsamples

We can observe that when using the $\mathbf{\Gamma}$ matrices from the education attainment subsamples the implied mobility-duration profiles are very similar to the obtained in our main analysis. When we use high skilled sample to measure miscoding in all occupations, including those that have very few college workers in them, the associated profile is only by 1-1.5 percentage point different from the baseline one. We also obtain a very similar conclusion when using the low skilled sample across all occupations. The mobility-duration profile corrected by using only information on those with the dominant education in a given occupation also track the original mobility duration profile closely. Moreover, when using the $\mathbf{\Gamma}$ matrices implied by the self-interviewed or proxy subsamples, we once again obtain mobility-duration profiles that hardly differ from the one obtained in our main analysis. This evidence thus suggests that differences in education attainment or interview status (or their interaction) do not affect the conclusion obtained from our main correction analysis: occupational mobility is high, a little over 40%, and increases moderately with unemployment duration.

4.1.2 Time invariance of $\mathbf{\Gamma}$

As previously discussed we estimate the $\mathbf{\Gamma}$ correction matrix using the 1985-1986 SIPP panels and then apply it to occupational mobility data up to 2014. An important concern that arises from this application is whether the estimated coding errors remain relevant in the later years of our sample as implied by assumption $A1$. To directly evaluate whether this is the case we would need to re-apply our correction method at a later date and compare the more recent coding error correction matrix to our baseline one. To perform this exercise we will require a US data set that switched from dependent to independent interviewing with respect to occupations in the 2000s. However, we are not aware of any data set in which this re-design took place. Due to this barrier we instead take a different approach. We consider a group of workers who are likely occupation stayers, but coded independently such that some of them would be observed as occupational movers. We then evaluate the extent to which our $\mathbf{\Gamma}$ matrix can predict these workers' observed occupational mobility, particularly in more recent years. A high predictive power would suggest that coding errors estimated using data from 1985-1986 remain relevant throughout our sample.

Motivated by Fujita and Moscarini (2017) we use temporary laid-off workers to approximate this group of occupational stayers. These authors' empirical work suggests that workers in temporary layoff have a very low chance of an occupation switch once recalled by their previous employers. Therefore it is not unreasonable to assume that this set of workers are *largely* made up of true occupation stayers and can provide a good approximation to the latter group. Instead of using the SIPP to measure temporary layoffs (as Fujita and Moscarini, 2017), however, we use the Current Population Survey (CPS). The main reason for this choice is that dependent interviewing in the CPS only applies when a person is employed both in the current month and the month before (see e.g. the CPS interviewing manual 2015). This implies that workers who are *unemployed on temporary layoff* will have their occupations coded independently. In contrast, as interviews in the SIPP are conducted every four months, one cannot guarantee that temporary layoffs with spells of unemployment of at most 4 months will have their occupations independently coded at re-employment.

	(Cell-by-Cell) Correlation across Transition Matrices			
	All, <13 weeks		All 3-dgt Industry Stayers	
CPS years	All cells	Diagonal	All cells	Diagonal
1994-2021	0.991	0.844	0.993	0.805
1994-2004	0.986	0.816	0.990	0.823
2004-2014	0.989	0.790	0.992	0.742
2014-2021	0.989	0.812	0.990	0.710

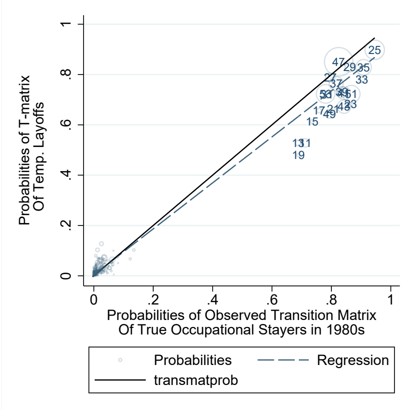
Table 9: Correlations Observed Transition Matrix of Temporary Layoffs with Spurious Transitions (from $\mathbf{\Gamma}$)

Therefore, if the vast majority of temporary layoffs are independently-coded true occupation stayers *and* the $\mathbf{\Gamma}$ coding errors persist over time, we would observe a positive correlation between the elements of the transition matrix $\hat{\mathbf{T}}_s^{\mathbf{I}}$ estimated in our main analysis and the elements of the observed transition matrix of those workers returning to work out of a temporary layoff, even if we consider temporary layoffs three decades later. Table 9 presents the results of such an exercise using several time periods post the CPS 1994-redesign. The first two columns refer to all those workers who were in temporary layoff for less than 13 weeks before re-employment; while the second two columns refers to the subset of temporary layoffs with less than 13 weeks in unemployment what were also observed as industry stayers at re-employment when considering a 3-digit industry aggregation. Although the latter group reduces the sample size, it is more likely to contain true occupational stayers. This occurs as occupation and industry mobility tends to go hand in hand. We return to this point below.[16]

