Confused about Careers? Untangling Occupational Mobility, Miscoding and Distance [*†]

Carlos Carrillo-Tudela
University of Essex
CEPR, CESifo, IZA, and Stone Centre

Saman Darougheh
Danmarks Nationalbank

Ludo Visschers
U. Carlos III de Madrid,
CESifo, IZA, Stone Centre,
Edinburgh Futures Institute

March 2025

Abstract

Occupational mobility and its relation with economic fundamentals is obscured by mistakes when assigning occupational codes. We correct ('de-garble') occupation patterns, using the heterogeneous probabilities with which a worker in a certain occupation appears as working in another. This leads to stronger empirical patterns of occupational mobility with 'task distance' based on O*NET, across age, the business cycle and with wage changes. Miscoding between occupation pairs reflects task similarities and can be used as a distance measure itself, and appears helpful to distinguish among occupations close in task space. Overall, taking into account miscoding, occupations and tasks tend to matter (even) more for economic outcomes than standard approaches suggest.

*Ludo Visschers gratefully acknowledges grants CEX2021-001181-M and ATR2023-145734 funded by MICIU/AEI /10.13039/501100011033 (Unidades de Excelencia Maria de Maeztu, and Programa ATRAE, of the Spanish Government). The views in this article are solely those of the authors and should not be interpreted as reflecting the views of Danmarks Nationalbank, or any person associated with the European System of Central Banks.

†We provide the data necessary to correct occupational flows under https://www.samandarougheh.com/miscoding.

1 Introduction

Occupational mobility shapes the working life of individuals and the performance of economies in the aggregate. However, it is hard to measure accurately: in standard surveys occupations are often 'miscoded', i.e. incorrectly reported. Miscoding drastically inflates observed occupational mobility and distorts the origins or destinations of true occupational moves, changing the career paths observed in the data. As a result, there may be quite a difference between the patterns suggested by the survey data and the underlying role of occupations in economic outcomes.

In this paper, we correct occupational mobility for miscoding at the level of origin-destination occupation pairs and find that this can considerably strengthen the inferred link between occupational mobility and economic fundamentals and outcomes. Methodologically, we explain how to correct observed occupational mobility patterns – using the heterogeneous probabilities with which a worker in a certain occupation is observed as working in (specific) other occupations. We are able to identify these probabilities, even though we cannot observe the true occupations directly, using survey design changes of the Current Population Survey (CPS), at the level of the three-digit Census Occupational Classification.[1] Applying this method, we are able to revisit a number of prominent empirical findings in the literature.

First, by building our correction method on the underlying 'garbling' process with which the worker's true occupation at times appears as a different one, we are able to apply our correction to very different sets of workers. We find that the same miscoding probabilities imply very different, but typically considerable, amounts of spurious flows and therefore very different relative corrections of mobility rates. The underlying miscoding probabilities are high and rather heterogeneous across occupation pairs. As a result, in the basic CPS, occupational mobility of employer changers (at the three-digit level) falls from over 60% to around 50% after correcting for miscoding, while year-to-year occupation mobility falls from 45% to around 20%.

Second, we investigate the relation between occupational mobility and task distance in the face of miscoding. Observed task distance is distorted additionally because miscoding affects the start and end points of true occupational moves, on top of the issue of spurious

[1] At higher levels of aggregation, we combine the data from the Current Population Survey with the (smaller) data from a similar survey redesign of the Survey of Program Participation.

flows. Even though we have just highlighted that miscoding is common, the literature on task distance tends not to take it into account. Implicitly, this could be justified when miscoding involves replacing the true occupation by another occupation that is still very close in terms of their O*NET task portfolio.[2] We instead find that occupation miscoding can substantially distort distance measurement. Spurious flows, for example, are present even among transitions over large task distances. Moreover, a substantial portion of transitions are corrected into distances in different quantiles of the distance distribution when not spurious. After correcting for miscoding, we observe stronger relations between occupation task distance with fundamentals, such as age or the business cycle, and with economic outcomes, such as wages.

Third, we observe that occupational miscoding itself is also rooted in similarity of tasks and work activities across occupations: more similar occupations are more likely to be miscoded into each other. It therefore follows that miscoding itself, too, can be used for measurement of 'task distance'. We define a miscoding-based distance 'measure' and show that it does well — better than our O*NET tasked-based measure at short and medium distances — in predicting which occupational transitions take place. It also predicts as well the wage losses of the unemployed, given their occupational change, with a suggestion of improvement at the short distance range. To be clear, these findings are observed *after* we have corrected our data for miscoding.[3] The conclusion is that task similarity across occupations can be more relevant for economic outcomes than the raw (uncorrected) data *and* O*NET-based task distance measures suggest. Moreover, it is suggestive that there may be room to improve on O*NET based task similarity, particularly with respect to distinguishing 'roughly similar' occupations from 'very similar' occupations in terms of tasks, especially in terms of predicting mobility.

Let us now put our paper in broader context. It has been known for considerable time that occupation coding is prone to considerable error (see e.g. Mellow and Sider (1983) and Mathiowetz (1992)). The effects of this depend on the survey design. When longitudinal surveys use independent interviewing and coding, i.e. respondents are asked to answer without reference to information gathered in previous interviews and when their answers

[2]Indeed, sometimes this is used as an explicit justification, however, without evidence.

[3]Hence they are not driven by the distortion of miscoding that would bias the *uncorrected* flows in the same direction of the miscoding-based distance. Imprecision in our estimates of miscoding probability will create an attenuating effect in the relation between miscoding-corrected mobility and miscoding-based distance. See the discussion in section 4.

are edited or codified by only taking into account information from the current interview.[4] This inflates mobility: an occupation miscode before and after a correctly coded interview of a true occupation stayer creates two spurious occupation moves. Given high error rates, this will be a common and serious issue (discussed e.g. in the literature mentioned below).

In part in response to this, household surveys, like the CPS and the SIPP (Survey of Income and Program Participation), have switched to the use of dependent interviewing, where interviewees are asked if answers in the last interview still apply. As we formally show below, the changes observed with these survey redesigns provide us with the means to infer underlying miscoding probabilities. At the same time, the redesigns leave an important part of the issue unresolved: those who report activity or employer changes (direct or via unemployment) in the CPS or SIPP are subsequently still independently interviewed (and coded) with respect to their occupation.[5] Since an important part of economic reallocation occurs via these routes, it is important to take miscoding into account and correct it especially for these workers – and indeed a large part of our paper will look at these subset of workers.

A large body of economic research emphasizes that occupational mobility shapes working lives, employer mobility, wage inequality and aggregate outcomes. Prominent papers in this literature, such as Kambourov and Manovskii (2008; 2013), Moscarini and Thomsson (2007) and Neal (1999), have taken seriously the issue of spurious mobility caused by miscoding. A typical strategy to address this is to mark certain observed transitions as spurious by the presence or absence of certain correlates. For example, occupation changes without employer changes, industry changes or large enough wage change are deemed spurious.[6]

There are two types of drawbacks to this strategy. First, this strategy is not useful for analyses where occupation identities matter beyond identifying an occupation move, such as the measurement of occupation similarity in terms of tasks (task distance). Second, some changes considered spurious by these criteria will be genuine, while spurious flows will still occur among the population not ruled out by these criteria. We will derive in section 2.4

[4] In the case of occupations, this means each time workers get interviewed, they are asked to describe anew their line of work. Professional coders will then independently code workers' lines of work by only using these descriptions to assign occupation codes

[5] So do occupation codes between interview 4 and 5 of the CPS.

[6] A related strategy considers any occupational move immediately followed by a move in the opposite direction is spurious. However, dependent interviewing as used e.g. in the CPS, makes it this criterion inapplicable: after a coding mistake, the occupation code is carried over across interviews until the worker changes activity or employer, or the rotation ends.

that a portion of employer (and therefore industry) stayers does change occupations, while spurious occupation changes have considerable presence among employer movers. These type I and II errors may be substantial and do not necessarily cancel out.

A sophisticated way of dealing with the drawbacks of this strategy is to validate, for the sets of workers considered, the size of the adjustment for spurious flows with historical data or survey redesigns. Kambourov and Manovskii (2008), Kambourov and Manovskii (2009) and Moscarini and Thomsson (2007) do this using the retrospective recoding of occupations in the 1970s PSID or the 1994 CPS redesign. Vom Lehn, Ellsworth, and Kroff (2022) combine the retrospective occupation information from CPS supplements next to the year-to-year observed occupational mobility in the basic monthly CPS and are consequently able to look at miscoding at multiple points of time.[7] A drawback is that the application of the corrections derived in these papers or rules-of-thumb based on these remain likely specific to samples similar to those selected in these papers. Although our approach is different, we share with these papers that we leverage differences in survey designs to learn about miscoding.

Rather than focusing on treating symptoms of miscoding, such as spurious flows, we want to go closer to the root of the problem and uncover the statistical process by which miscoding arises, to counteract it – towards "miscode noise-cancellation". In this vein, we build on Abowd and Zellner (1985), and Poterba and Summers (1986) in particular who use a garbling matrix that captures the errors made in classifying workers across (three) labor force statuses.[8] Their critical assumption is that CPS reconciliation interviews, done in the 1980s, provide the true individuals' labor market status.[9] We do not have similar auxiliary information on occupations that allows us to directly recover the much larger garbling matrix of one-, two- and three-digit occupation codes. In the section 2.1 of this paper, we cover how instead it is possible to distill this from the before and after of aforementioned

[7] In a sense, their comparison of CPS supplement with basic monthly CPS is similar to using a survey redesign, and one that is repeated at multiple points in time, allowing them to study miscoding over time.

[8] We use the term "garbling" because the relation between the observed occupation identities relates to the underlying occupation identities through a Markov matrix that adds noise to the underlying occupations and satisfies the conditions laid out in Marschak and Miyasawa (1968) in the process of formalizing the term "garbled information" and goes back to Blackwell (1951).

[9] In recent work, Dvorkin (2025) assumes the PSID retrospective recode uncovers the true occupations, points out that in principle this could lead us directly observe the garbling matrix for his purposes if the PSID sample were large enough, but (given the relatively small size of the sample) opts for a strategy where instead he uses the retrospectively recoded PSID information is used to contribute a set of miscoding-free occupation/industry observations to the structural model, which helps identify structural parameters that include miscoding probabilities inside the structural model.

survey redesigns, without the necessity of assuming that true information is revealed in any particular interview.[10] [11]

In Carrillo-Tudela and Visschers (2023c) we have summarized and used the method of section 2 specifically to correct occupational mobility rates between major occupation groups and, even more aggregate, four task-based occupation supercategories, based on the one-digit correction matrix estimated from the 1986 SIPP redesign. By using the 1994 CPS redesign in this paper, we are able to estimate corrections at the three-digit occupational level. With this level of detail (and besides looking at the occupational mobility across different sets of workers), we will study task-based distances while offsetting miscoding (in section 3) and derive a miscoding-based distance measure (in section 4).

A large literature (following early and seminal papers as Poletaev and Robinson (2008) and Gathmann and Schönberg (2010), among others) takes the perspective of occupations as bundles of tasks and defines a distance metric between these bundles. It then finds interesting and important relations between 'task distance' to other economic outcomes (such as wage dynamics, propensities to change, etc.). One underlying idea is that human capital is less transferable between occupations when they share fewer tasks, i.e. are more distant. Following the early papers, a large amount of research, also on the macroeconomic side (e.g. Lise and Postel-Vinay (2020), Guvenen, Kuruscu, Tanaka, and Wiczer (2020) and Baley, Figueiredo, and Ulbricht (2022) has further emphasized the relevance of occupational distance in understanding career and wage mobility patterns.

The issue of the 'garbling' of origin and destination occupation in investigations of task distances, however, seems important yet under-investigated, with (to our knowledge) only a few papers addressing it more than cursorily. Abraham and Spletzer (2009) highlight the discrepancy between the CPS and the Occupational Employment Statistics regarding the

[10] More generally, a second strand of the literature interested in occupational misclassification estimates "garbling" matrices in structural models, where the observed choices of economic agents with respect to further economic variables like wages help identify miscoding (see e.g. Sullivan (2009) and Roys and Taber (2017).) However, the interplay between (true) occupational mobility and individual-level occupational wage changes is an active research area that has not yielded a clean a consensus on how true occupation changes and wage changes co-occur that then could be used to identify or validate the identification of which occupation changes are spurious. In addition, many of these models are demanding to solve, which limits the amount of miscoding heterogeneity can be estimated, to often a single parameter. The aforementioned Dvorkin (2025) fits in this "structural" category, but is able to make considerable progress capturing the heterogeneous miscoding probabilities in his structural model.

[11] Further related is Feng and Hu (2013) who also identify a 'garbling' matrix of 3-state labor market status, but for the situation that miscoding can bias the size of the stocks (e.g. the unemployment rate) with different assumptions, while in our paper we assume independent interviewing get the size of occupations right at the population level.

intensity of tasks and the level of wages, pointing to coding error as one explanation. Speer (2016) is the only paper we are aware of that explicitly sets out to measure the impact of occupational miscoding on the distribution of task distances. However, he considers all occupation changes of employer changers true, and all non-management occupation occupation changes of employer stayers spurious (following the literature discussed above). Our paper adds two advantages relative to his analysis: by using the underlying miscoding process we will respect that both spurious transitions among employer changers and true transitions among employer stayers are common. And second, and perhaps more crucial: beyond pointing out the importance of miscoding, our method allows us to address it and "clean it out". Section 3 investigates how mobility and wage patterns, as a function of task distances, change after this is done.

More than garbling information alone, section 4 shows that miscoding contains valuable information on task distance itself: miscoding occurs more frequently when task similarity is high. The performance of this distance measure suggests that tasks as unit of analysis may be even more important than analyses based on O*NET descriptions alone suggest. Section 5 concludes.

2 Measuring and Correcting Occupational Miscoding

In this section, we present how one can infer the probabilities that a worker in occupation i is reported instead in occupation j, for all occupation pairs i, j, even though we cannot observe worker's true occupation directly, following Carrillo-Tudela and Visschers (2023a).[12]

The changes of survey design in the Current Population Survey (CPS) and the Survey of Income and Program Participation (SIPP) allow us to estimate these. We then present how to correct observed occupation flows, and apply this correction to different sets of workers.

[12]This estimation method was originally developed to deal with the presence of spurious occupational flows for the unemployed and can be found in working paper Carrillo-Tudela and Visschers (2023b). However, as it was subsidiary to the main point of that paper, it was covered briefly on page 1123 in the main text of the published version, Carrillo-Tudela and Visschers (2023c), with the statement of key propositions relegated to the Online Supplemental Appendix and the proofs not part of the publication. Continuing investigations led us to see the importance and broader usefulness of the method (e.g. for task distance at the heart of this paper). Here we provide exposition, theorems and most important proofs of theory behind the miscoding correction in section 2.1, and the remainder of the proofs are included in the appendix of this paper, making the methodology self-contained.

2.1 Identification of Miscoding between Occupation Pairs

Let $\mathbf{M}$ denote the matrix that contains workers' *true* occupational flows, where element m_{ij} is the flow of workers from occupation i to occupation j. Let γ_{ij} represent the probability that a worker who is in reality in occupation i instead is reported to be in occupation j. This probability is an element of an $O \times O$ 'miscoding' or 'garbling' (transition) matrix $\mathbf{\Gamma}$, with O occupations and $\sum_{j=1}^{O} \gamma_{ij} = 1$.

Given $\mathbf{\Gamma}$ and assuming a large enough number of observations, under the *independent interviewing* survey design $\mathbf{M}$ will be observed as $\mathbf{M^I} = \mathbf{\Gamma}'\mathbf{M}\mathbf{\Gamma}$.[13] The pre- and post-multiplication by $\mathbf{\Gamma}$ takes into account that the observed occupations in consecutive interviews would be subject to coding error.

Alternatively, under *dependent interviewing*, respondents would answer questions that refer back to information gathered in previous interviews. In the survey designs we will consider in this paper, dependent interviewing is implemented as follows. Workers are asked if their employer or tasks have changed since the last interview. If they indicate that neither has changed, the occupation coded for the most recent interview will simply be carried forward, avoiding a potential spurious transitions due to miscoding. Differently, if workers report a change of employer or tasks, they will be asked to describe anew their line of work, which will then be independently coded into an occupation. Thus, *conditional on a reported task or employer change*, the main surveys under consideration revert back to independent interviewing and coding with respect to occupations.

Now suppose that we were to subject a population of true employer+activity stayers to independent interviewing as described above. Let $\mathbf{M_s^I}$ denote the matrix that contains these workers' *observed* occupational transition *flows*. In this case $\mathbf{M_s^I} = \mathbf{\Gamma}'\mathbf{M_s}\mathbf{\Gamma}$, where $\mathbf{M_s}$ is the diagonal matrix that captures the *true* distribution of employer+activity stayers over occupations on its diagonal. Thus, each of the off-diagonal elements of $\mathbf{M_s^I}$ will represent a spurious flow. Similarly, let $\mathbf{M_m}$ denote the matrix that contains the *true* occupational transition *flows* of employer+activity changers. The diagonal of $\mathbf{M_m}$ describes the distribution of true occupational stayers among employer/activity changers. The off-diagonal elements contain the flows of all true occupational movers. Under independent interviewing we observe $\mathbf{M_m^I} = \mathbf{\Gamma}'\mathbf{M_m}\mathbf{\Gamma}$.

[13] Here and in the exposition immediately below, we assume that the law of large numbers holds, i.e. no finite sample randomness. See Poterba and Summers (1986) for a similar formulation of misclassification across labor market states (employed, unemployed, out of the labor force).

We will show that if a large population were interviewed under both survey designs we can isolate a set of spurious occupation flows for true occupation stayers from $\mathbf{M}_{\mathbf{s}}^{\mathbf{I}}$, using the difference $\mathbf{M}^{\mathbf{I}} - \mathbf{M}_{\mathbf{m}}^{\mathbf{I}}$, where $\mathbf{M}^{\mathbf{I}}$ denotes the matrix that contains the aggregate occupational transition flows across two interview dates under independent interviewing.[14]In what follows we assume that under dependent interviewing employer stayers who truly change occupation answered affirmatively to changing their work activities. That is, a worker cannot truly switch occupations without changing their work activities, and will report this.[15] Further, for the matrix $\mathbf{M}_{\mathbf{s}}^{\mathbf{I}}$ to enable us to identify and estimate $\mathbf{\Gamma}$ we make three additional assumptions.

(*A1*) *Independent classification errors*: We assume that, conditional on the true occupation, the realization of an observed occupational code does not depend on workers' labor market histories, demographic characteristics or the time it occurred in our sample. This assumption is present in Poterba and Summers (1986) and Abowd and Zellner (1985) and is consistent with independent interviewing and coding.[16] This also implies that $\mathbf{\Gamma}$ errors in the individuals' verbatim responses are fully captured by the nature of the job these individuals are performing: they only depend on their *true* occupation.[17] The assumption also implies we can use the estimated $\mathbf{\Gamma}$ to correct observed occupational flow matrices at

[14]This is possible as all true movers would be present in both $\mathbf{M}^{\mathbf{I}}$ and $\mathbf{M}_{\mathbf{m}}^{\mathbf{I}}$, independently coded in both surveys, and thus differenced out. Importantly, under our assumptions, this approach remains valid even if $\mathbf{M}_{\mathbf{m}}^{\mathbf{I}}$ includes some true occupation stayers who are independently coded. Such situations arise when independent coding is triggered for reasons other than an occupation change (at the considered level of aggregation), such as when a worker changes employers while maintaining their occupation. Note that $\mathbf{M}_{\mathbf{m}}^{\mathbf{I}}$ might also contain reported activity changers who nevertheless stayed within the same occupation. In other words, we do not need to assume that $\mathbf{M}_{\mathbf{m}}^{\mathbf{I}}$ consists exclusively of true movers or that $\mathbf{M}_{\mathbf{s}}^{\mathbf{I}}$ captures all spurious moves in the data.

[15]We allow change of work activities that do not imply a change of occupation. Indeed, the presence, post CPS redesign, of independently interviewed employer or activity changers that do not truly change occupations creates a challenge for identification as detailed below.

[16]In the standard practice 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. 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.

[17]A seeming concern would be that errors introduced by respondents could be correlated with their characteristics. However in our context this is only a concern if this remains *after conditioning on their true occupation*. Two comments are in order. First, Mathiowetz (1992) shows that *coder error* is the main source of classification error when coding occupations. She shows that the importance of coder error is two to five times larger than the importance of the respondent error, depending on the level of occupational code aggregation. This limits the importance of the respondent error that may vary with respondent characteristic and perhaps shifts it more to which occupations are more easily to classify than others. Second, correlations at the population level between miscoding propensity (say, male, Hispanic ethnicitiy, non-white, skill level, as e.g. highlighted in Vom Lehn, Ellsworth, and Kroff (2022)) and worker characteristics not in contradiction with this assumption: "laborers" are very prone to miscode but also have a selected set of worker characteristics.

different times than the time window over which $\mathbf{\Gamma}$ was estimated. We think of this as an abstraction that helps us isolate the first-order impact of miscoding that are the focus of this paper.[18]

(*A2*) *"Detailed balance" in miscoding*: $\mathbf{diag}(\mathbf{c_s})\mathbf{\Gamma}$ is symmetric, where $\mathbf{c_s}$ is a Ox1 vector that describes the distribution of true employer+activity stayers across occupations and $\mathbf{diag}(\mathbf{c_s})$ is the diagonal matrix of $\mathbf{c_s}$. 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.[19] Detailed balance allows us to invert $\mathbf{\Gamma}$. Detailed balance is weaker assumption than e.g. the assumption made Keane and Wolpin (2001) and Roys and Taber (2017), who also incorporate 'unbiased misclassification' that does not change the size of occupations: there the probability of miscoding a worker in occupation i is assumed to be uniform across all occupations $j \neq i$, and *all* miscoding is captured by one parameter.

(*A3*) *Strict diagonal dominance*: $\mathbf{\Gamma}$ is strictly diagonally dominant in that $\gamma_{ii} > 0.5$ for all $i = 1, 2, ..., O$. This assumption is also present in Hausman et al., (1998) and implies a minimal level of competence for respondent and coder. That is, it is more likely to correctly code a given occupation i than to miscoded it.

Together these assumptions allow us to estimate miscoding probabilities specific to any pair of occupations. In particular, they allow for miscoding to be much more frequent between certain occupation *pairs* than others, rather than e.g. a function of only true, or only observed, occupation. The occupation pair-specific miscoding will be important in the sections below.

We can now relate the implied spurious flow matrix of employer+activity stayers, $\mathbf{M_s^I}$ to the garbling matrix $\mathbf{\Gamma}$. Note that $\mathbf{M_s} = \mathbf{diag}(\mathbf{c_s})$. Given Assumption *(A2)* we can observe $\mathbf{c_s}$ directly from the observed occupational distribution of dependently interviewed employer+activity stayers.[20] By virtue of the symmetry of $\mathbf{M_s}$ and Assumption (*A2*), we have

[18] Figure 1 below, which compares SIPP and CPS, illustrates relative stability of miscoding probabilities between mid 1980s to mid 1990s.

[19] We highlight that we only need to apply this assumption to the set of employer+activity stayers, $\mathbf{diag}(\mathbf{c_s})\mathbf{\Gamma} = \mathbf{\Gamma}'\mathbf{diag}(\mathbf{c_s})$, which as an assumption is perhaps more appealing when it is applied to the overall population. Nevertheless, we verify that in our data the set of employer+activity stayers is a large enough proportion of the overall population such that the difference between the occupational distribution of employer+activity stayers and the occupational distribution of the population is small.

[20] To show how we recover $\mathbf{c_s}$ from the data, let $\mathbf{c_s^D}$ denote the O-sized vector that describes the *observed*

that $\mathbf{\Gamma}'\mathbf{M_s} = \mathbf{\Gamma}'\mathbf{M_s}' = \mathbf{M_s}\mathbf{\Gamma}$. Since Assumption *(A1)* implies $\mathbf{M_s^I} = \mathbf{M^I} - \mathbf{M_m^I} = \mathbf{\Gamma}'\mathbf{M_s}\mathbf{\Gamma}$, substituting back yields $\mathbf{M_s^I} = \mathbf{M_s}\mathbf{\Gamma}\mathbf{\Gamma}$. Also note that $\mathbf{M_s^I} = \mathbf{M_s}\mathbf{T_s^I}$, where $\mathbf{T_s^I}$ is the occupational transition probability matrix of the employer+activity stayers in this population *observed* under independent interviewing. Substitution yields $\mathbf{M_s}\mathbf{T_s^I} = \mathbf{M_s}\mathbf{\Gamma}\mathbf{\Gamma}$. 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 $\mathbf{\Gamma}$,

$$\mathbf{T_s^I} = \mathbf{\Gamma}\mathbf{\Gamma}. \tag{1}$$

The 'double' miscoding of the occupations across two (independent) interviews only works out to be the squared application of the miscoding matrix $\mathbf{\Gamma}$, if the detailed balance assumption holds. Towards estimation of the $\mathbf{\Gamma}$ matrix, we would take the matrix root of $\mathbf{T_s^I}$. In general, we cannot guarantee the uniqueness of the root of a transition matrix (see Higham and Lin, 2011). However, we can show that, given Assumptions *(A2)* and *(A3)*, the root is indeed unique and can be constructed by diagonalization. The following proposition states and proves this claim.

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

Proof. We proceed in two steps: first we construct $\mathbf{\Gamma}$ as a solution to equation (1), given $\mathbf{T_s^I}$, with the proporties stated in the proposition. Then we show that this solution is unique under assumptions *(A2)* and *(A3)*.

STEP I, CONSTRUCTION: 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 Assumption *(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

occupation distribution of employer+activity stayers under dependent interviewing. 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. Then $\mathbf{c_s^D} = \mathbf{\Gamma}'\mathbf{M_s}\overrightarrow{\mathbf{1}} = \mathbf{M_s}\mathbf{\Gamma}\overrightarrow{\mathbf{1}} = \mathbf{M_s}\overrightarrow{\mathbf{1}} = \mathbf{c_s}$, where the second equality uses Assumption (A2) and the third that $\mathbf{\Gamma}$ is a transition matrix. We can likewise observe $\mathbf{c_s}$ from $\mathbf{M_s^I}$ by considering the vector that collects all flows that originated in each observed occupation, such that $\mathbf{c_s} = \mathbf{diag}(\mathbf{M_s^I}\overrightarrow{\mathbf{1}})$, where $\overrightarrow{\mathbf{1}}$ refers to the unit vector.

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 its 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, $\mathbf{T_s^I} = \mathbf{diag}(\sqrt{\mathbf{c_s}})^{-1}\,\mathbf{G}\,\mathbf{G}\,\mathbf{diag}(\sqrt{\mathbf{c_s}}) = \mathbf{diag}(\sqrt{\mathbf{c_s}})^{-1}\,\mathbf{Q}\,\mathbf{\Delta}^2\,\mathbf{Q}^{-1}\,\mathbf{diag}(\sqrt{\mathbf{c_s}})$, and hence $\mathbf{T_s^I}$ is also orthogonally diagonalizable, and has a root $\mathbf{P\Lambda}^{0.5}\mathbf{P}^{-1}$, where $\mathbf{\Lambda}$ is the diagonal matrix of eigenvalues of $\mathbf{T_s^I}$, and $\mathbf{P}$ the associated orthogonal matrix with eigenvectors of $\mathbf{T_s^I}$.

STEP II: UNIQUENESS Since Assumption *(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 and has all eigenvalues between 0 and 1 (as has $\mathbf{\Gamma}$). It follows that $\mathbf{G}\,\mathbf{G} = \mathbf{S}$ is also positive definite, and $\mathbf{diag}(\sqrt{\mathbf{c_s}})^{-1}\,\mathbf{S}\,\mathbf{diag}(\sqrt{\mathbf{c_s}}) = \mathbf{T_s^I}$ 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.

Given Assumptions *(A2)* and *(A3)* and the properties derived above, to now show the uniqueness of the root of $\mathbf{T_s^I}$, we 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 $\mathbf{K^{-1}H\Delta^2 H^{-1}K} = \mathbf{\Delta^2}$ 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 multi-

plicity of the eigenvalues δ_i itself. But then it must be true that $\mathbf{K^{-1}H\Delta H^{-1}K = \Delta}$. Then, $\mathbf{G = H\ \Delta\ H^{-1} = KK^{-1}H\ \Delta\ H^{-1}KK^{-1} = K\ \Delta\ K^{-1} = Y}$ which leads to a contradiction. ■

Hence, if we observe $\mathbf{T_s^I}$ associated with a garbling process $\mathbf{\Gamma}$, we can uncover $\mathbf{\Gamma}$. A key step towards consistent estimation of $\mathbf{\Gamma}$ is that under assumptions *(A2)-(A3)*, we can operate in the space of transition matrices that are similar (in matrix sense) to a symmetric positive definite matrix, $PDT(\mathbb{R}^{O\times O})$. In that space, the matrix root mapping is continuous, as established by Lemma 1 and proved in Appendix A.

Lemma 1 *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 in the spectral matrix norm.*

Let us define $\hat{\mathbf{M}}_\mathbf{s}^\mathbf{I}$ as the estimated flow matrix from the data. Construct its symmetrized version $\hat{\mathbf{M}}_{\mathbf{s,sym}}^\mathbf{I} = 0.5\hat{\mathbf{M}}_\mathbf{s}^\mathbf{I} + 0.5\big(\hat{\mathbf{M}}_\mathbf{s}^\mathbf{I}\big)^\mathbf{T}$ and call $\hat{\mathbf{T}}_\mathbf{s}^\mathbf{I}$ the transition matrix associated with the latter. Then we have

Proposition 2 $\mathbf{\Gamma}$ *is consistently estimated from* $(\hat{\mathbf{T}}_\mathbf{s}^\mathbf{I})^{0.5} \in PDT(\mathbb{R}^{O\times O})$*, the space of* $O \times O$ *transition matrices that are similar to symmetric positive definite matrices, such that* $\hat{\mathbf{\Gamma}} = (\hat{\mathbf{T}}_\mathbf{s}^\mathbf{I})^{\mathbf{0.5}} = \hat{\mathbf{P}}\hat{\mathbf{\Lambda}}^{\mathbf{0.5}}\hat{\mathbf{P}}^{\mathbf{-1}}$*. That is, as the number n of observations increases,* $plim_{n\to\infty}\hat{\mathbf{\Gamma}} = \mathbf{\Gamma}$.

It may be tempting to think that one can simply compare the levels of occupational mobility before and after the change of survey design from independent to dependent interviewing, to measure miscoding and correct miscoding. In the survey designs we cover in this paper, a dependent interviewing question may trigger a set of questions that follow independent interviewing. (Moscarini and Thomsson (2007) refer to this as 'conditionally independent interviewing') In particular, employer movers and activity changers are subject to independent interviewing (and coding). If there are true occupation stayers among these respondents, miscoding may still create spurious flows for these workers. Hence, the observed flows under dependent interviewing will still be overestimating true flows.[21] How important this is depends on how much independent interviewing of true occupation stayers still is triggered under the dependent interviewing design; hence, simply comparing occupational transition matrices (or, in the aggregate, rates) before and after the redesign is not sufficient.

To illustrate this identification problem, we consider a very simple example where suppose that every occupation has the same probability $1-\gamma$ of being coded correctly. Further,

[21] Further, if one is interested in the subset of employer and activity changers specifically, then the switch to the new dependent interviewing leaves the independent interviewing of this subset unaffected.

assume that the miscoding probability is uniform across very many occupations, so we can abstract from the probability that a true stayers is miscoded twice into the same occupation (and hence still observed as a stayer) or a mover is by chance miscoded as a stayer. Let a fraction s of the population be a true occupation stayer but a proportion x of these true stayers be an employer/activity changer and hence independently interviewed ($x = 1$ with independent interviewing for all). Assuming that occupational mobility under the independent interviewing design dropped from f_I to f_D under the above-specified dependent interviewing design, the following two equations hold $(1-s)+s2\gamma = f_I$ and $(1-s)+sx2\gamma = f_D$, where these equations capture that independently interviewed stayers have a 2γ probability of being observed as movers. It is clear that without knowledge of x (which we do not directly observe, in practice), one cannot identify s and γ.[22] However, we can identify these by using the reported proportion of worker that are interviewed independently in the dependent interview design, which is given by $\pi = (1-s)+sx$. Subtracting the second from the first equation above and using the expression for π yields $(1-\pi)2\gamma = f_I - f_D$ from which γ follows directly, and subsequently we can derive s, x.

In the general case, if $\mathbf{M_m}$ has unknown mass on its diagonal, $\mathbf{\Gamma}$ cannot be identified from $\mathbf{M^I}$ and $\mathbf{M^D}$ alone, where $\mathbf{M^D} = \mathbf{M_m^I} + \mathbf{M_s^D}$ 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. Building on the simple example above, for the general case, we can subtract $\mathbf{M}^I$ from $\mathbf{M}^D$. Then we observe a matrix $\mathbf{M^{diff}} = (\mathbf{M_s^D} - \mathbf{M_s^I})$, which has $0.5n(n-1)$ exogenous parameters given the symmetry of Assumption A2. (It needs to be symmetric and the i-th diagonal element equals the negative of the sum of all nondiagonal elements in row (or column) i.). Matrix $\mathbf{\Gamma}$ relates to $\mathbf{M^{diff}} = \mathbf{M_s} - \mathbf{\Gamma}'\mathbf{M_s}\mathbf{\Gamma}$, which has $0.5n(n+1)$ unknowns (again, given assumption A2).[23]

However, when a survey redesign allows us to identify all workers who are subjected to independent interviewing (with observed transition matrix $\mathbf{M_m^I}$) and this set of workers

[22] Though from the presence of occupation stayers among employer or activity changers, we can bound it from below: it seems save to conclude that it some occupation stayers are present among this group.

[23] In addition to $\mathbf{M^D} - \mathbf{M^I} = \mathbf{M_s} - \mathbf{\Gamma}'\mathbf{M_s}\mathbf{\Gamma}$, one observes $\mathbf{M^D}$ which equals $\mathbf{\Gamma}'\mathbf{M_m}\mathbf{\Gamma} + \mathbf{M_s}$, where $\mathbf{M_m}$ is mobility of the independently interviewed that contains all true movers, but not necessarily exclusively. When $\mathbf{M_m}$ has mass on its diagonal, this additional system of equations has n^2 additional exogenous variables on the LHS and n^2 unknowns (arising from $\mathbf{M_m}$, given $\mathbf{\Gamma}, \mathbf{M_s}$) on the RHS. If we knew the diagonal elements on $\mathbf{M_m}$ would equal zero, we would be able to identify $\mathbf{M_m}, \mathbf{M_s}, \mathbf{\Gamma}$, but empirically occupation staying among employer/activity movers is common.

contains all true movers (but not necessarily exclusively so), we can estimate $\mathbf{M}_{\mathbf{s}}^{\mathbf{I}}$ from $\hat{\mathbf{M}}^{\mathbf{I}} - \hat{\mathbf{M}}_{\mathbf{m}}^{\mathbf{I}}$. That is, the above discussion is summarized as

Corollary 1: *If* $\mathbf{M_m}$ *has mass on its diagonal,* $\mathbf{\Gamma}$ *cannot be identified from* $\mathbf{M}^{\mathbf{I}}$ *and* $\mathbf{M}^{\mathbf{D}}$ *alone. However,* $\hat{\mathbf{\Gamma}}$ *is consistently estimated from* $\hat{\mathbf{T}}_{\mathbf{s}}^{\mathbf{I}}$ *when the latter is estimated from* $\hat{\mathbf{M}}^{\mathbf{I}} - \hat{\mathbf{M}}_{\mathbf{m}}^{\mathbf{I}}$.

We now implement this estimation using two redesigns of common household surveys in the US.

2.2 Inferring Miscoding from Standard Labor Market Surveys

Two major survey redesigns contain information that can be used to estimate $\mathbf{\Gamma}$: the 1994 Current Population Survey (CPS) and the 1985-1986 Survey of Income and Program Participation (SIPP) redesigns.[24] To keep the occupation classification as comparable as possible across time, we map all original occupations into the harmonized 1990 Census Occupation Classification variable, provided by IPUMS (Flood et al., 2023).[25] We consider three levels of aggregation: the original level of 3-digit aggregation, the 25 major occupational groups (which we will refer to as 2-digits) and 13 major occupational groups, having merged all service occupations into one (which we will refer to as 1-digit).

Survey of Income and Program Participation (SIPP) Redesign (1986). The 1986 SIPP panel introduced dependent interviewing, contrasting with the independent interviewing used in the 1985 panel, with which it overlaps for more than a year. The survey design changes follow the one discussed in the theoretical exposition. Until the 1985 panel all workers were asked to describe their job anew at the moment of an interview, without

[24] We have also investigated further data sources, where stronger additional assumptions need to be made for identification. For example, we investigated the CPS job and occupational tenure supplement to isolate occupation stayers that are independently coded across interviews 4 and 5 of the basic monthly survey. The sample size of this exercise is smaller and Vom Lehn, Ellsworth, and Kroff (2022) highlight that the tenure supplement tends to overestimate occupational staying. We also considered temporary layoffs under the assumption that workers return to their old occupation, but are coded independently upon their return. This is a much smaller sample. Also perhaps some of those in temporary layoff transition to permanent layoff, followed by a true occupational change. Finally in ongoing work, we are comparing the observed occupational mobility between interview 4 and 5, and try to remove true occupational mobility in the intervening month by investigating the patterns observed in interviews 1-4 and 5-8 of other rotations that coincide with the missing 8 months between interview 4 and 5. However, one needs to make structural assumptions of how flows correlate across individuals during 8-month periods when we only observe 4-month snapshots.

[25] Specifically, the *occ1990* variable. Some parts of our analysis requires combining our occupational flow data with the Department of labor's O*NET project. Since the latter does not capture all *occ1990* codes, we follow Guvenen, Kuruscu, Tanaka, and Wiczer (2020) in aggregating several individual occupations *occ1990* in a manner that allows us to link it to the O*NET data. We are left with 333 occupations at the 3-digit level of aggregation.

reference to answers given at an earlier interview. In the 1986 panel, instead, the practice changed such that 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.

We exploit the time overlap between these panels from February 1986 to April 1987, thus we capture the same population under different survey designs. We focus on full-time workers with a single employer, applying additional restrictions to ensure comparability.[26] We use the flow matrices of dependently and independently coded individuals in the two adjacent waves to estimate $\hat{\mathbf{M}}^{\mathbf{I}}_{\mathbf{s}}$.[27]

The Current Population Survey's Redesign of 1994. In the Current Population Survey (CPS), respondents are interviewed every month for four consecutive months, rotated out for eight months, and then interviewed again for four more consecutive months. Before the 1994 redesign, the CPS collected occupation and other data in each interview month independently, similar to the SIPP before its 1986 redesign. Respondents provided information about their current occupation (or past occupation in case of recent job loss) in each month, with the CPS independently coding this information for each interview. After the 1994 redesign, respondents were interviewed dependently about their occupations in interviews 2-4 and 6-8. As with the SIPP redesign, the occupational code would only be asked again if either a change in employer or work activities was reported, otherwise the previous month's occupation was carried over.

By treating the period just after the redesign as informative about $\hat{\mathbf{M}}^{\mathbf{D}}$ and the period

[26] This is, with small changes, the sample on which the *1-digit* miscoding correction used in Carrillo-Tudela and Visschers (2023c) was estimated. We include only workers who remained in full-time employment throughout two waves and reported having only one employer at any point in time. We exclude workers who experienced non-temporary layoffs with short unemployment episodes, those with imputed occupations, and those enrolled in school. Additionally, we restrict our sample to individuals between 19 and 66 years old. After applying these restrictions, we obtain 28,302 wave/individual observations for the 1985 panel, 27,801 for the 1986 panel, and 5,922 for the 1987 panel.

[27] In particular, we employ a two-step process: First, we calculate the occupation flow matrix across two adjacent waves for workers in the 1985 panel, using SIPP-provided individual sample weights. We then subtract from this matrix the occupational flow matrix of independently interviewed individuals across two adjacent waves. This calculation yields $\hat{\mathbf{M}}^{\mathbf{I}}_{\mathbf{s}} = \hat{\mathbf{M}}^{\mathbf{I}}_{\mathbf{85}} - \hat{\mathbf{M}}^{\mathbf{D}}_{\mathbf{86,87}}$. To address issues arising from finite sample size, we average the flows between occupations i and j in both directions in $\hat{\mathbf{M}}^{\mathbf{I}}_{\mathbf{s}}$. Finally, we set any remaining negative elements to zero, a step particularly important for more detailed classifications where the number of elements in the flow matrix is larger relative to the number of observations.