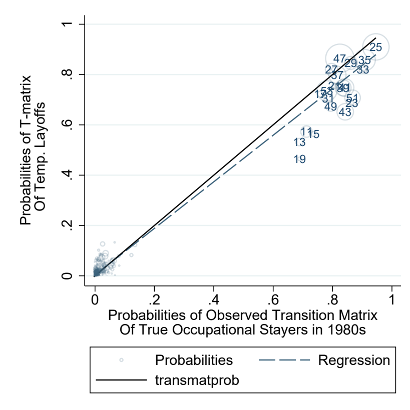
Across both samples of temporary layoffs Table 9 shows the correlation between their observed occupational transition matrix and $\hat{\mathbf{T}}_s^{\mathbf{I}}$, obtained from the 1985-1986 SIPP data, when using the 2000 SOC. One can immediately see the very high correlations in the observed occupational mobility patterns among those in temporary layoff and the one implied by the $\mathbf{\Gamma}$ matrix. Moreover, the value of the correlation is nearly unchanged over time even when focusing on the years 2014-2021, the period after our SIPP analysis ended. In particular, the correlations are nearly one when taking all elements of the matrices (or, very similar but not shown here, all cells with positive probabilities). Note, however, that some degree of positive correlation may not be unexpected for this exercise, as the diagonal of both matrices will unsurprisingly consists of numbers closer to one, while some other cells will naturally be closer or equal to zero. Therefore, a more stringent test is to

[16]The post-1994 sample size is about 45,000 for all temporary layoffs with duration less than 13 weeks and about 30,000 for temporary layoffs who are industry stayers. The decadal samples (1994-2004, 2004-2014, 2014-2021) are about 1/3 of these numbers.

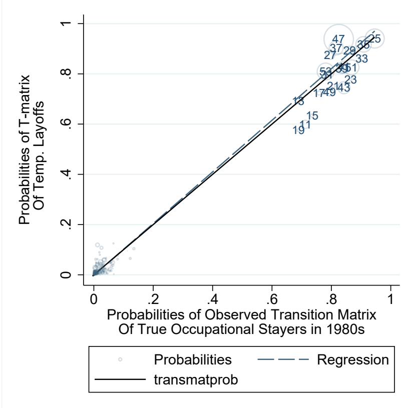
consider the correlation only of the diagonal elements. Here we once again observe high correlations ($\rho > 0.7$ across all periods). Indeed coding errors according to the $\mathbf{\Gamma}$ matrix explain about two-thirds of the variance of heterogeneity on the diagonal of occupational transition of temporary layoffs in the second half of the CPS sample ($\rho = 0.80$).



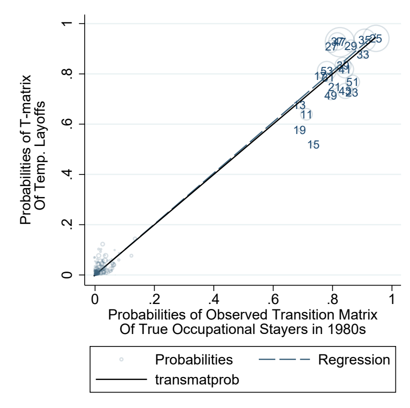
(a) Temp Layoffs, Occ. Mobility, 1994+



(b) Temp Layoffs, Occ. Mobility, 2014-21



(c) Temp. Layoffs, Ind. Stayers, 1994+



(d) Temp. Layoffs, Ind. Stayers, 2014-21

Figure 7: Comparing temporary layoffs in the CPS with the spurious transition matrix from the SIPP

Figure 7 displays the same exercise as in the previous section, where we now depict the values of each element of the occupation transition matrix of temporary layoffs on the y-axis relative to the associate value of the transition matrix implied by $\mathbf{\Gamma}$. The top row considers all temporary layoffs (unemployed for less than 13 weeks), while the bottom row considers the subgroup of temporary layoffs that are also industry stayers. We observe that the same occupations most prone to miscoding according to $\mathbf{\Gamma}$ are also the ones observed among temporary layoffs: managers, business and financials operations, computer and math occupations, and physical sciences. Conversely, many of the occupations that are least likely to be miscoded according to $\mathbf{\Gamma}$ are also the ones among temporary layoffs: healthcare practitioners, protective services, food preparation services and above all educators. Figures 7.b and Figure 7.d consider temporary layoffs for the more recent period 2014-2021 and show that these patterns are largely maintained over time (if a bit more noisy).