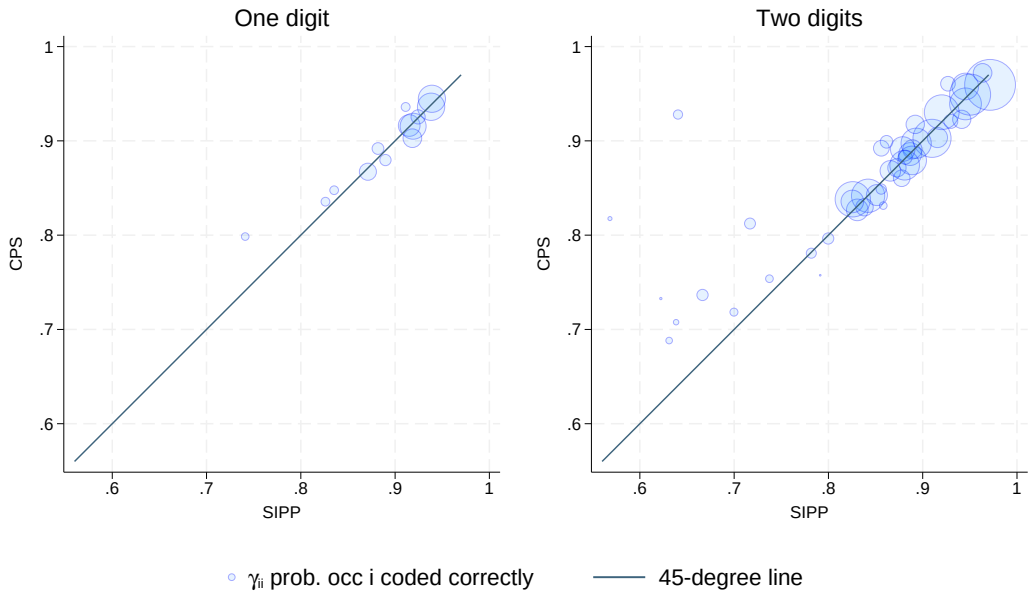
just before as informative about $\hat{\mathbf{M}}^{\mathbf{I}}$, we can derive $\hat{\mathbf{M}}^{\mathbf{I}}_{\mathbf{s}}$ from the changes in occupation flows associated with the 1994 CPS redesign. In Appendix B.1, we described our approach on these data in more detail. We also present suggestive evidence that the fall in spurious flows during 1994 likely is due to the introduction of dependent coding, and not due to other structural changes in the U.S. economy during that period. The changes in mobility rate across the four and fifth interview, which remain been independently coded, are small – at least a magnitude smaller around the redesign than those observed across other interviews (affected by the survey redesign). Overall, observed occupational mobility at the 3-digit level across interviews 4 and 5 stayed close to 40% from 1987-2002. In addition, the vast majority of workers who are employed in two adjacent months are employer and activity stayers, which suggests that pre-redesign most observed mobility is driven by miscoding. Indeed (as shown in the appendix), the observed month-to-month occupational mobility post-redesign (which still includes miscoding of employer/activity switchers) fell by more than 80%.[28]

The CPS exercise is based on considerably more observations than the SIPP exercise, and allows us to estimate miscoding at the level of over 300 three-digit occupations, in addition to the more aggregate classifications. In the SIPP, we can estimate miscoding at the one- and two-digit level, using the lower number of observations when keeping to the cleanest exercise of literal sample design overlap in time.

Estimated Miscoding At the one- and two-digit classification level, we first estimate the garbling matrix $\hat{\mathbf{\Gamma}}$ separately across the two survey redesigns. This allows us to get a sense of consistency and persistence during nearly a decade of the miscoding probabilities. Figure 1 plots the diagonal elements of $\hat{\mathbf{\Gamma}}$ based on the CPS 1994 redesign (y-axis) with those estimated using the 1986 SIPP redesign (x-axis). The key take away is that the probabilities of coding a one-digit occupation correctly estimated using these different sources of data are highly correlated, aligning well with the 45-degree line. Deviations from this line can reflect small sample noise (in particular, in the SIPP and for small occupations) and true

[28]We use 2.18 million of month-to-month worker (mobility) observations after the redesign, and 2.04 million before. Only 3.7% of the after observations are independently coded, which suggest that well over 95% of pre-redesign observations are independently coded occupation stayers, pre-redesign. This lines up with Moscarini and Thomsson (2007), whose calculation yields that nearly 90% of month-to-month occupational mobility pre-redesign is spurious, which also suggessts that over 90% of month-to-month observations concerns true occupation stayers. The size of this CPS sample allows us to study miscoding at the three-digit level.

Figure 1: Correlation of Diagonal Probabilities γ_{ii} of $\mathbf{\Gamma}$ matrices from the SIPP and CPS



X-axis: γ_{ii} *from* $\mathbf{\Gamma}_{\mathbf{SIPP}}$*; y-axis:* γ_{ii} *from* $\mathbf{\Gamma}_{\mathbf{CPS}}$*. Left panels: one-digit occupational aggregation (13 occupations). Right panels: two-digit aggregation (45 occupations). Dashed line: 45-degree line. The size of each circle is inversely proportional to the product of the bootstrapped standard errors of the SIPP and CPS estimation. When weighted by the 1994 population of employed workers, the average one-digit miscoding rates are 9.5% (CPS) and 9.7% (SIPP). The two-digits miscoding rates are 12.6% (CPS) and 13.5% (SIPP).*

Table 1: Estimated Miscoding $\mathbf{\Gamma}$ Matrix based on CPS & SIPP (pooled, one-digit classification)

	mgmt	spec	tech	sale	admn	serv	farm	mech	cnst	prod	oper	trsp	labr
managers	**86.1**	2.4	0.4	3.2	4.3	1.0	0.2	0.4	0.8	0.6	0.3	0.2	0.1
prof spec.	2.2	**93.3**	1.3	0.4	1.3	0.8	0.1	0.2	0.1	0.2	0.2	0.0	0.0
technic.	1.5	5.5	**84.3**	0.3	2.9	1.6	0.1	1.7	0.3	0.5	1.1	0.1	0.2
sales	3.6	0.5	0.1	**91.1**	2.0	0.8	0.1	0.3	0.1	0.3	0.1	0.3	0.7
admin	3.5	1.1	0.6	1.5	**90.6**	0.9	0.1	0.2	0.0	0.2	0.4	0.2	0.7
service	1.0	0.9	0.4	0.7	1.1	**94.0**	0.2	0.4	0.2	0.2	0.2	0.2	0.4
farm/fish	1.0	0.3	0.2	0.3	0.4	1.0	**93.8**	0.4	0.5	0.1	0.4	0.6	0.9
mechanics	1.5	0.7	1.5	1.0	0.7	1.4	0.3	**87.1**	1.7	1.3	1.5	0.5	1.0
constructn	2.7	0.3	0.3	0.2	0.2	0.7	0.3	1.6	**88.6**	0.9	0.9	0.9	2.4
prec.prod	2.8	0.8	0.5	1.1	1.1	0.7	0.1	1.7	1.1	**83.0**	5.7	0.4	1.1
operators	0.5	0.4	0.6	0.2	0.9	0.4	0.2	0.9	0.5	2.6	**89.8**	0.4	2.6
transport	0.7	0.1	0.1	0.9	1.0	0.5	0.4	0.5	0.9	0.3	0.6	**92.2**	1.9
laborers	0.5	0.2	0.2	2.4	3.4	1.4	0.7	1.1	2.9	0.9	5.0	2.3	**79.0**

changes in miscoding probabilities. That the largest deviations from the 45-degree line appear for small occupations, is suggestive of the former. Hence, aggregating across these data sets appears beneficial. Below, for one- and two-digit classification levels, we will use the miscoding matrix aggregated across both survey redesigns (with the aggregation procedure explained in the Appendix).

Table 1 shows the estimated garbling matrix at a 1-digit level obtained from the above

procedure. The rows represent the origin occupations, and the columns the destination occupations at a 1-digit level of aggregation.

The diagonal of the matrix shows that different occupations have different propensities to be assigned the correct code. For example, we find that individuals whose true occupation is "laborers" have a 79% probability of being coded correctly, while individuals whose true occupation is "professional speciality (prof spec.)" have a 93% probability of being coded correctly. The off-diagonal elements instead show that, given a true occupation, some coding mistakes are much more likely than others. This can be visualized through the heat map associated with each row. For example, workers whose true occupation is "laborers" have a much larger probability to be miscoded as "machine operators (operators)" (5%), "admin. support" (3.4%) or "construction" (2.9%) than as "managers" (0.5%) or "professionals" (0.2%).

Taken together, 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 either the source or destination occupation is miscoded, the probability of observing a spurious transition for a true occupational stayer is around 20%, even at the 1-digit level. Thus, our methodology suggests that coding error is indeed substantial under independent interviewing, even at the one-digit level.

We can investigate how our estimation method relates to other approaches to reduce the impact of miscoding. For example, a common approach when using pre-1994 CPS data or longitudinal survey data (like the NLSY or PSID) is to assume miscoding when encountering transitions from occupation i to j immediately followed by a return to i (see Neal (1999) and Moscarini and Thomsson (2007)).[29] Rather than 're-coding' these "move-and-immediate-return" observations as true occupation stays, the *frequency* with which these patterns occur can actually be taken as informative of miscoding probabilities. Simply put: if miscoding between occupations A and B is more frequent than occupations A and C *and* most of this type of return mobility is spurious mobility, we would expect to observe more $A \to B \to A$ return moves than $A \to C \to A$, in close proportion.

To investigate the relation between our miscoding probabilities and return mobility, we consider CPS data prior to 1994 and focus on continuously employed workers with completed four-month interview sequences of the form A-A-B-A, A-B-A-A, and A-A-A-A, where

[29] Moscarini and Thomsson (2007) observe a significant drop in these type of patterns after the 1994 CPS redesign, suggesting that under independent coding most of these patterns are indeed spurious.

A and B are different occupations. The details are in Appendix B.3, but to summarize: we indeed document a strong positive relationship, close to one-for-one, between the share of A-A-B-A and A-B-A-A patterns among all completed four-month interview sequences and the probability that occupations A and B are miscoded.[30] This supports that, simultaneously, this type of mobility is mostly spurious in the CPS and that the implications of our miscoding probabilities are in line with the data (along a dimension that was not used in our estimation).[31]

2.3 Correcting Occupational Flows

To use the estimated miscoding matrix $\hat{\mathbf{\Gamma}}$ and correct observed occupational flows occuring between occupations i, j (including $i = j$, i.e. occupational stays), we start from an observed occupation flow matrix under independent interviewing $\hat{\mathbf{M}}^{\mathbf{I}}$. Then we pre- and post-multiply this matrix with the inverse of the $\hat{\mathbf{\Gamma}}$ matrix (or its transpose), such that our estimate of the true flow matrix, $\hat{\mathbf{M}}$, is given by[32]

$$\hat{\mathbf{M}} = (\hat{\mathbf{\Gamma}}')^{-1}\, \hat{\mathbf{M}}^{\mathbf{I}}\, \hat{\mathbf{\Gamma}}^{-1}, \tag{2}$$

where we recall that the miscoding matrix $\mathbf{\Gamma}$ is invertible by virtue of the properties derived in Proposition 1. Given the law of large numbers and Assumption *(A1)*, we also have that

$$\operatorname*{plim}_{n\to\infty} \hat{\mathbf{M}}_{\mathbf{n}} = (\mathbf{\Gamma}')^{-1}\, \mathbf{\Gamma}'\mathbf{M}\mathbf{\Gamma}\, \mathbf{\Gamma}^{-1} = \mathbf{M}.$$

That is, the sample estimator $\hat{\mathbf{M}}$ is a consistent estimator of $\mathbf{M}$.

2.4 Corrected Occupational Mobility

We now apply our miscoding correction model to measure occupational mobility in the United States, for different subsets of workers. Figure 2 presents uncorrected and corrected

[30] Formally we estimate the relationship $s_{a,b} = \mathbf{\Gamma}_{a,b} + \epsilon_{a,b}$, where $s_{a,b} \equiv \frac{N_{a,b}}{\sum_{b'} N_{a,b'}}$ and $N_{a,b}$ is the number of observed 4-interview transitions for origin occupation a and temporary occupation b, $\mathbf{\Gamma}_{a,b}$ is the miscoding probability between a and b, and $\epsilon_{a,b}$ is assumed to be Gaussian error. We find that the fit of this relationship (R^2) varies from 0.96 at a 1-digit level to 0.86 at a 3-digit level.

[31] In a companion paper (Carrillo-Tudela, Darougheh, and Visschers, 2025), we use the Danish matched employer-employee wage-payment based administrative data and find that the occupational mobility patterns of EE movers in these data line up perhaps surprisingly well with their US Current Population Survey counterparts, after miscoding correction.

[32] Poterba and Summers (1986) use this correction for miscoding of the (three-state) labor force status, where they could observe the garbling matrix directly.

occupational mobility rates of different sets of workers, at the three-digit level.[33] In the left panel, we consider workers who are switching employers, directly (EE) and through unemployment (UE). In the raw (uncorrected) data we observe that nearly two-thirds of the UE and 60% of EE movers change their three-digit occupation. While spurious mobility is important (about one-fifths for UE and one-fourth of EE movers), the underlying corrected mobility remains high at around 55% for UE and 45% for EE movers.

Correcting miscoding increases the net mobility rate instead of decreasing it. (Here, occupation identities matter and our 'degarbling' method goes beyond adjusting spurious flows.) Higher net mobility makes a minor contribution to the lower difference, post-correction, between gross and net mobility rates. However, for employer movers this difference, excess occupational mobility, remains substantial – even as miscoding is addressed. Conclusions that excess occupational mobility is simply a measure of the extent of spurious mobility and miscoding, appear wholly unwarranted.

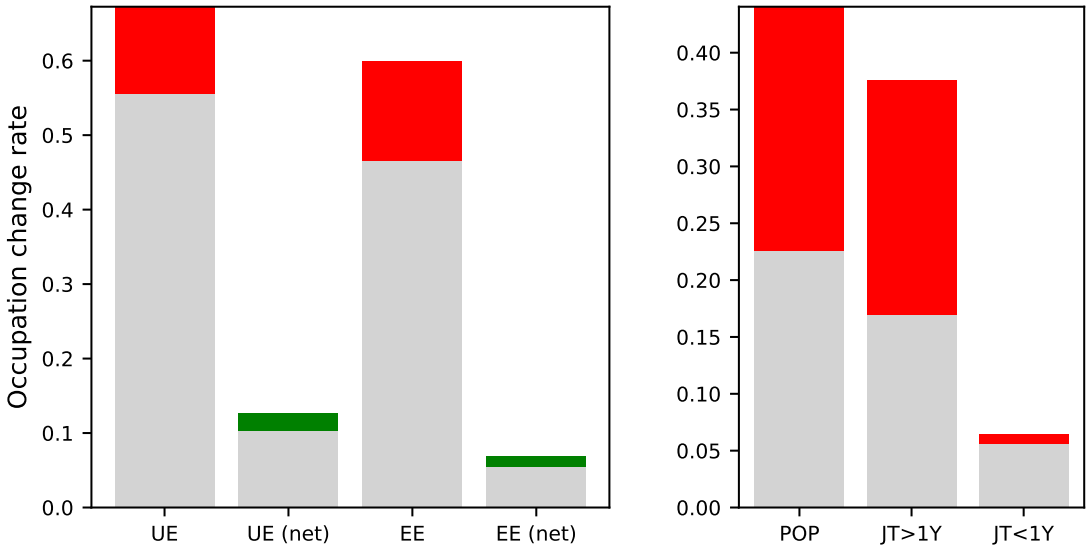
If, instead of concentrating on employer movers, we consider the occupational mobility of all workers over a year, the impact of miscoding is very different. In the right panel, we compare the occupation of an employed worker at a point in time with their independently coded occupation one year before. We find a high mobility rate in the raw data (close to 45%), but considerably lower mobility after correction, around 22%. This is rather close to the around 20% mobility that Kambourov and Manovskii (2008) derive for year-to-year occupational mobility at the three digit level.

We can split the workers in the right panel into two subsets, those who are one year or longer with their employer and those who are not. We plot the contribution of these subsets to the overall mobility rate (POP). Most of the occupational mobility observed across 12 months takes place among those who remain with their employer in the raw data. A large fraction of these flows – about half – are spurious. In contrast, a much larger proportion of the observed occupational mobility of recent employer movers is genuine (the right-most bar in the right panel). Still, because those workers with an employer tenure of more than a year are a much larger part of the population, they contribute more than 75% to the economy's 12-month occupational mobility.

Given the same miscoding probabilities, the relative correction for spurious mobility

[33] This figure concerns the 1980s, specifically 1983 and 1987: when the CPS used independent interviewing but job tenure and EE information can be gained from CPS tenure supplements. Similar figures can be drawn for later periods.

Figure 2: Occupational Mobility, Raw and Corrected, across Worker Subsets



Gray: corrected data. Red: reduction of mobility due to correction. Gray+Red: uncorrected data; Green: increase of mobility due to correction. Both panels: occupational mobility shares at the three digit aggregation. ***Left panel****: Occupational Mobility Rates of Employer Changers. UE: Unemployment-to-employment movers, excluding (temporary) layoff. EE: Employment-to-employment movers.* ***Right panel****: Year-to-Year Occupational Mobility. POP: Mobility rate of the population of all workers that were employed in both years. JT>1: Contribution to this rate of the subpopulation of workers with job tenure larger than one year – job stayers. JT<1: Contribution of workers that stayed at the same job for less than a year – job movers. We use the reported job tenure in the CPS tenure supplements from 1983 and 1987 to identify EE movers (in the left panel) and split by job tenure (in the right panel), combined with data from the basic monthly CPS during these years.*

varies by the set of workers considered. Behind this, miscoding tends to change a true occupational stay into an observed occupational mover, but for a true occupational mover typically changes the origin or destination without turning them into an occupational stayer. Using $\mathbf{\Gamma}^{-1}$ in equation (2) to counteract miscoding takes this into account, while rule-of-thumb adjustments (e.g. "60% of mobility is spurious") do not.

3 Occupational Miscoding and Task Distance

Beyond whether an occupational move occurs or not, the 'quality' of the move also matters. A sizeable body of literature has found that 'task distance' measures that quantify occupational task (dis)similarities between origin and destination occupations predict wage changes, relative frequency of transitions, match longevity, etc.[34]Hence, to fully capture the importance of task distances, it appears essential to measure sufficiently accurately the

[34]The standard explanation is that workers can retain a significant portion of their human capital if the new occupation involves similar activities and skills as their previous one (see e.g. Poletaev and Robinson (2008), Gathmann and Schönberg (2010) and Guvenen, Kuruscu, Tanaka, and Wiczer (2020)).

start and end points of workers' occupational transitions.[35] A few papers highlight potential biases due to inaccurate measurement of task distances due to miscoding, e.g. Abraham and Spletzer (2009) and Speer (2016).

However, using the miscoding matrix $\mathbf{\Gamma}$ that contains the bilateral miscoding probabilities, $\gamma_{i,j}$, we will now investigate the impact of miscoding on task distance measurement, illustrate how this can affect the empirical relation between task distance and economic outcomes and, last but not least, how we can apply our miscoding correction to address this. Even with the same elements on the diagonal of the miscoding matrix $\mathbf{\Gamma}$, the impact of miscoding on task distance measurement may vary greatly, depending on whether or not miscoding mainly takes place between occupations that are very close in (O*NET-based) task distance.

Below, we focus especially on the sample of occupational changers among employer switcher, separating them by EE and UE transitions. We chose to focus on employer and occupation movers because, to this day, their occupational mobility in standard household surveys remains affected by independent interviewing. In addition, these transitions are interesting in themselves: they typically involve a much larger range of task changes and economic outcomes than those who changed occupations with the same employer (who most likely following a narrower career progression).

3.1 Occupational Miscoding, Distance and Mobility

We begin by studying the relationship between task distance and worker mobility, and how miscoding impacts their observed relationship. Consider the observed occupational mobility flows (or lack thereof) of a particular subset of workers of interest, for example those who changed employers directly $\hat{\mathbf{M}}^{\mathbf{I}}_{\mathbf{EE}}$. Correcting for miscoding yields $\hat{\mathbf{M}}_{\mathbf{EE}} = (\hat{\mathbf{\Gamma}}')^{-1}\,\hat{\mathbf{M}}^{\mathbf{I}}_{\mathbf{EE}}\,\hat{\mathbf{\Gamma}}^{-1}$. For simplicity and generality (because the procedure will be repeated for workers who find jobs through unemployment), we drop the EE subscript and refer to $(i,j)^{th}$ element of $\hat{\mathbf{M}}_{\mathbf{EE}}$ as $\hat{m}_{ij}$. With each flow ij, we associate a distance measure $d(i,j)$ that we construct using standard methods of the literature detailed next. With that done, we will relate distances

[35] These investigations typically proceed in two parts: one considers workers' labor market outcomes over time, including coded occupations. The second part establishes the mapping from occupation codes into a 'task distance' measure. For the US, the data used is typically from the US Department of Labor's O*NET project, which reports a large amount of occupation descriptors that include information on task and work activities (as well as knowledge and skill requirements). See `https://www.onetcenter.org/overview.html`

Table 2: Principal Component Analysis (PCA): Summary by Level of Aggregation

Level of aggregation	Number of PCs	%Variance explained
One	5	85.1
Two	25	95.8
Three	25	79.8

to flows.

Measuring Task Distance We follow previous studies and use the O*NET dictionary to assign tasks to each 3-digit occupation. We reduce O*NET's 120 task dimensions through principle component analysis (PCA), and use these principle components to characterize each occupation as a n-dimensional vector of task intensities. Table 2 lists the number of principal components and the share of variance explained across different levels of occupational aggregation, such that each 1-digit occupation is described by a 5-dimensional vector and each 2- and 3-digit occupation by a 25-dimensional vector.[36]

We follow Guvenen, Kuruscu, Tanaka, and Wiczer (2020) and employ a Euclidean distance metric over task intensities to evaluate how close or far apart are corrected and uncorrected occupational moves. In particular, we take the task distance between two occupations i and j to be determined by

$$d(i,j) = \left(\sum_{n=1}^{N} (q_{i,n} - q_{j,n})^2 \right)^{1/2}, \tag{3}$$

where $q_{i,n}$, $q_{j,n}$ correspond to the task intensity along the n^{th} principal component (out of N components) in occupation i and occupation j, respectively.[37] This distance captures how different two occupations are in their task requirements: occupations that use similar combinations of tasks will have small distances, while occupations with very different task requirements will have large distances.

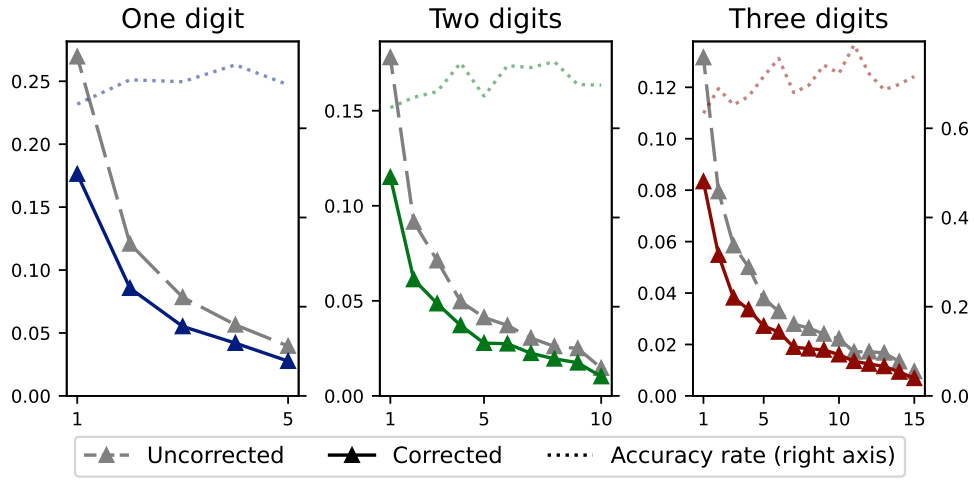
Occupational Flow Frequency and Distance To relate occupation flows to distances, we divide the task distance spectrum into quantile distance bins $B(n)$; this means that we focus on the ordinal aspect of task distance, facilitating comparisons with other distance defini-

[36] All principal components are obtained after orthogonalizing them via a varimax rotation. Appendix **??** describes the most important loadings of all components.

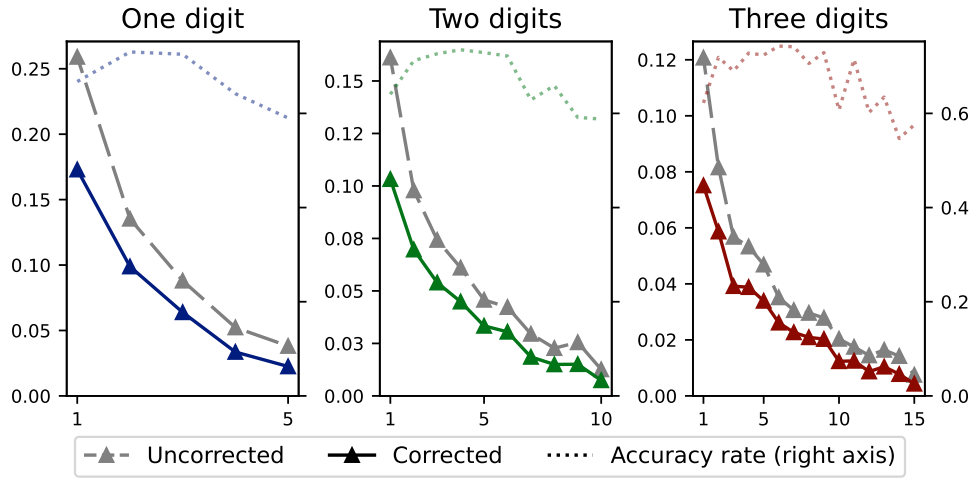
[37] The 1- or 2-digit distances are constructed from the underlying 3-digit occupations, where we used the relative frequency of 3-digit transitions to compute the weighted average at the 1-digit and 2-digit aggregation. We find that across these levels of aggregation, other task-based distances such as the Angular distance and the Manhattan distance deliver very similar results.

Figure 3: Raw and Corrected Flows of Employer Changers across Task Distance Quantiles

(a) Employer-to-Employer (EE) Movers



(b) Hires from Unemployment (relative to previous occupation)



The figure plots the proportion of employer transitions with a task distance that falls within a given distance quantile, for all quantiles. One digit: 5 quantiles; two digit: 8 quantiles; three digit: 15 quantiles. The gray lines depict the raw data with independent interviewing. The colored lines depict the miscoding corrected data. The dotted lines indicate the accuracy rate – the ratio of corrected flows relative to uncorrected flows. The graphs do not show the proportion of workers for whom occupation and tasks are identical before and after transition. We weight each occupation pair by the number of workers in the origin occupation. Data: UE transitions (1982m1-2024m9) and EE transitions (1994m2-2024m9) in the CPS.

tions, e.g. in section 4. To be precise, $B(n)$ refers to the n-th $\bar{n}$-quantile of the distribution of all possible non-zero distances $d(i,j), i \neq j$, each distance weighted by the proportion of the sample originally in occupation i (i.e. $\sum_j \hat{m}_{ij}$). Thus, with five bins, the first bin $B(1)$ corresponds to the 20% shortest-distance transitions available to workers, the second bin $B(2)$ corresponds to the 20% to 40% shortest-distance transitions available to workers, and so on, given these workers' origin occupation and weighing each worker equally.

Figure 3 displays the relationship between the amount of observed flows and (non-zero) task distance, aggregated in the quantile bins $B(n)$. We use 5 bins for distances between one-digit occupations, 10 bins for two-digit occupations and 15 bins for three-digit occupations.[38] In the top panel, we consider EE moves while in the second panel we consider UE moves.

We observe across all aggregations and type of employer changes, that closer task distances are associated with substantially more mobility.[39] Further, the miscoding correction lowers mobility substantially, from the dashed gray line to the colored solid line. Adding up the vertical differences between the observed flows without miscoding correction and after miscoding correction corresponds to the amount of spurious flows. The dotted line in Figure 3 shows that the proportional reduction in mobility upon correction is relatively uniform across distance bins, more so than one may perhaps have expected. Concretely, post-correction mobility hovers around 65% of the raw mobility (right axis) for both EE and EU movers. This finding suggests that the impact of miscoding goes much beyond creating only spurious flows that exclusively have very short observed task distances. Instead, miscoding wrongly increases the amount of flows at almost all task distances.

How does Miscoding shift Observed Flows and Distances? Behind the impact of miscoding documented in Figure 3, inflated observed mobility across all distances, can still be two types of explanations. On the one hand, miscoding could turn occupation stayers into low-distance occupation movers, low-distance movers into higher-distance movers, etc., moving a large amount of workers up by perhaps relatively small distances, to create shift of the entire distribution of distances. On the other hand, miscoding might create spurious flows that at times have small, at other times also have large observed task distances. In this case, upon observing an occupation change with large task distance, we cannot necessarily be sure that the worker is in fact changing occupations at all.

To gauge these two explanations, we now investigate the probability distribution of true

[38] To exemplify the distance bins, consider the first bin of the 15 three-digit task distance bins (corresponding to the right-most patterns of Figure 3), which contains e.g. financial managers who move to other management positions, carpenters who take up other construction trades, or nurses that become physicians. Proximity in terms of tasks does not automatically imply that it is always easy to switch occupations in practice. To illustrate the later, the seventh bin (which corresponds to the median of distances of all *possible* transitions - much beyond the median distance of *actual* transitions), contains financial managers who become bank tellers, carpenters who take up secondary school teaching or nurses who move to secretarial jobs.

[39] Note that if workers randomly chose among destinations (if they were changing occupations), we would observe a flat line.

occupational distances given a collection of observed distances in the data that is subject to miscoding.[40]We again analyze this probability at the level of quantiles of quantile task distances. Formally, define $QBF(n,n')$ as the amount of true flows (indexed by (i,j)) in the n-th distance quantile bin $B(n)$ that instead –after miscoding– are observed (in the raw data) to have distances (k,l) in the n'-th quantile bin, $B(n')$. To distinguish inferred distances after miscoding correction from observed distances in the raw data, from now on we denote the latter by $B^{raw}(n')$. We then have that

$$QBF^{miscode}(n,n') = \sum_{(k,l)\in B^{raw}(n')} \left(\sum_{(i,j)\in B(n)} \hat{\gamma}_{i,k}\hat{\gamma}_{j,l}\hat{m}_{ij} \right).$$

To include spurious flows, we define $B(0)$ as the set of occupation pairs that are occupational stays, i.e. $B(0) = \{(i,i), i = 1,\ldots,O\}$. Then, let $QBF^{miscode}(0,n)$ refer to the mass of true occupation stayers that appear (in the raw data) to have moved a task distance in the n-th distance quantile bin. The conditional probability of a true distance falling in quantile bin n given an observed flow in a quantile bin n', defined as $\mathbb{P}(B(n)|B^{raw}(n'))$ with abuse of notation, is then given by

$$\mathbb{P}(B(n)|B^{raw}(n')) = QBF^{miscode}(n,n') / \sum_{\tilde{n}\in 0,1,\ldots,\bar{n}} QBF^{miscode}(\tilde{n},n'). \tag{4}$$

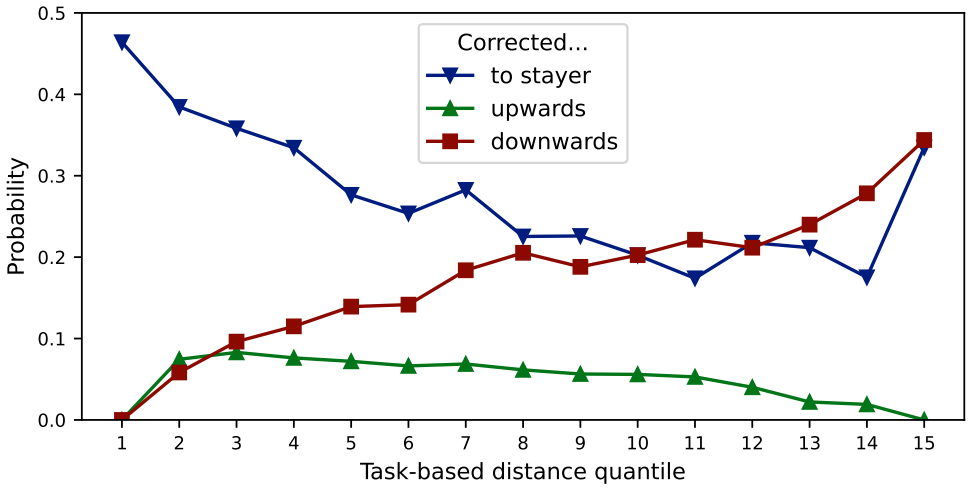
Figure 4 displays in blue (with downward pointing triangles) the probability that a flow in each distance bin of the raw data (with miscoding) corresponds to a true occupational stay, i.e. $\mathbb{P}(B(0)|B^{raw}(n'))$, across 3-digit occupations.[41] We observe that this probability is decreasing in task distance, such that for the shortest distances over 40% of observed occupational movers are actually occupational stayers and for larger distances about 20% are occupational stayers. The large proportion of actual occupation stayers among observed movers over the largest distances is quite striking and suggests that miscoding has the potential of seriously attenuating the economic implications obtained from uncorrected task distance analysis.

To continue the distance shifts caused by miscoding, the red line with square markers

[40]Note that, generally, we cannot simply read from the garbling matrix $\mathbf{\Gamma}$ alone the probability distribution over true occupation flows i,j for a worker making an observed transition from occupation k to l (under independent interviewing). Behind this is the usual Bayesian inference problem, where the aforementioned (posterior) outcome depends on the underlying distribution of true occupation flows that could give rise to an observed flow from k to l. However, combining the whole set of observed flows under independent interviewing with the miscoding matrix $\mathbf{\Gamma}$, we do have the information needed.

[41]The matrix of miscoding flows implied by equation (4) is presented in Appendix D, Table D.2.

Figure 4: Directions of Miscoding-Correction of Distances



In this figure, we plot for the three-digit occupation pairs within each of the 15 quantile of distance, the shares of observed transitions (in the raw data) that miscoding-correction moves to: (i) occupation stayers, (ii) to a smaller distance quantile ("downwards"), and (iii) to a higher distance quantile ("upwards"). Corrected downwards excludes corrections to occupation stayers. Data: all EE transitions in the CPS, 1994m2-2023m9.

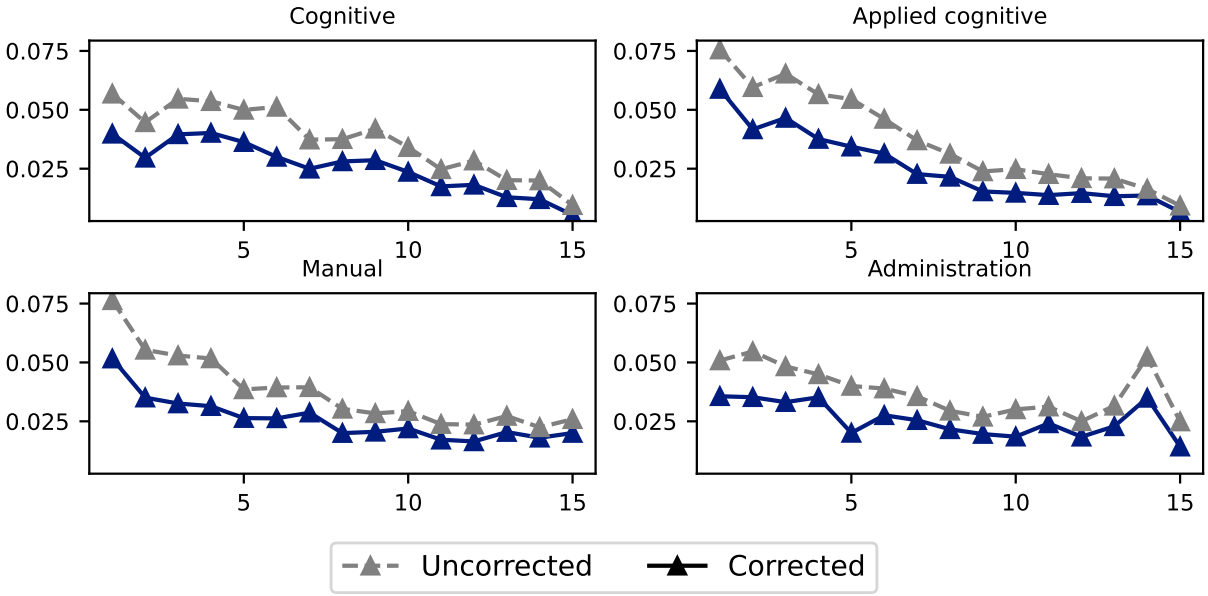
captures the estimated probability that a flow observed in the raw data in bin $B^{raw}(n')$ corresponds to a true flow, but one that is in a lower distance bin: $\sum_{0<n<n'} \mathbb{P}(B(n)|B^{raw}(n'))$. This probability increases with task distance, to 20-30% in the higher quantiles. This shows that a considerable share of observed large-distance transitions are occupational moves but across (sometimes much) smaller task distances. Again, with significant mismeasurement of task distances, the relevance and the responsiveness of economic outcomes to actual task distances may be obscured.

The third direction of distance distortion of miscoding is captured by the green line, the estimated probability that a flow observed in the raw data in bin $B^{raw}(n')$ corresponds to a true flow in a higher distance bin, $\sum_{n'<n<\bar{n}} \mathbb{P}(B(n)|B^{raw}(n'))$. Beyond the first few bins, this probability is decreasing, lies below the red line and does not appear as quantitatively important as the other two distortions. To complete the discussion, we estimate (not shown in the graph) that in more than 99.2% of the cases observed occupation stayers are true occupation stayers. This implies that miscoding very rarely leads a true occupational move to be observed as an occupational stay.[42]

Task components Our distance measure $d(i,j)$ aggregates differences across 120 task dimensions, via 25 principal components (PCs). One may wonder whether closeness along

[42]We have focused here on EE movers, but very similar conclusions apply when we consider the observed and inferred underlying task mobility of workers hired from unemployment.

Figure 5: Uncorrected and Corrected Flows of Employer-to-Employer Occupation Changers across principal components of tasks



These figures follow Figure 3, but here task-based distance is the Euclidean distance over a single principal component (PC). The four panels correspond to the top four PC. All data is at the three-digits aggregation, using EE transitions only.

certain PC task dimensions is especially related to occupational flows and whether miscoding could be obscuring this: some task dimensions could be more prone to miscoding than others.

To investigate this, we now redefine distance $d(i, j)$ to measure differences in task intensity between occupations i and j along *individual* PC task dimensions, rather than in the full multidimensional space. For each single-dimension distance measure, we again create bins corresponding to quantiles of its distance distribution (again weighted by the number of workers in the 'origin' occupation), allowing for consistent comparison across different task dimensions and with our previous results. Following the analysis in Figure 3, we display the mobility patterns with distance, both corrected and uncorrected for miscoding, for selected task dimensions in Figure 5.

Specifically, we plot this relationship for the four most important principal components in terms of explaining variation in O*NET task descriptors. These are the "cognitive" component (which involve learning and processing new information to take decisions); the "applied cognitive" component (which emphasizes the application of existing knowledge/skills), the "manual" component (which refers to the use of physical strength) and the "administrative" component (which involve clerical tasks, management and persua-

sion). Among all 25 principal components, these four dimensions exhibit the stronger relationship with mobility (including the strongest), as well as being among the most affected by miscoding in absolute terms.

Comparing Figures 3 and 5, we observe that the aggregate task distance, which takes into account all dimensions according to equation (3), displays a much stronger relationship with mobility than individual principal component dimensions. This is also the case after taking into account miscoding. Miscoding itself is relevant across all task components, rather than only for a subset. (See Figure E.1 in Appendix E for an overview across all 25 principal components.) Besides these four dimensions, the strongest impact of miscoding occurs along the "language and communication" and "interpersonal" task dimensions. The latter has the largest correction of constant and slope in the mobility-distance relationship of Figure 5 (but an overall weaker relationship with mobility than the aforementioned four dimensions). With the growing attention given to the interpersonal dimension of jobs and its relationship with economic outcomes (see e.g. Borghans, Ter Weel, and Weinberg (2006)), it may be important to consider the accurate measurement of interpersonal intensity when it is based on occupation codes in household surveys.

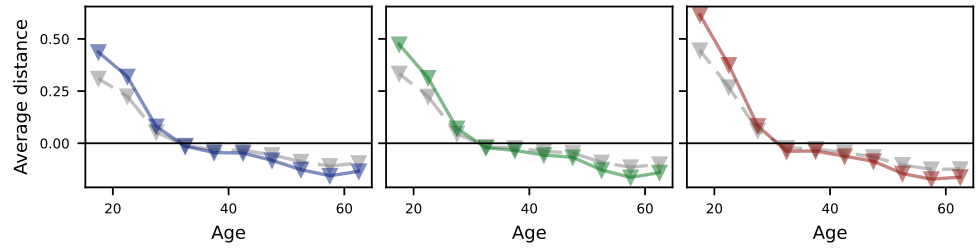
Overall, occupational miscoding is pervasive, it appears to affect many, if not all, task dimensions, while the impact on distance measurement goes far beyond mere perturbations. As a consequence, there is a large scope for it to obscure relations between occupational mobility, task distance and economic outcomes. Consequently, our views of the nature and mechanics of worker reallocation may be affected by miscoding. To look into this, we consider three aspects of labor markets that are often tied closely to reallocation: the life cycle in section 3.2, the business cycle in Appendix section C.1 and the evolution of wages in section 3.3, and investigate their relation with occupational and task reallocation – after taking into account miscoding.