Some caveats Even if coding mistakes were completely persistent, the comparison between the observed transition matrix of temporary layoffs and our correction matrix may be affected by other factors not considered in the previous section.

First, and perhaps most importantly, the comparison of the occupational transitions behind $\mathbf{\Gamma}$ with those of temporary layoffs relies on the assumption that the latter set of workers captures both the whole set of occupation stayers (and its miscoding), but simultaneously no other type of worker, i.e. true movers. Deviations from this requirement will likely lower the observed correlation. One way to investigate the presence of true movers among temporary layoffs is to consider their amount of net mobility – as miscoding inflates excess mobility rather than net mobility. When considering all temporary layoffs, the average net occupational mobility rate (as defined in Section 2.3 of the paper) is significantly lower for these workers than for all the unemployed, only about 1/3 of the latter. If we assume that true gross mobility and net mobility move roughly in proportion, this would suggest that a much larger share of temporary layoffs are true stayers. However, since net mobility does not drop to zero, some true mobility must remain. The latter seems a likely explanation of why, even though the top row of Figure 7 shows a high correlation between the two transition matrices, the slope of the regression line lies a bit further from the 45-degree line relative to the cases studied in Section 4.1.1. However, if we focus on the subgroup of temporary layoffs that are also industry stayers we find that their average net occupational mobility rate drops by another third relative to all temporary layoffs.[17] This strongly suggests that among this subgroup of temporary layoff workers the vast majority are indeed occupational stayers and hence provide a more accurate way to evaluate the persistence of the coding errors in $\mathbf{\Gamma}$. Consistent with this conclusion, we observe in the bottom row of Figure 7 that the regression line of temporary layoffs who are also three-digit industry stayers lies much closer to the 45-degree line.[18]

Second, there may be some differences between the CPS and SIPP that affect the occupational information and its coding, e.g. (slight) differences in the questions about occupations.[19] Given many similarities in terms of occupations across both surveys (and across time), we expect this factor not to be of great importance. Third, there may be finite sample considerations. The sample sizes of the 10-year CPS windows are lower than the sample on which the $\mathbf{\Gamma}$ is estimated in the SIPP (see footnote 15 and Section 1.3, respectively). We observe some indication that for the smaller CPS 10-year windows, the sample size may leave some role for finite sample noise. In particular, the overall correlation of the temporary layoff transition matrix over 1994-2020 is higher (at 0.843) than the correlation with respect to the underlying subsets 1994-2004, 2004-2014, 2014-2020 (around 0.80).

Having highlighted these factors, it is noteworthy that the coding error correction matrix estimated in the 1985-1986 SIPP panels can explain between half and two-thirds of the *variance* along the diagonal of the transition matrix of temporary layoffs, even in the last decade. These high correlations occur even though the set of workers in the SIPP behind $\mathbf{\Gamma}$ and the temporary layoffs in the CPS are in a very different labor market situation (and formally even in a different employment status). Overall, this suggests that when using the $\mathbf{\Gamma}$-implied miscoding correction we gain a much better sense of miscoding, beyond a simple uniform level adjustment applied evenly across all occupation (which by construction has zero correlation with heterogeneity on the diagonal of spurious occupational mobility transition matrix).

Implications for measured occupational mobility As a final exercise we use the observed occupational transition matrix of the subset of workers who were on temporary layoff and observed as industry stayers instead of the $\mathbf{\Gamma}$ matrix in order to correct the occupational mobility-duration profile of the unemployed in our

[17]Note that in theory, it could be the case that a spurious occupation change correlates with the *realization* of a spurious industry change, something not ruled by assumption $A1$.

[18]As noted before, to the extent that some realizations of spurious industry mobility correlate with realizations of spurious occupation mobility (which is not ruled out by assumption $A1$), we may have restricted our sample too much to fully satisfy the condition that a representative set of true occupational stayers are included. However, the difference between the regression line and the 45-degree line remains small.