3.2 Age, Mobility and Occupational Distance

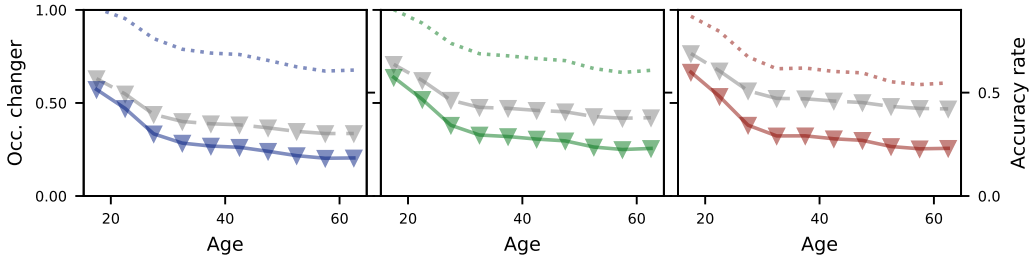
Consider first the relationship between age and workers' occupation and task mobility. From the seminal papers on task distance onwards, age and labor market experience have been a focus of attention, see e.g. Poletaev and Robinson (2008) and Gathmann and Schönberg

Figure 6: Task Distance of EE Movers, across Ages

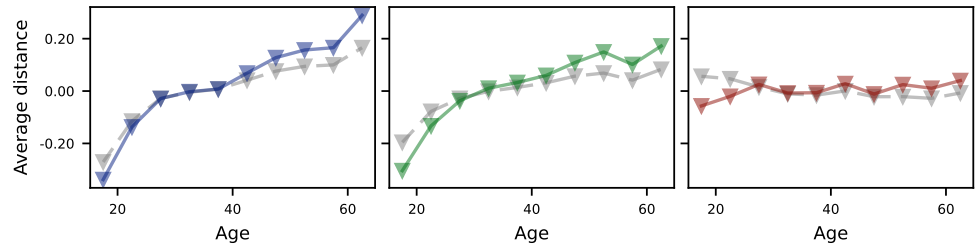
(a) Average task distance with age: all EE movers (incl. occupation stayers)



(b) Decomposition: Share of EE Movers who Change Occupations



(c) Decomposition: Average Task Distance of Occupation Changers Only



Colored solid lines: corrected data (blue: 1 digit, green: two digits, red: three digits). Gray dashed lines: uncorrected data. In the top panel, we plot the average task distance across age groups for all EE movers, including occupation stayers. The second panel plots the share occupation changers by age group. The dotted lines in this row indicate the accuracy rate of occupation movers – the ratio of true mobility to uncorrected mobility. The third panel excludes these occupation changers and repeats the first panel on the set of occupation stayers. We standardize task distances in row one using all occupations including stayers; in row three only using occupation movers. Standardizations are done pooled across all age groups, but separately for corrected and uncorrected data.

(2010). In what follows, we condition on direct employer-to-employer (EE) transitions.[43] To address the potentially differential impact of miscoding across age, we group EE transitions into ten age bins, from 18 to 65. We then calculate the matrix with all occupational flows of EE movers in an age bin a, $\hat{\mathbf{M}}^{\mathbf{I}}_{\mathbf{EE,a}}$, and correct these flows following equation (2), which results in $\hat{\mathbf{M}}_{\mathbf{EE,a}}$. Each element of this matrix is again associated with a distance according to equation (3). To interpret and compare distances across age groups, we consider all age groups ($\sum_a \hat{\mathbf{M}}_{\mathbf{EE,a}}$), and standardize the task-based distance $d(i,j)$ based on the

[43]Part of the overall decline in 12-month occupational mobility rate with age occurs because employer mobility itself declines with age. This component of decline, naturally, is not affected by the occupational miscoding issues highlighted in the paper.

mean and standard deviation of the pooled sample across all ages. We do this separately for the corrected and uncorrected data.

The first row of Figure 6 depicts the mean of the task distance changes (including the zero distances of occupation stayers) with age. We observe across all three levels of occupation aggregation that correcting for miscoding implies a stronger decline of mean task distance with age. In particular, for young age groups the age mean of task distances among EE transitions, after miscoding correction, lies about 10-15% (of the cross sectional standard deviation) higher in the all-age distribution than if we did not correct. As workers become older, the mean task distance after correcting for miscoding becomes instead lower relative to the uncorrected one. Overall, age becomes a more powerful factor to explain heterogeneity in task distances after correcting for miscoding.

The second row of Figure 6 considers the extensive margin of occupational mobility: whether workers of different ages change occupations (or not) when changing employers. While most of observed mobility rate of the young in the raw data remain after miscoding correction, of older workers close to 40%-50% of the observed occupational moves are spurious. Instead of a life-cycle drop in occupational mobility from a little over 60% (young) to a little under 40% (old), at the one-digit level, we observe –after correction– a drop from slightly under 60% to, much lower, about 20%. This considerably larger age elasticity of occupational mobility is shaped by a much larger absolute and relative impact of the miscoding correction on mobility rates at higher ages.

The third row of Figure 6 considers the intensive margin: the task distance covered by those who changed occupations. We observe again that the correction for miscoding leads to clearer or different patterns. Noticeably, the task distance of occupation changers moves in the opposite direction with age than the propensity to change occupation. Older workers who leave their old occupation move to a new occupation with tasks that are on average *more* different from their old one. In the case of the three-digit classification, we find that the correction changes the sign of the slope.[44]

Miscoding blurs the distinction between the extensive and intensive margin of occupation and task mobility, which appear to exhibit opposing patterns. Higher spurious mobility

[44]The linear relation between age and (age-)mean task distance is negative and statistically significant in the raw data, and positive and marginally statistically significant (with t-value 1.91) in the corrected data. The effect of the miscoding correction on the gradient between age and the average distance of occupation movers is to make it more positively sloped, statistically significantly at p-values below 1% across all three classification levels, in the third row of Figure 6. A similar conclusion applies to the first row in Figure 6.

of older workers blunts the underlying decline of occupational mobility with age along the extensive margin, while e.g. if spurious occupation changes come with on average lower distances, along the intensive margin, the slope of task distance of occupation movers with age may be biased downwards, in the bottom right panel of Figure 6. The joint behavior along the intensive and extensive margins guides us to potential economic mechanisms, and thus miscoding may obstruct our understanding.[45]

A very similar miscoding mechanism is at work over the business cycle. In Appendix section C.1 we document how the miscoding correction reveals that the occupational mobility rate in EE and UE transitions is more *pro*cyclical than appears in the raw data. At the same time, the corrected task distance of occupation movers in both EE and UE transitions is more *counter*cyclical than the raw data suggests: times of high unemployment are times in which occupation changers (among EE or UE movers) change their tasks more, i.e. move over larger task distances.

Further Discussion Garbling matrix $\mathbf{\Gamma}$ is assumed to be independent of observable characteristics other than occupation (including age) – yet it implies differential corrections of occupational mobility and task distance across observable characteristics. This is not contradictory. There are multiple reasons for this, some of which we highlight now. Among them is that removing miscoding noise (captured by $\mathbf{\Gamma}$) increases the relative importance of fundamentals like age or experience, to explain heterogeneity in task distance. The direction of this effect is mechanical, but it is important to quantify how strong this effect is – how important 'noise' is, vs 'signal'. This mechanism at work in the first row of the figure, where we have normalized distances by the cross-sectional variance.

Another reason, in particular in relation to the occupational mobility rates in the second row, is that the underlying *true* mobility much higher for the young. For true movers, with the same 'garbling' process, $\mathbf{\Gamma}$, one may get the identities of occupations of movers wrong but not the fact that a change of occupation has taken place. For occupational stayers, however, a miscode will almost always turn them into a spurious mover. Hence, young workers with higher underlying mobility have a much smaller relative and absolute correction in

[45] As an example of an economic mechanism: it may be the case that older workers have less employment opportunities in occupations that are close in task distance and build on their previous occupational experience, but require some measure of learning and adaptation. As a result, they would need much more to stick to their old occupation or, if they are forced to change, move to occupations with lower skill requirements at further task distances.

the second row of Figure 6 than older workers. A third reason we should highlight is that workers of different ages are in different occupations, with different miscoding probabilities. Our miscoding matrix $\mathbf{\Gamma}$ captures this heterogeneity.

This showcases an advantage of our miscoding-correction approach, which is based on capturing the miscoding process itself (under assumptions, of course), relative to ex-post rule-of-thumb corrections ("x% of mobility is spurious"). Because true mobility and occupational composition shift with fundamentals such as –here– age, the impact of miscoding also shifts. Ex-post rules-of-thumb that do not take this into account, conversely, should therefore only be applied to samples that are similar enough to those they were estimated on.

3.3 Occupational Distance and Wage Changes

A significant part of the empirical literature on individuals' task changes relates these to their wage outcomes.[46] Miscoding can obscure or diminish the importance of tasks in wage changes. We now illustrate how one can apply our correction method to address miscoding when relating task to wage changes and study the resulting relationship between task distance and wages.

We will focus here on the wage changes across workers' unemployment spells and how these vary with our O*NET-based task distance. To do so, we will compare the workers' last reported (log) hourly wage rate before unemployment to their first reported (log) wage rate after reemployment, of workers that are paid by the hour.[47] We use the SIPP 1986 to 2008 panels (spanning 1985-2013) for this, which provides higher-frequency wage observations than the CPS, and focus on workers who change occupations during an employment-unemployment-employment (EUE) transition.

[46]The list of papers is long and includes many of the papers discussed in the introduction and throughout the paper. Starting again from the seminal contributions by Poletaev and Robinson (2008) and Gathmann and Schönberg (2010), it is relevant among many areas, for example related to mismatch between task-specific human capital supply and match/firm-specific task demand, e.g. Fredriksson, Hensvik, and Nordström Skans (2018), Guvenen, Kuruscu, Tanaka, and Wiczer (2020), Lise and Postel-Vinay (2020), and Baley, Figueiredo, and Ulbricht (2022).

[47]The restriction to this set of workers is for simplicity: it allows us to compare hourly wages immediately before and after the unemployment spell, without making further assumptions and transformations that take into account the potentially partial nature of the last or first month on the job. The disadvantage of this approach is that workers who are paid by the hour are more often than average in blue-collar or less skilled jobs. A more general exercise is possible, which would include those that have a monthly salary, but also take into account e.g. that jobs do not always start at the first day and end the last day of the month, investigate the role of hours (part-time vs full-time) changes, etc.

Discussion of Correction Procedure by Wage levels One issue is that *in isolation* we cannot assign a true task distance or, probabilistically, a distribution of true task distances to a single observation of an individual worker moving from occupation k to occupation l, with only information on $\mathbf{\Gamma}$. Workers who are observed to move from occupation k to l with a large wage change may have a different underlying distribution of true occupation changes than similarly moving workers with a small wage change. However, when we have a large enough population of workers with a certain wage change, we can apply our $\mathbf{\Gamma}$-based correction to the entire occupational flow matrix of these workers, and consistently estimate the true flow matrix of this population.

To spell out the procedure, we collect the occupational mobility (or absence thereof), observed under independent interviewing, of all workers in sample who have a certain wage outcome, say a wage change Δw, and denote the associated flow matrix by $\hat{\mathbf{M}}^{\mathbf{I}}_{\mathbf{UE},\Delta\mathbf{w}}$. We then apply the correction procedure of section 2.3 to all observed occupational mobility in $\hat{\mathbf{M}}^{\mathbf{I}}_{\mathbf{UE},\Delta\mathbf{w}}$, to yield an estimate of the true flows $\hat{\mathbf{M}}_{\mathbf{UE},\Delta\mathbf{w}} = (\hat{\mathbf{\Gamma}}')^{-1}\, \hat{\mathbf{M}}^{\mathbf{I}}_{\mathbf{UE},\Delta\mathbf{w}}\, \hat{\mathbf{\Gamma}}^{-1}$. As before, every element ij of this matrix has an associated task distance $d(i,j)$. [48] In the practical implementation of this approach, we apply the above to derive the corrected flows for a set of wage change *segments*. Specifically, we consider 50 wage-change quantiles of EUE movers and correct the occupation flows associated with each of these quintiles. We then reduce the individual-level raw data to a dataset of (occupation pair, wage change segment)-combinations, each with an associated frequency of worker flows (taken from the aforementioned flow matrix) and the distance $d(i,j)$. From this, we can calculate statistics on wage (change) outcomes as a function of distance.

Results We focus on the relation between task distance and the absolute value of (log) wage changes, because distance between occupations does not have a notion of 'up' or 'down' direction. Do further task distances imply, on average, that wages change more? For occupational transitions, e.g., from i to j, we calculate the weighted mean over the absolute value of the midpoints of segments Δw, using the ij-th element of the flow matrices $\hat{\mathbf{M}}^{\mathbf{I}}_{\mathbf{UE},\Delta\mathbf{w}}$

[48] Given that we have estimated the *entire* true occupational flow matrix of those with a certain wage change, we can use it to calculate the probability distribution of underlying true moves behind observing (with potential miscoding) a transition from k to l with this wage change. This calculation proceeds along the same lines as discussed around equation (4), the probability of a true flow from k to l conditional on observing i to j is given by $m_{kl}\gamma_{ki}\gamma_{lj} / \sum_{k',l'} m_{k'l'}\gamma_{k'i}\gamma_{l'j}$, where m_{kl} refers to the kl element of the (inferred) true flow matrix (e.g. $\hat{\mathbf{M}}_{\mathbf{UE},\Delta\mathbf{w}}$) and γ_{ij} refers to the ij element of garbling matrix Γ. Therefore, our correction method allows us to derive these posterior probabilities.

and $\hat{\mathbf{M}}_{\mathbf{UE},\Delta\mathbf{w}}$ as weights for the uncorrected resp. miscoding-corrected case, and relate it to task distance $d(i,j)$.[49]

The miscoding correction effectively moves occupational transitions back to the diagonal, which across wage changes levels provides an additional validation of our correction method. Spurious occupation changers should have the wage changes of occupation stayers. Before miscoding-correction, 30.3% of the workers moving through unemployment are observed as occupation stayers in our SIPP sample, with on average an absolute wage change of 14.4%. After correction, 43.9% are occupation stayers, a size increase of this category of nearly 45%. Yet, the mean absolute wage change is hardly changed at 14.3%.[50] The correction thus puts relatively more mass back on the diagonals of the occupational flow matrices of different wage changes in such a way that the mean absolute wage change of the new mass on the diagonal is nearly identical to the corresponding absolute wage change of *stayers* in the raw data (as is, also, the mean wage change of this mass).[51]

For occupation movers, we plot in Figure 7 the smoothed relationships between absolute wage change and task distance, with the green solid line corresponding to the (smoothed) miscoding-corrected data, and the gray dashed line to the (smoothed) raw data. In addition, the green triangles indicate the absolute wage changes per (5%-)quantile in the corrected data and the gray diamonds indicate the same relationship in the uncorrected data.[52] We see that after correcting miscoding, a (true) occupational transition through unemployment of a given distance has a roughly 10-15% higher (hourly) wage change in absolute value than suggested by the raw data, given the pollution of the raw data by spurious flows and, in addition, distortions of start- and endpoints of true occupational moves.

Occupational transitions with higher task distances remain associated with larger absolute wage changes than those with shorter distances, also after correction. The smoothed relationship in the uncorrected data between absolute wage change and task distance increases, in a relatively steady manner, from 20% to around 26% from the 5th to the 95th

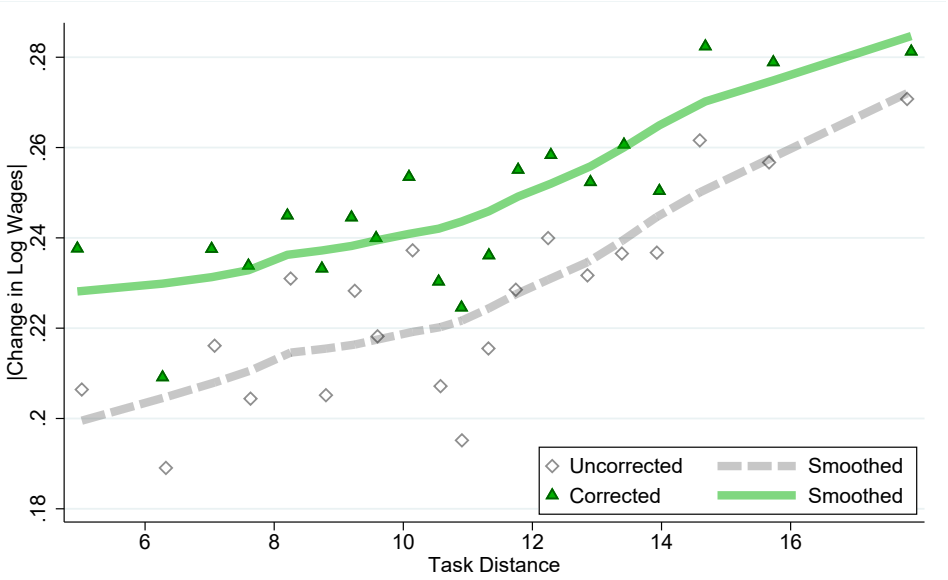
[49] As a technical point, we deal with occupation pairs that (in small samples) are associated with a negative flow after correction by resetting these to 0. We ensure that the total mass of workers with a certain wage change stays unchanged by proportionally scaling down the remaining flows.

[50] In terms of average wage changes (not in absolute values), the average wage change of occupation stayers in the raw data is small, -0.23%, and after correction nearly unchanged at-0.20%

[51] See also Figure E.2 in Appendix E.

[52] Note that now each quantile bin covers 5% of occupational transitions that *actually do occur*, different from section 3.1, where each bin covered a percentage of the *potential transitions* that workers could make, where the former clearly (and by much) stochastically dominates the latter. Further, the mean absolute wage change in each bin is calculated by again weighting each occupation pair by the frequency that this flow occurs and is plotted against the mean distance of each bin with the same weights.

Figure 7: Task Distance and Mean Absolute Wage Change, With & Without Miscoding-Correction



We plot the average absolute wage change against task distance for twenty task distance quantile bins. Our sample includes all EUE transitions in the SIPP where the new occupation ("destination") differs from the previous occupation ("origin"), with occupations at the three-digit level and wages paid by the hour. Wage change is computed as the change in (log) hourly wages. The gray dashed and green solid lines indicate the locally smoothed uncorrected and corrected estimates.

percentile. For the corrected data, we observe that this profile has shifted up across the board with a somewhat flatter slope: absolute wage changes rise from 23% to 28% from the 5th to the 95th percentile.

Figure 7 suggests that any occupational mobility, even at the relatively smallest task distances, comes with a significantly larger absolute wage change than an occupational stay. This is suggestive of a wage change discontinuity at zero task distance, perhaps related to finding of Cortes and Gallipoli (2018) that a large part of the 'transition costs' between occupations does not appear to be related to tasks as measured by O*NET or its precursor, the Dictionary of Occupational Titles (DOT).[53] If we were to take this discontinuity seriously, yet at the same time also hold that frequent miscoding only creates small deviations in task distance, including nonzero task distances close to zero for spurious movers, then we should expect miscoding mainly to seriously bias the mean absolute wage change down

[53] Strictly speaking, to be more comparable to the notion of 'transition costs', we should consider the average wage change (rather than the average of its absolute value). On average, we observe that movers through unemployment with short distances lose about 2% of their hourly wages in the raw data, and this loss is about 20% larger (around 2.4%) after correction. Beyond the mean distance we see increasingly negative mean wage changes with task distance, in line perhaps with one's ex ante intuition about moves through unemployment.

at low distances. Consequently, the slope of absolute wage changes with respect task to distance would be lower than estimated while ignoring miscoding. Instead, because spurious mobility has a presence also at higher task distances, and at the same time the garbling of occupational start- and endpoints also overstates distances in the raw data, both discussed in section 3.1, we observe that the slope, after miscoding-correction, flattens but only by relatively small amount in Figure 7.

4 Miscoding-Based Distance

In the previous section, we have argued that occupation miscoding can substantially distort observed task distances. At the same time, miscoding probability itself is not a mere flipside of task proximity constructed from O*NET. That is, it is not necessarily the case that the highest miscoding occurs for the occupations that are most similar according to O*NET. For example, we have seen that miscoding can make 'true' occupation stayers appear spuriously as movers over *large* task distances. Overall, we find a modest negative correlation between miscoding probability and task distance, at only -0.16 (unweighted).

However, at a fundamental level both miscoding probabilities and O*NET-based task similarities *are* closely related. O*NET captures a consensus of the relevance of a large set of task dimensions in an occupation. Occupations can be thought of as 'close' when this set of relevant tasks closely aligns. Miscoding also captures similarity of occupations. This happens across two survey stages: (i) respondents describe work activities in different occupations in very similar terms, and (ii) coders have difficulty assigning an occupation code unambiguously, given a description.

The key notion in this section is that the information on occupational similarity contained in miscoding probabilities themselves can be used to define an occupation distance. This information is not necessarily already reflected in O*NET-based task distance, especially given the only modest correlation between miscoding probabilities and O*NET-based task distancec. To investigate this further, we formalize a miscoding-based occupation 'dis-

tance' by defining the *semimetric* between two occupations i and j, $d_m(i,j)$, as:

$$d_m(i,j) = \begin{cases} 0 & \text{if } i = j \\ \infty & \text{if } \gamma_{ij} = 0, \gamma_{ji} = 0 \\ -\log \frac{1}{2}(\gamma_{ij} + \gamma_{ji}) & \text{else .} \end{cases} \tag{5}$$

The log transformation keeps the relation between distance and probabilities interpretable, while keeping continuity as miscoding probabilities approach 0. Averaging γ_{ij} and γ_{ji} ensures symmetry in our measure. We assigns zero distance when origin and destination occupation are the same.[54] [55]

Let us now highlight some reasons why the miscoding-based distance and O*NET-based task distance may diverge. Both condense task information into a much lower-dimensional object. O*NET measures many task dimensions across all occupations, but reports its relevance with a relatively low resolution: many dimensions are condensed to a 0-7 or 1-5 discrete scale. Our miscoding measure is based on survey responses, in particular a few lines of free-flow description that can only cover the job dimensions that the respondent finds most important. Lack of verbal precision of an individual in their interview can create additional noise.[56] Another difference is that miscoding distance by definition already is one-dimensional. In contrast, to reduce the many task dimensions in O*NET to a distance, a set of assumptions is required for dimensionality reduction (in our case, those behind principal component analysis, and the weight each of the remaining dimension). In practice, this typically involves linearity assumptions that side-step more complicated interactions among dimensions.

The remainder of this section explores our novel miscoding-based distance measure and contrasts it with the O*NET task-based distance discussed in the previous section. For brevity, we refer to the latter as 'task-based distance', even though both miscoding and

[54] Using 'distance' for this function is somewhat loose since we do not impose the triangle inequality, part of the standard mathematical definition.

[55] We can give some examples of task distance in the different miscoding quantile bins with occupations covered earlier in footnote 38. Consider financial managers, occupations in the lowest distance bin are accountants and auditors, other financial specialists, managers and administrators, not elsewhere classified (n.e.c.).; in bin 8 (median of all potential distances) are truck drivers, retail sales workers and typists. For carpenters, in the first bin there are construction workers, supervisors of construction workers and painters; in bin 8 are admin support jobs, n.e.c. and customer service reps. For nurses, in the first bin there are licensed practical nurses, nursing aides and managers of health occupations, while in bin 8, cashiers, cooks and dispatchers.

[56] O*NET also involves survey responses but reports averages those, and hence can be expected to be less subjected to idiosyncracies of individual respondents.

task-based distance are ultimately grounded in task similarities. Section 4.1 highlights the usefulness of miscoding-based distance for understanding occupational flows. Section 4.2 relates the miscoding distance of an occupational transition to wage outcomes.

4.1 Miscoding Distance and Occupational Mobility

We now investigate the relation between miscoding distance $d_m(i,j)$ and the frequency of occupational transitions, following in the footsteps of section 3.1. If miscoding distance were a better measure of occupational proximity, we would expect it to do a better job at predicting which flows actually occur.[57]

Figure 8 illustrates the relationship between the miscoding-corrected occupation flows of workers rehired after unemployment and both miscoding-based and task-based distance measures. For the x-axes in the panels in the first row we take 15 quantile bins of the 'potential' distance distribution, which considers *all* possible 3-digit transitions i, j and assigns them a distance. This means that we include those i, j with $\gamma_{ij} = 0, \gamma_{ji} = 0$, i.e. without miscoding, as part of this distribution. Also note that we again weight transitions from occupation i to j by the proportion of workers in sample originally in occupation i, so that e.g. the bottom bin corresponds to the 6.7% shortest distances available to workers, given their origin occupation.

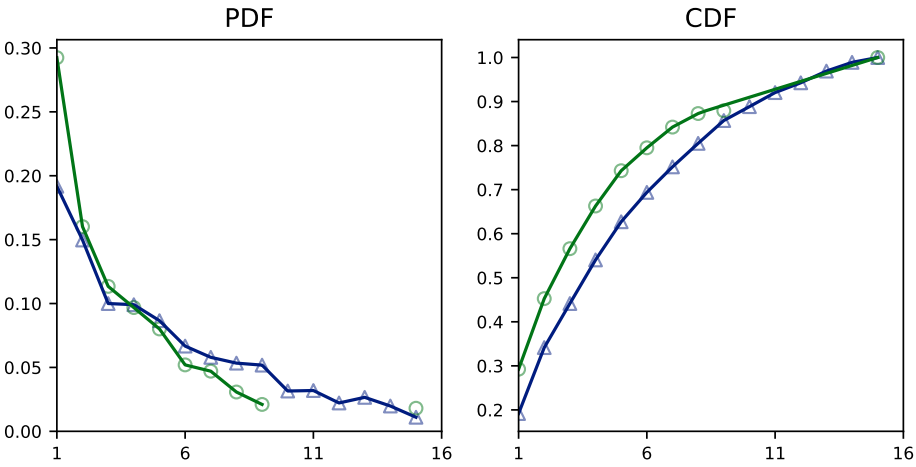
In the top row, we then plot both the density and cdf of all transitions that actually occur. We observe that the lowest quantile of occupation pairs is associated with noticeably more flows when using miscoding distance (green circles) than O*NET task-based distance (blue triangles). This extends to next three quantiles. This is also reflected in the cumulative density function of flows over quantile bins on the right. A clear gap opens up early in the cdf between miscoding and O*NET task-based distance that only closes in the last bin. The 10th to the 15th bin consists of occupation pairs with miscoding probability $\gamma_{ij} = 0, \gamma_{ji} = 0$, hence infinite miscoding distance, which cannot be ranked. We plot the density (normalized by the bin size), associated with these observations in the last bin.[58] Given these findings,

[57] The premise is that when a particular occupational move is easier or less costly, everything else equal, we would see more occupational moves occur. One reason that task similarity (whether picked up by O*NET task based distance or miscoding distance) could make transitions easier or less costly is that the worker can transfer part of his experience or ability (human capital) across tasks. More generally, occupational mobility patterns may also be driven by other dimensions than task similarity, for example shifts in labor demand due to technological change. However, we have no reason to suspect that these would be picked up better by a miscoding-based distance measure than an O*NET based distance measure.

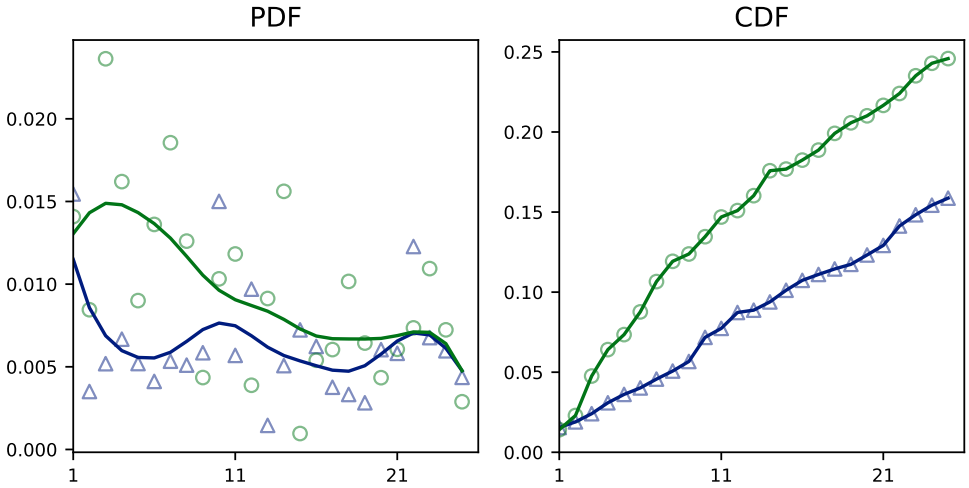
[58] To be clear, the total mass of transitions with infinite distances is about 6 (bins) times the level of the density displayed in the 15th bin. That is, around 12% of the actual transitions takes place across occupations

Figure 8: Miscoding-based distance correlates more with mobility at lower distances

(a) Entire distribution



(b) First 25/500 quantiles



We plot the PDF (left panels) and CDF (right panels) of UE transitions of occupation movers across distance quantiles. Top row: 15 quantiles; Bottom row: 500 quantiles of which we display only the first 25. Blue triangles: task-based distance. Green circles: miscoding-based distance. Solid lines in the bottom right panel correspond to smoothed values using a nonparametric kernel regression. See text for further explanation.

one can predict better whether occupation flows take place when using miscoding distances rather than O*NET task-based distances.

Since miscoding and O*NET task-based distance perform so differently in the lowest distance bin of the top panel of Figure 8, we zoom in on the shortest distances on the bottom row of this figure. Here, we have assigned, much more finely, both distance measures to 500 weighted quantile bins, and display the first 25 quantiles, i.e. the shortest 5% of distances. We can see that considerably more occupational flows are associated with the shortest miscoding distances, nearly twice as much as with O*NET task-based distances. In

that have zero miscoding between them and hence an infinite miscoding distance.

the cdf on the right, we see that a considerable gap opens up between miscoding and O*NET task-based distances as a result. The gradient of declining transition rates with distance is steeper for the miscoding-based measure in these short-distance comparisons, suggesting it captures well the fine gradations of similarity that influence workers' mobility patterns.

To provide appropriate context, in Figure 8 we focus on occupational flows that have been *corrected* for miscoding (which is based on equation (2) and further explained in section 3.1). This removes spurious flows and takes into account that miscoding affects distances of 'true' movers by garbling the origin and destination occupation. If we were to repeat the above exercise on uncorrected occupational mobility data, we would find an even stronger relationship between miscoding distance and flows than in Figure 8. However, in this case, high miscoding probabilities imply low miscoding distances but simultaneously are associated with more spurious flows. Hence, miscoding implies a structural bias in the relationship between *uncorrected* flows and miscoding distances. It is therefore crucial to correct for miscoding before we draw a conclusion about the relevance of miscoding distance.

Implementing the miscoding correction, one should be aware that noise affecting the miscoding probabilities will instead have an attenuating effect on the relationship between *miscoding-corrected flows* and miscoding distance in Figure 8. This may lead us to *underestimate* the strength of the relation between miscoding distances and flows. Intuitively, a too-high estimate of the miscoding probability between occupations i, j has two effects: we assign too short a distance to the amount of flows, $d_m(i, j)$, and we would typically consider a larger proportion of observed (i, j) and (j, i) flows to be spurious and overcorrect the amount of flows between i and j downwards. Both imply that the relationship between (over)corrected flows and miscoding distance appears less strong than it actually is. Notwithstanding this, Figure 8 shows that short miscoding distances are much more closely associated with occupation flows than short task-based distances. Behind this could be that higher miscoding probabilities are often relatively precisely estimated, while it also may be the case that the low resolution of the indices used in O*NET cannot distinguish well between "very close" and "extremely close". The latter may be the reason why we do not see a very pronounced negative slope at very low distances for the O*NET task-based distance, in the bottom row of Figure 8.

All this suggests that miscoding distance is a helpful construct to distinguish between,

especially, short and very short distances. In contrast, an infinite miscoding distance is assigned to around 12% of actual transitions and around to 40% of all possible origin/destination occupation pairs. Consequently, O*NET task-based distance may be better suited to distinguish among large distances, suggesting the relative merits of one distance measure over the other depends on the question at hand.

4.2 Miscoding Distance and Wage Changes upon Reemployment

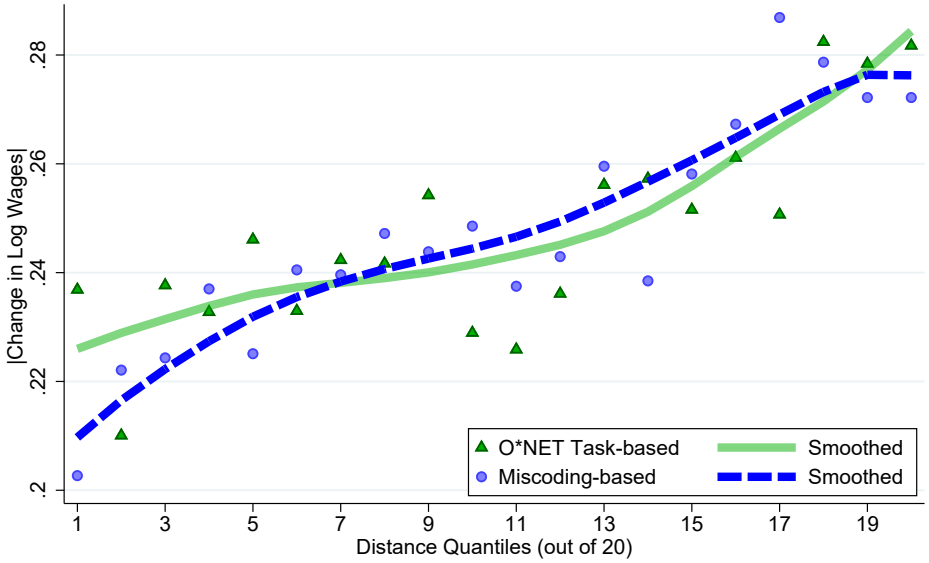
We now consider the relation between miscoding distance and the average size of (absolute) wage changes. We will do so also in comparison with O*NET task-based distance, studied already in section 3.3. If miscoding distance adds value beyond the distance from O*NET, this is suggestive that the role of task similarity can be larger than previously thought.

Concretely, to compare the two distance measures, we relate the mean absolute change of (log) wages to O*NET-based and miscoding-based *distance quantiles* on miscoding-corrected data. That is, following the procedure of section 3.3 we again split our data into 20 distance quantiles, such that each quantile contains approximately the same number of transitions. We then compute for each quantile the mean of the absolute change in log wages, weighted by the share transitions undergoing that wage change. We do this separately for both distance measures: for the miscoding distance, we plot these quantiles in blue circles in Figure 9; for the O*NET task-baed distance, we use green triangles. The dashed blue and solid green lines show the smoothed relations between absolute change in log wages and miscoding-based and O*NET task-based distance, respectively.

The shortest miscoding distances appear to come with a considerably lower absolute (log) wage change than the shortest O*NET task-based distances. The transitions with the lowest 20% of miscoding distances have absolute wage changes about 0.5-0.75 percentage point lower (statistically significant at the 5% level); for the lowest 5% of distances, this difference is even 3.4 pp (statistically significant, but with a relatively large standard error of 0.68 pp). Thus, the shortest miscoding distances pick up better those occupation transitions with (on average) smaller wage changes. Conversely, this is suggestive that very close task similarity matters more for the size of wage changes than the lowest O*NET task distances indicate.

We also observe that the absolute size of log wage changes increases more steeply with distance according to the miscoding-based measure than the task-based measure. If we

Figure 9: Lower Miscoding-based Distances Predict Lower Wage Changes



We plot the average absolute change in (log) wages by distance quantile. Green triangles/solid line: distance quantiles are defined by task-based distance. Blue circles/dashed line: distance quantiles are defined by miscoding-based distance. The solid and dashed lines correspond to smoothed values using a local polynomial regression. See text for further details.

estimate a linear relationship between the size of the wage change and the rank of the distance (between 0 and 1) over all distances, miscoding is steeper by 0.014 (with standard error 0.003) which means a predicted 1.4pp higher absolute change in log wages by moving the longest distance instead of the shortest. But, the figure already suggests a non-linear pattern, in line with the earlier intuition that miscoding distance allows us to distinguish particularly well among short distances. Thus, if we consider a linear relationship between the absolute size of log wage change and the rank of distances up to the mean, we find that moving the median distance instead of the shortest distance implies a predicted 2.48 pp larger absolute (log) wage change (with s.e. 0.56). The smoothed lines in the figure broadly tell the same story, but perhaps more easily digestible.

These patterns suggest that bilaterial miscoding propensities contain useful information about task similarities and can complement O*NET task-based measures, perhaps particularly so among close task transitions. They also suggest that the task dimension of the labor market may be more important than O*NET-based data by itself suggests. One should be careful not to draw definite negative conclusions from one comparison exercise, like the one is this section, but rather see it as opening the door for tasks to matter more than previously thought.

5 Conclusion

We have laid out a method for correcting miscoding in occupational flows by estimating the true underlying miscoding probabilities, specific to occupation pairs. With the miscoding 'garbling' matrix $\mathbf{\Gamma}$ in hand, it is straightforward to correct observed matrices of occupational flows for miscoding, converting spurious movers back into true stayers and uncovering the underlying start- and endpoints of occupational flows that were distorted by miscoding.

Our finding is that, after dealing with miscoding, occupational mobility often is more sensitive to economic fundamentals (such as age, or the business cycle) than the uncorrected data would suggest. The correction also changes the wage dynamics associated with occupational mobility of employer movers. Overall, in our analysis the importance of the task dimension of labor markets appears to increase after addressing occupational miscoding, and leaves the suggestion that miscoding may work to obscure this importance more generally in significant ways, but can be addressed by a reasonably simple 'degarbling' operation, detailed in this paper.

SUPPLEMENTAL APPENDICES

A Identifying Miscoding Probabilities

A.1 Estimation of $\mathbf{\Gamma}$

This section collects the remaining proofs that allow for consistent 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 1: *The function* $f : PDT(\mathbb{R}^{O\times O}) \to PDT(\mathbb{R}^{O\times O})$ *given by* $f(\mathbf{T}) = \mathbf{T^{0.5}}$ *is continuous with* $f(\mathbf{T_s^I}) = \mathbf{\Gamma}$ in the spectral matrix norm.

Proof. 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 $\|\mathbf{U_1} - \mathbf{U_2}\|_2 = \|\mathbf{U_1^{0.5}}(\mathbf{U_1^{0.5}} - \mathbf{U_2^{0.5}}) + (\mathbf{U_1^{0.5}} - \mathbf{U_2^{0.5}})\mathbf{U_2^{0.5}}\|_2 \geq |\mathbf{x'U_1^{0.5}}(\mathbf{U_1^{0.5}} - \mathbf{U_2^{0.5}})\mathbf{x} + \mathbf{x'}(\mathbf{U_1^{0.5}} - \mathbf{U_2^{0.5}})\mathbf{U_2^{0.5}}\mathbf{x}|$, 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 $|\mathbf{z'U_1^{0.5}}(\mathbf{U_1^{0.5}} - \mathbf{U_2^{0.5}})\mathbf{z} + \mathbf{z'}(\mathbf{U_1^{0.5}} - \mathbf{U_2^{0.5}})\mathbf{U_2^{0.5}}\mathbf{z}| \geq |\lambda|\ |\lambda_{min}^{0.5}(\mathbf{U_1}) + \lambda_{min}^{0.5}(\mathbf{U_2})| = \|\mathbf{U_1^{0.5}} - \mathbf{U_2^{0.5}}\|_2\ (\lambda_{min}^{0.5}(\mathbf{U_1}) + \lambda_{min}^{0.5}(\mathbf{U_2}))$, 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 $\hat{\mathbf{T}}_\mathbf{s}^\mathbf{I}$ denote the sample estimate of $\mathbf{T_s^I}$ and let $\hat{\mathbf{\Gamma}}$ be estimated by the root $(\hat{\mathbf{T}}_\mathbf{s}^\mathbf{I})^{0.5} \in PDT(\mathbb{R}^{O\times O})$ such that $\hat{\mathbf{\Gamma}} = (\hat{\mathbf{T}}_\mathbf{s}^\mathbf{I})^{\mathbf{0.5}} = \hat{\mathbf{P}}\hat{\mathbf{\Lambda}}^{\mathbf{0.5}}\hat{\mathbf{P}}^{-\mathbf{1}}$, where $\hat{\mathbf{\Lambda}}$ is the diagonal matrix with eigenvalues of $\hat{\mathbf{T}}_\mathbf{s}^\mathbf{I}$, $0 < \hat{\lambda_i}^{0.5} \leq 1$ and $\hat{\mathbf{P}}$ the orthogonal matrix with the associated (normalized) eigenvectors. We then have the following result.

Proposition 2: $\mathbf{\Gamma}$ *is consistently estimated from* $(\hat{\mathbf{T}}_\mathbf{s}^\mathbf{I})^{0.5} \in PDT(\mathbb{R}^{O\times O})$ *such that* $\hat{\mathbf{\Gamma}} = (\hat{\mathbf{T}}_\mathbf{s}^\mathbf{I})^{\mathbf{0.5}} =$

$\hat{\mathbf{P}}\hat{\mathbf{\Lambda}}^{0.5}\hat{\mathbf{P}}^{-1}$. *That is,* $plim_{n\to\infty}\hat{\mathbf{\Gamma}} = \mathbf{\Gamma}$.