[19]For example, in case the interviewee reports an occupation description that the CPS considers to be too general, the interviewer follows up by providing a list of more specific occupations from which the interviewee should chose from (Appendix 2 of the CPS Manual, 2015). Standard SIPP documentation does not report a similar routing.

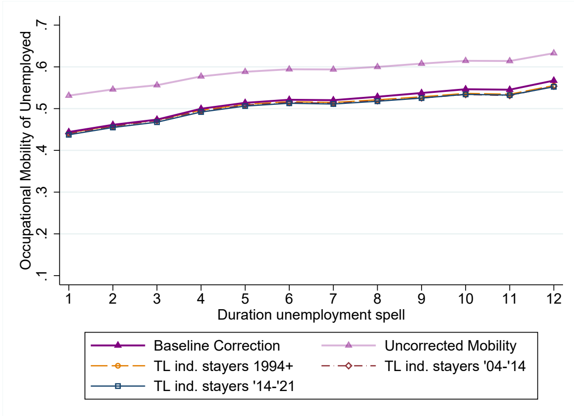


Figure 8: Corrected Mobility-Duration Profile from Temp Layoff and $\mathbf{\Gamma}$

SIPP sample. We consider this subset of workers as they have the largest proportion of true occupational stayers and provide a more accurate comparison with the mobility-duration profile derived from the $\mathbf{\Gamma}$ matrix. Figure 8 shows that the two corrected profiles are nearly indistinguishable, particularly during the first six month of the unemployment spell. This occurs irrespectively of whether we consider all industry-staying temporary layoffs as the base for the correction matrix, or focus on the subperiods 2004-2014, or even 2014-2021. In all these cases we find a high occupational mobility rate that increases moderately with unemployment duration as we did in our main analysis.

Taken together, the above evidence strongly suggests that the $\mathbf{\Gamma}$ coding errors do remain relevant even in the 2010s.

4.2 Assumption $A2$

This assumption requires that the number of workers whose true occupation i gets mistakenly coded as j and the number of workers whose true occupation j gets mistakenly coded as i and is needed to derive equation (1) and solve for $\mathbf{\Gamma}$. Although this assumption is clearly strong, it is important to note it is a weaker version of the one proposed by Keane and Wolpin (2011) and subsequently used in the structural estimation of discrete choice models. In particular, Roys and Taber (2017) used this assumption to correct for occupation classification error. They require that the number of workers whose true occupation i gets mistakenly coded as occupation j is independent of i and j for all $i \neq j$ and given by a constant. This implies that in their "garbling" matrix all off-diagonal elements are the same. In contrast, our approach allows (and recovers) different off-diagonal elements in $\mathbf{\Gamma}$, showing the large heterogeneity in bilateral occupation miscoding (see also Sullivan, 2009).

Unfortunately, we cannot directly test assumption $A2$ in our SIPP data. Instead, we can investigate whether there is evidence of an important implication of $A2$: coding errors do not change the distribution of occupations across workers. The PSID retrospective coding exercise is suitable for this exercise as it provides cleaned occupations codes for workers during the period 1968-1980, but does not correct for coding errors in the occupations assigned to the same workers during the period 1981-1997. Assumption $A2$ would imply that we should not observe any significant change in the contribution of each individual occupation across these two periods after controlling for time trends.

Table 10 reports the results of regressing the contribution of a given occupation across the period 1968-1997 on the "break" indicator and cubic polynomial time trend.[20] We can observe that classification errors across

[20] Using higher order polynomials generate similar results. We also exclude the categories referring to armed forces and agricultural

Table 10: Occupational mobility and unemployment duration

	Health prof, School/college teachers	Account, archit, legal tech., others	Managers, officials, proprietors	Clerical workers	Retail and sales workers
Break	-0.0002	0.0040	-0.0014	0.0034	0.0044
	(0.0027)	(0.0039)	(0.0043)	(0.0033)	(0.0030)
Constant	0.0458***	0.0774***	0.1459***	0.0966***	0.0510***
	(0.0040)	(0.0100)	(0.0064)	(0.0134)	(0.0121)
Time trend	X	X	X	X	X
R^2	0.8482	0.9619	0.8552	0.8294	0.8793
	Foremen workers	**Craftsmen kindred workers**	**Operatives kindred workers**	**Laborer workers**	**Other service workers**
Break	-0.0004	0.0012	0.0031	-0.0085***	-0.0031
	(0.0023)	(0.0029)	(0.0044)	(0.0017)	(0.0023)
Constant	0.0179***	0.0494***	0.1717***	0.0524***	0.0907***
	(0.0034)	(0.0043)	(0.0066)	(0.0015)	(0.0027)
Time trend	X	X	X	X	X
R^2	0.6108	0.7911	0.9110	0.8437	0.4639

Levels of significance: $^{*}p < 0.1$, $^{**}p < 0.05$, $^{***}p < 0.01$

occupational coding do not seem to meaningfully change the distribution of occupation across our PSID sample. We observe that the break indicator is nearly zero and not statistically significant, except for the unskilled laborer occupation. However, even in this case the effect of coding error is to decrease the contribution by an average of only 0.85%. Considering that this is a small occupation, contributing 4.20% in 1980 and 3.60% in 1981, the effect of coding error in this respect does not seem to be of first order importance.

References

[1] Abowd, J. and A. Zellner. 1985. "Estimating Gross Labor-Force Flows". *Journal of Business & Economic Statistics*, 3 (3): 254-283.

[2] Campanelli, P., K. Thomson, N. Moon, and T. Staples. 1997. "The Quality of Occupational Coding in the United Kingdom". In L. Lyberg, P. Biemer, M. Collins, E. de Leeuw, C. Dippo, N. Schwarz, & D. Trewin (Eds.), *Survey Measurement and Process Quality*, 437-453. New York: Wiley.

[3] Hausman, J., J. Abrevaya, and F.M. Scott-Morton. 1998. "Misclassification of the Dependent Variable in a Discrete-Response Setting", *Journal of Econometrics*, 87: 239-269.

[4] Hill, D. 1994. "The Relative Empirical Validity of Dependent and Independent Data Collection in a Panel Survey". *Journal of Official Statistics*, 10 (4): 359-380.

[5] Jäckle, A. and S. Eckman. 2020. "Is that still the same? Has that Changed? On the Accuracy of Measuring Change with Dependent Interviewing". *Journal of Survey Statistics and Methodology*, 8(4): 706-725.

[6] Jäckle, A. 2008. "Dependent Interviewing: Effects on Respondent Burden and Efficiency of Data Collection". *Journal of Official Statistics*, 24 (3): 1-21.

[7] Fujita, S. and G. Moscarini. 2017. "Recall and Unemployment". *American Economic Review*, Vol. 102(7): 3875-3916.

occupations as these are not included in our main analysis.

[8] Kambourov, G. and I. Manovskii. 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.

[9] Kambourov, G. and I. Manovskii. 2009. "Occupational Specificity of Human Capital". *International Economic Review*, 50 (1): 63-115.

[10] Kambourov, G. and I. Manovskii. 2008. "Rising Occupational and Industry Mobilty in the United States: 1968-97". *International Economic Review*, 49 (1): 41-79.

[11] Keane, M. and K. Wolpin, 2011. "The Effect of Parental Transfers and Borrowing Constraints on Educational Attainment". *International Economic Review*, 42 (4): 1051-1103.

[12] vom Lehn, C., C. Ellsworth and Z. Kroff, 2021. "Reconciling Occupational Mobility in the Current Population Survey", *Journal of Labor Economics*, *forthcoming*.

[13] Lynn, P. and E. Sala. 2007. "Measuring Change in Employment Characteristics: The Effect of Dependent Interviewing". *International Journal of Public Opinion Research*, 18 (4): 500-509.

[14] Mathiowetz, N. 1992. "Errors in Reports of Occupations". *Public Opinion Quarterly*, 56: 332-135.

[15] Magnac, T. and M. Visser. 1999. "Transition Models with Measurement Errors", *The Review of Economics and Statistics*, 81 (3): 466-474.

[16] Moscarini, G and K. Thomsson. 2007. "Occupational and Job Mobility in the US". *Scandinavian Journal of Economics*, 109 (4): 807-836.

[17] Poterba, J. and L. Summers. 1986. "Reporting Errors and Labor Market Dynamics". *Econometrica*, 54 (6): 1319-1338.

[18] Roys, N. and C. Taber. 2017. "Skill Prices, Occupations and Changes in the Wage Structure". Mimeo. Department of Economics, Royal Holloway, University of London.

[19] Sullivan, P. 2010. "Empirical evidence on occupation and industry specific human capital". *Labour Economics*, 17 (3): 567-580.

[20] Sullivan, P. 2009. "Estimation of an Occupational Choice Model when Occupations are Misclassified". *Journal of Human Resources*, 44 (2): 495-535.