Proof. From Lemma 1 and Proposition 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, from flow matrix $\hat{\mathbf{M}}^{\mathbf{I}}_{\mathbf{s}}$, which element by element converges to $\mathbf{M}^{\mathbf{I}}_{\mathbf{s}}$. Done the ijth element of $\hat{\mathbf{M}}^{\mathbf{I}}_{\mathbf{s}}$ by $\hat{m}_{ij}$. Transition probability $\hat{m}_{ij}/\sum_{k=1}^{O}\hat{m}_{ik}$ converges in probability to element ij of $\mathbf{T}^{\mathbf{I}}_{\mathbf{s}}$ (see Anderson and Goodman, 1957; Billingsley 1961, thm 1.1-3.) for all occupations, given assumptions *A2* and *A3*, and also transition probability $(0.5\hat{m}_{ij} + 0.5\hat{m}_{ji})/\sum_{k=1}^{O}(0.5\hat{m}_{ik} + 0.5\hat{m}_{ki})$, define this as $\hat{t}_{ij}$ and the whole matrix as $\hat{\mathbf{T}}^{\mathbf{I}}_{\mathbf{s}}$. Then $plim_{n\to\infty}\hat{\mathbf{T}}^{\mathbf{I}}_{\mathbf{s,n}} = \mathbf{T}^{\mathbf{I}}_{\mathbf{s}}$. Moreover $\hat{\mathbf{T}}^{\mathbf{I}}_{\mathbf{s}}$ is similar (in the matrix sense) to a symmetric matrix. Because $\mathbf{T}^{\mathbf{I}}_{\mathbf{s}}$ is a transition matrix that is similar (in the matrix sense) to a symmetric positive definite matrix and the space of such matrices is open, for n large enough n, $\hat{\mathbf{T}}^{\mathbf{I}}_{\mathbf{s,n}}$ is similar in the matrix sense to a symmetric positive definite matrix, i.e. $\in PDT(\mathbb{R}^{O\times O})$, assumptions *(A2)* and *(A3)* apply to it, $\hat{\mathbf{\Gamma}}_{\mathbf{n}}$ can be computed from $\mathbf{T}^{\hat{\mathbf{I}}}_{\mathbf{s,n}}$ according to $\hat{\mathbf{P}}_{\mathbf{n}}\hat{\mathbf{\Lambda}}^{\mathbf{0.5}}_{\mathbf{n}}\hat{\mathbf{P}}_{\mathbf{n}}^{\ -1}$, per Proposition 1. By continuity of the mapping in Lemma 1, it follows that $plim_{n\to\infty}\hat{\mathbf{\Gamma}} = \mathbf{\Gamma}$, and our estimator is consistent. ■

B Data appendix

B.1 Our miscoding sources in more detail

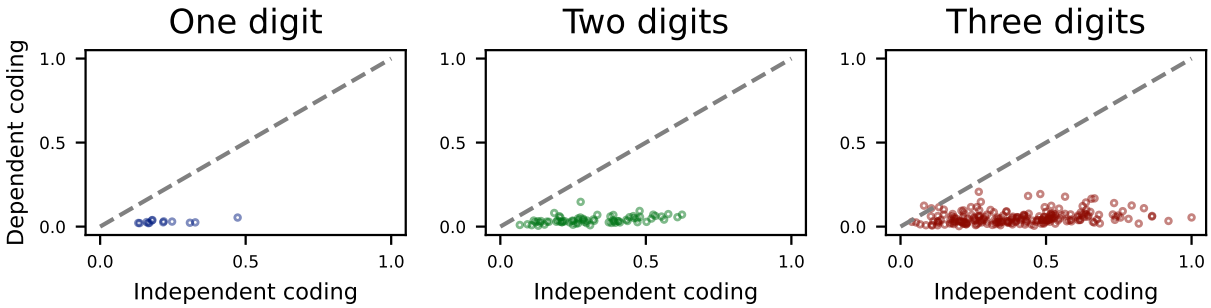
B.1.1 Survey of Income and Program Participation (SIPP) Redesign (1986)

The 1986 SIPP panel introduced dependent interviewing, contrasting with the independent interviewing used in the 1985 panel. We exploit the overlap between these panels from February 1986 to April 1987, representing the same population under different survey designs. We focus on full-time workers with a single employer, applying additional restrictions to ensure comparability.[59] The impact of dependent coding is evident in the drastic reduction of observed occupational mobility rates, from about 20% to 3% between waves.

Figure B.1 illustrates the significant impact of the introduction of dependent coding

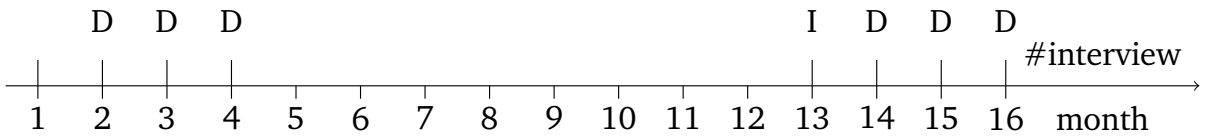
[59]This is, with small changes, the sample on which the miscoding correction used in Carrillo-Tudela and Visschers (2023c) was estimate. We include only workers who remained in full-time employment throughout two waves and reported having only one employer at any point in time. We exclude workers who experienced non-temporary layoffs with short unemployment episodes, those with imputed occupations, and those enrolled in school. Additionally, we restrict our sample to individuals between 19 and 66 years old. After applying these restrictions, we obtain 28,302 wave/individual observations for the 1985 panel, 27,801 for the 1986 panel, and 5,922 for the 1987 panel.

Figure B.1: Dependent coding reduces occupational transitions in the SIPP



We compare the monthly occupation-changing probability under independent and dependent coding. The unit of observation is an (origin) occupation. The level of aggregation for the three panels is one digit, two digits, and three digits, from left to right.

Figure B.2: Dependent and independent interviewing in the CPS since 1994



In the CPS, respondents are being interviewed for four consecutive months, rotated out for 8 months, and interviewed for 4 more consecutive months. Occupational transitions are dependently coded for months indicated with D and independently for months with I. There is no transition in the first month.

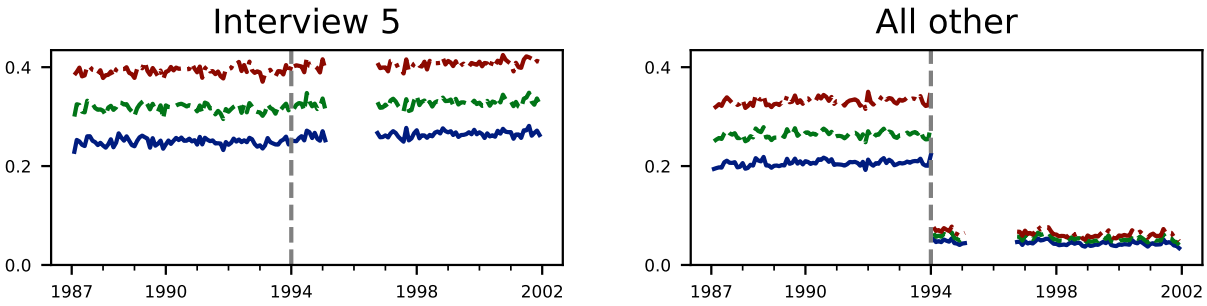
in the SIPP. The figure displays the share of occupation-changing observations for each origin occupation under both dependent and independent coding across our three levels of aggregation. Notably, the introduction of dependent coding led to a drastic reduction in the average share of occupation-changing observations, affecting virtually every origin occupation. We use the flow matrices of dependently and independently coded individuals in the two adjacent waves to estimate $\hat{\mathbf{M}}^{\mathbf{SI}}$. [60]

B.1.2 The Current Population Survey's Redesign of 1994

In the Current Population Survey (CPS) , respondents are interviewed for four consecutive months, rotated out for eight months, and then interviewed again for four more consecutive months (see Figure B.2) . Before the 1994 redesign, the CPS collected occupation and other data in each interview month independently, similar to the SIPP before its 1986 redesign. Respondents provided information about their current occupation (or past occu-

[60] To be more precise, we employ a two-step process: First, we calculate the occupation flow matrix across two adjacent waves for workers in the 1985 panel, using SIPP-provided individual sample weights. We then subtract from this the occupational flow matrix of independently interviewed individuals across two adjacent waves. This calculation yields $\hat{\mathbf{M}}^{\mathbf{SI}} = \hat{\mathbf{M}}^{\mathbf{I}}_{\mathbf{85}} - \hat{\mathbf{M}}^{\mathbf{DI}}_{\mathbf{86,87}}$. To address issues arising from finite sample size, we average the flows between occupations i and j in both directions in $\hat{\mathbf{M}}^{\mathbf{SI}}$. Finally, we set any remaining negative elements to zero, a step particularly important for more detailed classifications where the number of elements in the flow matrix is larger relative to the number of observations.

Figure B.3: Dependent coding reduces occupational transitions in the CPS



We plot the share of occupation changers among three samples in the CPS, and across three levels of aggregation (red, green, blue: three, two, one digits). Left panel: all observations. Center panel: respondents in the fifth interview. Right panel : workers experiencing an unemployment-to-employment transition. The left plot shows that the occupation-switching probability fell drastically with the introduction of dependent coding in 1994. The remaining two panels show occupation-switching probabilities for two populations that continuously are encoded using independent coding, highlighting that there was no structural change in transition probabilities for these samples.

pation in case of recent job loss) in each month, with the CPS independently encoding this information for each interview.

The 1994 redesign altered this methodology. Post redesign, respondents were interviewed dependently about occupations in interviews 2-4 and 6-8, i.e., whenever they were also interviewed in the previous month. If no change in employer or work activities/tasks was reported, the previous month's occupation was carried over. However, if a change was reported, the interviewer would ask the respondent to describe their work activities, after which a coder would assign an occupational code, similar to the pre-redesign process.

Unlike the SIPP redesign, the CPS changes were implemented simultaneously for all waves in 1994, without an overlap period of different survey designs. The utility of the CPS redesign for our analysis thus hinges on the assumption that no other structural changes in the U.S. economy significantly affected occupational mobility during this period. A key feature of the CPS design is that the fifth interview has always been independently coded, both before and after the 1994 redesign, allowing us to judge the extent to which the US economy and occupational mobility changed simulatenously with the redesign of the CPS.

Figure B.3 illustrates these points. The left panel shows occupation changes according to the fifth interview across time. The stability of these rates through the 1994 redesign suggests that underlying true occupational mobility remained relatively constant during this period and supports our use of the redesign to estimate miscoding.

The right panel of the figure reveals the extent of spurious mobility in the data. It displays a sharp drop in observed occupational mobility between interviews in adjacent

months after the 1994 redesign. The observed monthly occupational mobility for workers employed across two months decreased from over 20% to about 5% for 1-digit occupation groups, and from about 35% to about 8% at the 3-digit level. This substantial reduction indicates that a large portion of the pre-redesign occupational mobility among these workers was likely spurious, resulting from inconsistencies in independent coding rather than actual occupation changes.

The extent of spurious mobility evident in the right panel, combined with the stability of underlying mobility shown in the left panel, indicates that comparing mobility patterns immediately before and after the redesign can provide information about miscoding.[61] [62]

B.2 Aggregation of SIPP and CPS miscoding

To take into account any potential differences in estimates from the two sources, we consider the following aggregate approach.

Let $\mathbf{\Gamma}_{\mathbf{k}}^{\text{ds}}$ be the garbling matrix computed using a given data source ds (CPS, SIPP), bootstrapped iteration k. We then compute source-specific weights for occupation-pair i, j as

$$w^{\text{ds}}(i,j) = \frac{v^{\text{ds}}(i,j)}{\sum_{\text{ds}} v^{\text{ds}}(i,j)}, \text{ with } v^{\text{ds}}(i,j) = 1/Var_k(\mathbf{\Gamma}_{\mathbf{k}}^{\text{ds}}(i,j))$$

with the variance being computed over k, where the second expression normalizes the weights to sum up to one for each occupation pair. Denote by $\mathbf{M}^{\text{ds}}$ the spurious flows from a given source and episode and $\mathbf{M}^{\text{ds, norm}}$ the spurious flows where all elements of the matrix have been rescaled to sum up to 1. We then aggregate these normalized spurious flows as

$$\mathbf{M}^{\text{agg}} = \sum_{\text{ds}} w^{\text{ds}} \mathbf{M}^{\text{ds, norm}}$$

and use $\mathbf{M}^{\text{agg}}$ to obtain a $\hat{\mathbf{\Gamma}}$.

[61] By treating the period just after the redesign as informative about $\mathbf{M^{DI}}$ and the period just before as informative about $\mathbf{M^{I}}$, we can derive an implied $\mathbf{M^{SI}}$ from the changes in occupation flows associated with the 1994 CPS redesign.

[62] Given that true mobility appears to constitute a small portion of the observed pre-redesign month-to-month occupational mobility, even substantial relative shifts in true mobility between the pre- and post-redesign periods would have a limited effect on inferred spurious mobility. The stability of observed mobility in the left panel supports the use of a pre-introduction period from 1987 and a post-introduction period up to December 2002. We do not extend the window beyond December 2002 due to the change in the occupational coding in the CPS in 2003. For several months in the post-introduction period, identifiers cannot be linked. We choose selectively months from the pre-introduction window to ensure that each calendar month appears as often in the pre-treatment window as in the post-introduction window.

B.3 Comparison to other miscoding correction

Neal (1999) and Moscarini and Thomsson (2007) assess occupational miscoding by examining transitions from occupation i to j immediately followed by a return to i (we call this 'return mobility' across occupations). Moscarini and Thomsson (2007) observe a significant drop in return mobility frequencies after the 1994 CPS redesign, suggesting that most return mobility under independent coding is spurious. We expect a high correlation between our estimated miscoding probabilities and the relative frequency of return mobility.

To test this, we analyze CPS data prior to 1994, when occupations were independently coded in every interview. We focus on continuously employed workers with completed four-month interview sequences, specifically AABA, ABAA, and AAAA patterns, where A and B are potentially different occupations. We investigate whether, conditional on observing return migration starting in occupation A, it is more likely towards occupation B if A and B have a higher miscoding probability.

We estimate the relationship:

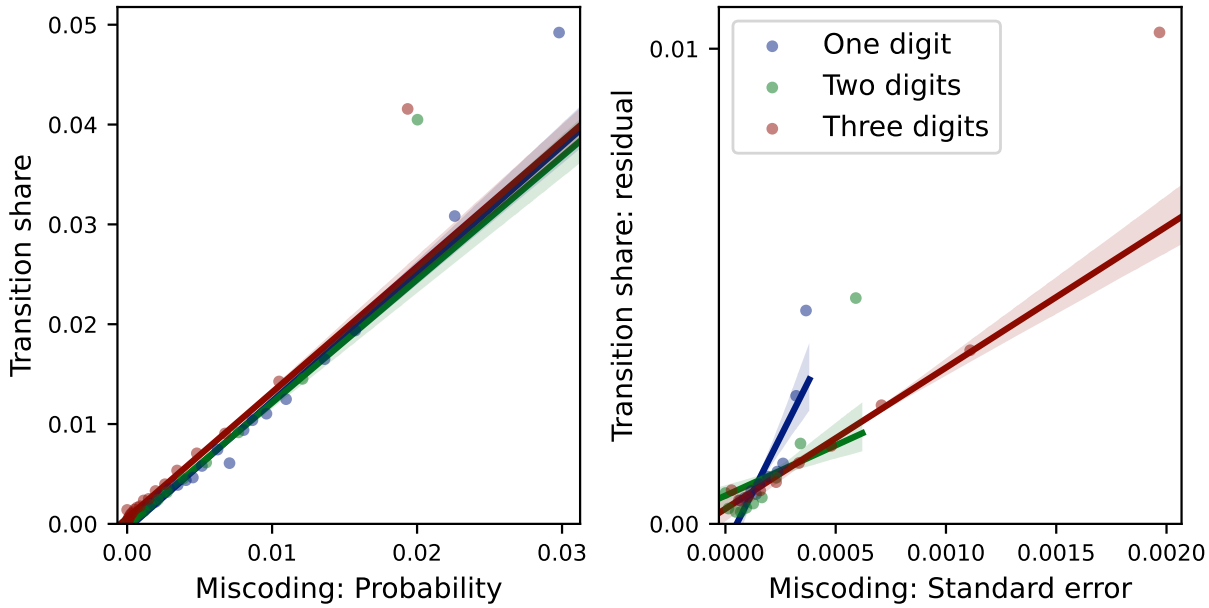
$$s_{a,b} = \mathbf{\Gamma}_{a,b} + \epsilon_{a,b}$$
$$s_{a,b} \equiv \frac{N_{a,b}}{\sum_{b'} N_{a,b'}},$$

where $N_{a \to b}$ is the number of observed 4-interview transitions for origin occupation a and temporary occupation b, and $\mathbf{\Gamma}_{a,b}$ is the miscoding probability between a and b.

Figure B.4 illustrates this relationship. The left panel shows a strong correlation between miscoding probability and observed return migration at the one-digit level, with a slope close to 1 - as the theory would have predicted when return mobility is spurious. Table B.1 shows that the fit in this relationship is remarkably high, varying between 0.91 and 0.42. Results at the three digits are quite noisy: restricting the relationship to sequences with at least 1000 observations improves the fit significantly. Conceptually, at least two factors prevent a perfect fit. First, some return migration might be actually be taking place. Second, miscoding probabilities are estimated on a finite sample, and thus with noise. To address the second point, we correlate the absolute residuals in this regression (i.e. a mistake in either direction) with bootstrapped standard errors. The right panel in Figure B.4 shows that noisier estimates lead to larger deviations (R^2 : 0.091).

Overall, the findings confirms that return migration completed within three interview

Figure B.4: Broken spells (AABA) more likely to be among high-miscoding probability pairs



Left: higher miscoding probability correlates with higher return migration. Right: Higher noise in the miscoding estimate correlates with higher residual (absolute value) in the previous regression.

months under independent interviewing is likely spurious, validating the interpretations of Neal (1999) and Moscarini and Thomsson (2007). Moreover, we show a direct linear relationship between observed return mobility frequency and miscoding probability between occupation pairs.

While intuition suggests quick return mobility is unlikely, one would either need to ask workers directly or an explicit quantitative theory of miscoding to be able to rule out this possibility: we can do the latter here, on the basis of our estimate $\mathbf{\Gamma}$. Conversely, our exercise could be taken to validate our measure of miscoding probabilities, as even on the level of occupation pairs, it produces a close relation to the observed data.

Table B.1: Miscoding and return migration for different levels of aggregation

	(1)	(2)	(3)	(4)
miscoding_prob	1.279***	1.234***	1.258***	1.232***
	(0.0192)	(0.00714)	(0.00575)	(0.00420)
Obs.	156	1487	12990	8354
R2	0.966	0.953	0.787	0.911
Aggregation	One digit	Two digits	Three digits	Three digits
Minimum Obs				1000

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Return migration is more likely among highly miscoded occupation pairs. The three columns: one-digit, two-digit, three digit aggregation.

B.4 Consistency with returns to tenure

Our analysis of the Survey of Income and Program Participation (SIPP) and the Current Population Survey (CPS) reveals that true occupational mobility is significantly lower than raw data series suggest. In models of the labor market (Kambourov and Manovskii, 2008, See for example), workers typically accumulate occupation-specific human capital which translates to higher wages. In these models, occupational changes result in the depreciation of this specific human capital, causing wage losses.

Occupational mobility in response to shocks (e.g., occupation- or industry-specific) represents a trade-off between forfeiting occupation-specific human capital and enduring a temporarily depressed labor market. To the extent that these models are calibrated using spuriously high occupational mobility rates, they underestimate the significance of occupation-specific human capital, as workers appear to relinquish it more frequently than they actually do.

Accurately quantifying the returns to occupation-specific human capital is crucial for our understanding of labor market dynamics. It informs critical policy questions: How should we approach occupation switching? Should retraining be advocated during recessions? To what extent is the labor market segmented across industries or occupations?

To complement our indirect inference argument, we conduct a Monte Carlo exercise examining the impact of miscoding on estimates of returns to occupational tenure. We simulate the occupational mobility of 1,000 workers, assuming that each worker's wage increases by xOur simulation results demonstrate that for each level of aggregation, the true estimates accurately capture the monthly returns to tenure consistent with Kambourov

and Manovskii (2008). However, the observed estimates diverge by an order of magnitude: estimating returns to occupational tenure using a miscoded dataset would lead researchers to significantly underestimate its importance!

Table B.2: Estimated return to occupational tenure on true and noisy data.

Aggregation Level	2-Year Return	Switch Prob.	Observed estimates Median	5th	95th	True estimates Median	5th	95th
One digit	3.68%	0.005	**0.001**	-0.000	0.002	**0.018**	0.018	0.018
Two digits	4.96%	0.004	**0.002**	0.001	0.003	**0.025**	0.024	0.025
Three digits	5.39%	0.005	**0.003**	0.003	0.004	**0.027**	0.027	0.027

We simulate monthly individual career patterns consistent with observed flows in the CPS and estimated 2-year returns from Kambourov and Manovskii (2008). In each month, workers draw a switching probability consistent with their occupation-specific switching rate estimated using the tenure supplement to the CPS. The average of these across all occupations is reported under Switch Prob.*. If the worker switches occupation, they draw a new occupation from the corrected transition matrix. When starting at a new job, their wage is normalized to one. It then grows monthly consistent with the 2-year returns estimated by Kambourov and Manovskii (2008). When a worker changes occupation, their wage is reset to one. We compute two monthly tenure variables: (i) true tenure based on actual mobility, and (ii) observed tenure based on observed mobility, where we miscoded the workers' occupations month-by-month according to* $\mathbf{\Gamma}$*. Finally, we regress either observed tenure or true tenure on monthly wages. We bootstrap the corresponding point estimates 100 times and report the percentiles under* Observed estimates *and* True estimates*.*

C Additional Investigations, Section 3

C.1 Aggregate distance and cyclicality

The above suggests that miscoding could also obscure some of workers' task distance dynamics over the business cycle. In Table C.1, Panel B, we present how the average task distance of occupation switchers, whether directly moving from one employer to another (EE) or through unemployment (UE), responds to cyclical conditions, before and after the miscoding correction.

Specifically, we standardize the task distances of the pooled sample of occupation movers across the entire sample period, such that the mean distance is normalized to zero, and distance 1 corresponds to one (pooled) cross-sectional standard deviation above the mean. We do this separately for the uncorrected and miscoding-corrected flows.

We then calculate the occupational flow matrices at a quarterly frequency, and correct each of these using $\mathbf{\Gamma}^{-1}$ as before. From these quarterly flow matrices, we calculate the log of the quarterly mean distance of occupation movers (given the standardization factors calculated above). After this, we bandpass-filter the quarterly time series. In Table C.1, in

Table C.1: Occupational Mobility and Task Distance over the Business Cycle

	Time Series Elasticity with Unemployment Rate (BP-Filtered)						Volatility	
	Raw Occ Mobility		Corrected Occ Mob		Difference Coeff		Raw	Corr.
	β_u	(SE)	β_c	(SE)	$\beta_c - \beta_u$	p value	(st.dev.)	(st.dev.)
	Panel A: Cyclicality of Proportion of Occupation Switchers among EE/UE Movers							
EE	-0.10	(0.02)	-0.15	(0.04)	0.05	0.004	0.020	0.042
UE	-0.10	(0.02)	-0.18	(0.05)	0.08	0.000	0.028	0.056
	Panel B: Cyclicality of Mean Task Distance of Occupation Switchers							
EE	0.05	(0.02)	0.13	(0.05)	-0.08	0.001	0.021	0.052
UE	0.07	(0.04)	0.23	(0.11)	-0.16	0.076	0.042	0.111

Panel A: Coefficients β_u, β_c of regression of log quarterly Proportion of EE, resp. UE movers, that changes occupation, on log unemployment rate, all bandpass filtered. Panel B: Coefficients β_u, β_c of regression of task-based distance traveled of EE, resp. UE movers, on log unemployment rate, all bandpass filtered. Sample: workers that change occupations. Distance is standardized simultaneously for the population of EE and UE movers, so that coefficients have the same magnitude, and the mean distances sum up to zero, appropriately weighted. We report the p-value of no difference in coefficients β_u, β_c from a seemingly unrelated regression (SUR) analysis with robust standard errors. Period 1982q1-2019q4; the EE series begin with introduction of dependent coding in 1994.

the left-most part, we focus on the time series regression of this quarterly series on the bandpass filtered log aggregate unemployment rate. In addition,in the right-most two columns report the standard deviation of the quarterly logged and detrended mean distance time series.

Comparing the miscoding-corrected vs raw data we see that mean task distance of occupational movers, both for EE and UE movers, is more strongly countercyclical after correcting for miscoding (and more volatile over the cycle, in the right-most column). The cyclical sensitivity increases by a factor 2.5-3 after miscoding correction. For EE occupation movers, the difference between the regression coefficient on corrected and uncorrected data is statistically significant at a p-value below 1%; For UE occupation movers, it is marginally statistically significant with a p-value of 7.6%.[63]

We can contrast this with the *pro*cyclicality of the proportion of occupation switchers among EE/UE movers, in panel A of Table C.1.[64] Along similar mechanics as we discussed

[63]Carrillo-Tudela, Summerfield, and Visschers (2025) also documents countercyclical increase in task distance of EE occupation switchers in Canada, but ignore miscoding and its effects.

[64]Carrillo-Tudela and Visschers (2023c) establishes the procyclicality of the UE occupational mobility proportion, using SIPP data at the major occupational group level. The look at this also using the miscoding correction described in this paper. Here, Panel A confirms their conclusion using CPS data, rather than the SIPP data that spans a smaller time window. In terms of uncorrected mobility in the CPS, Carrillo-Tudela, Hobijn, Visschers, et al. (2014) this, also for EE movers. None of these papers look at task distance, or correct observed task distance for miscoding.

with age, miscoding weakens the opposing cyclical patterns along the extensive and intensive margin of occupational/task mobility: in recessions, a lower proportion of EE and UE movers switches occupations, but those that do, cover a larger task distance. Miscoding blunts both opposing patterns, in part by creating more spurious mobility in recessions with relatively lower distances, and thereby makes it harder to gauge the mechanics of cyclical reallocation.

D Additional Tables

	Component 1	Loading 1	Component 2	Loading 2	Component 3	Loading 3	Component 4	Loading 4	Component 5	Loading 5
1	Active Learning	0.832	Mathematics Knowledge	0.856	Trunk Strength	0.882	Clerical	0.904	Technology Design	0.801
2	Systems Analysis	0.830	Troubleshooting	0.853	Explosive Strength	0.874	Administration and Management	0.875	Originality	0.790
3	Mathematics SKill	0.791	Quality Control Analysis	0.843	Extent Flexibility	0.848	Transportation	0.842	Operations Analysis	0.762
4	Information Ordering	0.788	Repairing	0.822	Gross Body Coordination	0.848	Persuasion	0.766	Problem Sensitivity	0.730
5	Operations Analysis	0.780	Operation and Control	0.803	Dynamic Strength	0.836	Management of Financial Resources	0.764	Inductive Reasoning	0.701
6	Learning Strategies	0.774	Multilimb Coordination	0.797	Gross Body Equilibrium	0.835	Management of Personnel Resources	0.732	Time Management	0.701
7	Systems Evaluation	0.773	Programming	0.785	Dynamic Flexibility	0.810	Management of Material Resources	0.721	Learning Strategies	0.698
8	Time Management	0.765	Equipment Maintenance	0.758	Response Orientation	0.803	Service Orientation	0.719	Information Ordering	0.695
9	Negotiation	0.761	Finger Dexterity	0.652	Near Vision	0.797	English Language	0.698	Active Learning	0.673
10	Speaking	0.756	Judgment and Decision Making	0.643	Static Strength	0.793	Systems Evaluation	0.692	Systems Evaluation	0.658
11	Written Expression	0.752	Installation	0.638	Stamina	0.742	Time Management	0.669	Oral Expression	0.615
12	Fluency of Ideas	0.731	Coordination	-0.609	Finger Dexterity	0.684	Systems Analysis	0.652	Management of Personnel Resources	0.610
13	Foreign Language	0.726	Design	0.604	Speaking	-0.577	Instructing	0.648	Active Listening	0.603
14	Active Listening	0.723	Control Precision	0.588	Active Listening	-0.576	Written Expression	0.643	Systems Analysis	0.580
15	Inductive Reasoning	0.715	Mathematics SKill	-0.579	Glare Sensitivity	0.565	Mathematics SKill	0.619	Number Facility	0.571
16	Written Comprehension	0.710	Dynamic Flexibility	0.573	Manual Dexterity	0.561	Active Learning	0.612	Monitoring	0.570
17	Reading Comprehension	0.709	Manual Dexterity	0.564	Foreign Language	-0.541	Operations Analysis	0.601	Building and Construction	0.552
18	Monitoring	0.708	Foreign Language	-0.560	Fluency of Ideas	-0.518	Monitoring	0.597	Fluency of Ideas	0.541
19	Writing	0.700	Response Orientation	0.560	Visualization	0.513	Reading Comprehension	0.592	Clerical	0.537
20	Originality	0.697	Writing	-0.554	Mathematics Knowledge	0.512	Coordination	0.589	Speaking	0.526
21	Service Orientation	0.696	Dynamic Strength	0.552	Wrist-Finger Speed	0.507	Written Comprehension	0.589	Flexibility of Closure	0.521
22	Instructing	0.694	Sociology and Anthropology	-0.537	Depth Perception	0.503	Originality	0.585	Social Perceptiveness	0.521
23	English Language	0.690	Explosive Strength	0.516	Operations Analysis	-0.479	Negotiation	0.583	Written Comprehension	0.519
24	Clerical	0.685	Written Expression	-0.514	Troubleshooting	0.477	Sociology and Anthropology	0.576	Administration and Management	0.508
25	Coordination	0.653	Geography	-0.509	Sales and Marketing	-0.477	Learning Strategies	0.574	Mathematics SKill	0.507

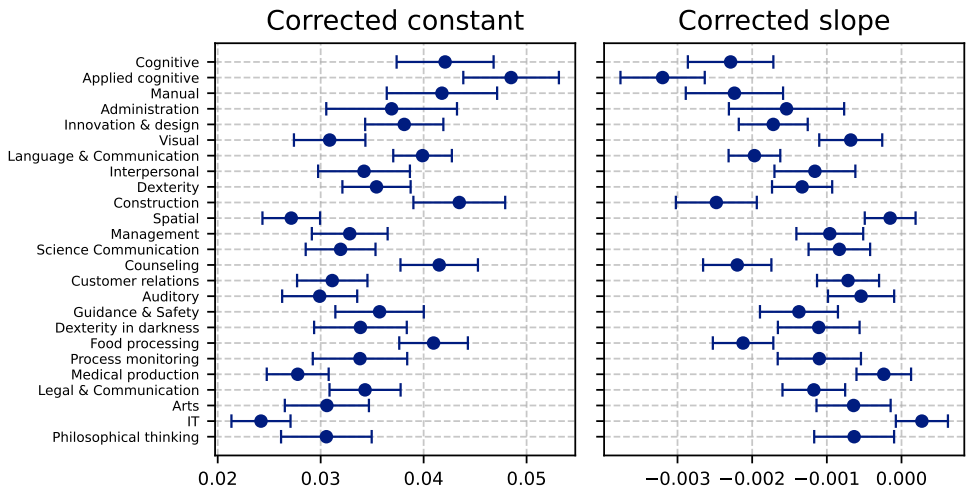
Table D.1: First five loadings of the principal components underlying the task-based distance.

Table D.2: Conditional Distributions of Observed Distances (Raw Data) – Miscoding Distortions of Distance

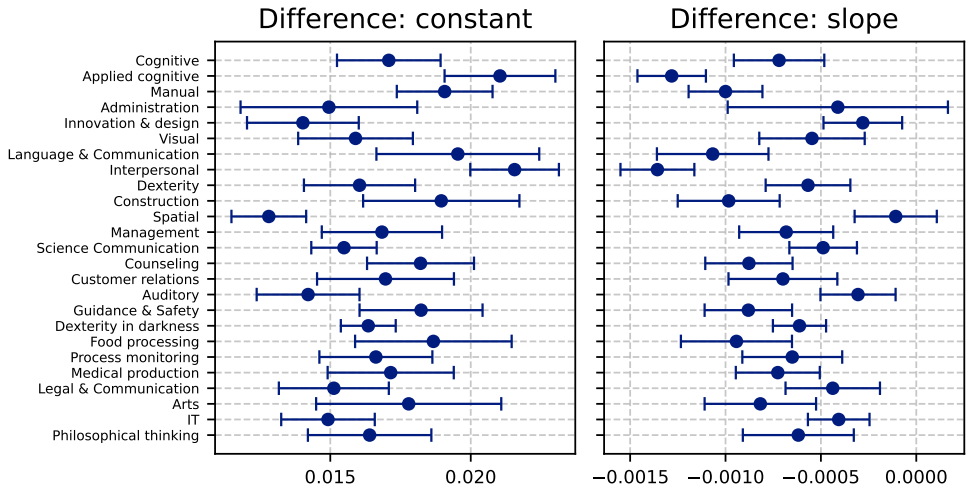
			underlying occupation distance (miscoding corrected, binned)														
		stay	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
observed distance (raw, binned)	stay	99.3	0.4	0.2	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-
	1	46.4	46.9	2.7	1.3	0.7	0.6	0.4	0.3	0.1	0.2	0.2	0.1	0.1	0.1	0.1	-
	2	38.5	5.8	48.3	2.1	1.6	1.0	0.8	0.6	0.4	0.3	0.2	0.2	0.1	0.1	0.1	-
	3	35.8	5.3	4.3	46.3	2.3	1.5	1.4	0.9	0.6	0.6	0.4	0.3	0.3	0.2	0.1	-
	4	33.4	4.5	3.8	3.2	47.4	1.9	1.4	1.2	0.9	0.7	0.5	0.4	0.3	0.2	0.1	-
	5	27.7	4.5	3.7	2.9	2.8	51.2	1.8	1.2	1.1	1.0	0.7	0.5	0.4	0.3	0.2	0.0
	6	25.4	3.9	3.0	2.6	2.6	2.0	53.8	1.7	1.1	1.1	0.8	0.6	0.7	0.5	0.2	-
	7	28.3	4.3	3.5	3.0	3.1	2.4	2.1	46.5	1.8	1.4	1.2	1.0	0.7	0.5	0.3	0.0
	8	22.5	4.5	3.3	2.8	2.8	2.3	2.5	2.3	50.8	1.7	1.3	1.2	1.0	0.6	0.3	0.0
	9	22.6	3.0	2.8	2.5	2.3	2.4	2.1	1.9	1.8	52.9	1.8	1.4	1.2	0.9	0.3	-
	10	20.2	3.3	2.4	2.0	2.0	2.0	2.0	1.9	2.1	2.5	53.9	1.9	1.6	1.3	0.5	0.2
	11	17.4	3.3	2.6	1.9	1.9	1.6	1.7	2.1	2.0	2.4	2.8	55.2	2.2	2.0	0.9	0.2
	12	21.8	3.2	2.2	1.8	1.3	1.5	1.6	1.6	1.5	2.0	2.1	2.3	53.1	2.5	1.1	0.4
	13	21.2	3.1	2.5	1.7	1.5	1.3	1.8	1.5	1.7	1.9	2.0	2.4	2.6	52.6	1.5	0.7
	14	17.5	3.3	2.4	1.5	1.4	1.2	1.3	1.9	1.3	1.8	1.9	2.4	3.2	4.1	52.7	1.9
	15	33.5	6.0	3.8	2.6	2.3	1.8	1.4	1.6	1.2	1.5	1.5	2.3	2.4	2.6	3.4	32.1

E Additional Figures

Figure E.1: Effect of Correction on Flows by Distance



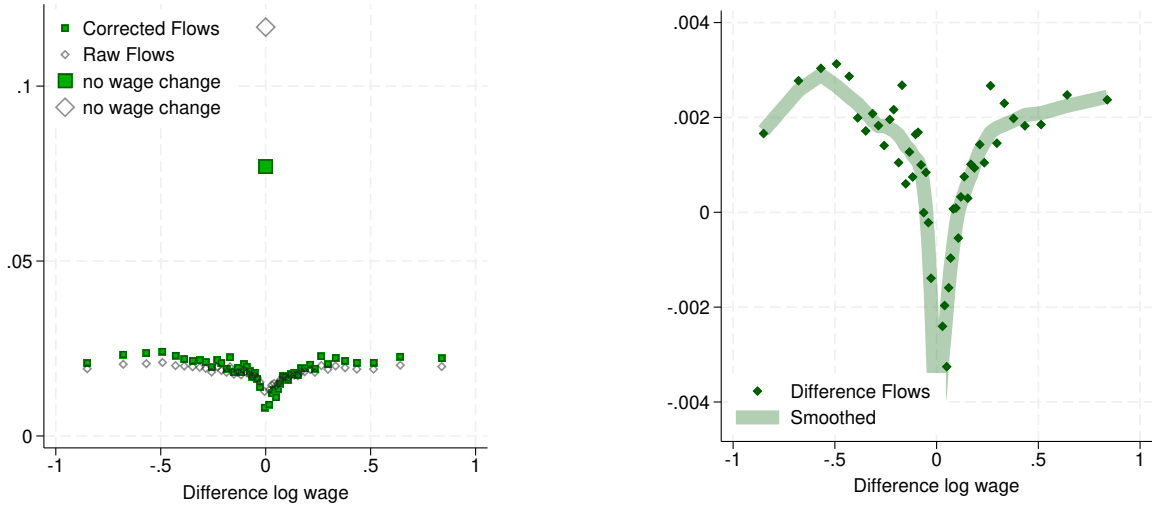
(a) Effect of Correction on Constant and Slope



(b) Constant and Slope Coefficient under Corrected Data

We regress flows on task-based distance, where the distance is computed only using a single principal component (PC). Top panel: We plot the constant and slope coefficient under the miscoding-corrected data, together with their 95% confidence bands. Bottom panel: we run a seemingly unrelated regression analysis to estimate whether the miscoding correction significantly changes the constant and slope coefficient. Confidence bands are at 95%. PCs are listed in order of importance. We find that the correction increases the constant and decreases the slope, meaning that we disproportionally reduce spurious flows for shorter distances. This effect is stronger when distance is defined according to more important PCs.

Figure E.2: Uncorrected and Corrected Flows by Wage Change

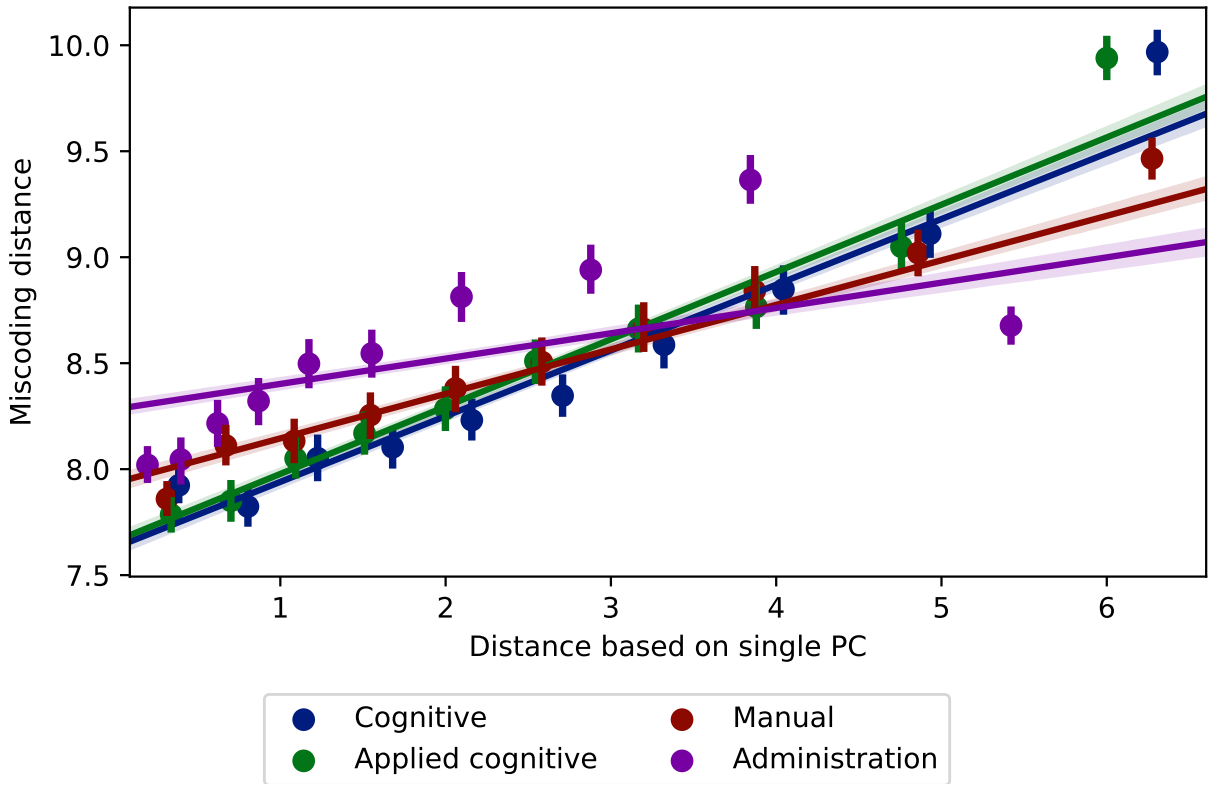


(a) Flows by wage change

(b) Effect of correction

Left panel: we plot the mass of occupation changers over wage changes of both corrected and uncorrected data. The mass of workers changing occupations with very small to no wage change is corrected downwards significantly ("no wage change" markers), suggesting that many of these are spurious flows. The second panel plots the change in flows due to the miscoding correction – excluding workers without wage changes. We can see that the correction not only lowers flows of workers with a wage change very close or exactly zero – more generally, the correction lowers observed flows with wage changes relatively close to zero, and increases those with larger absolute wage changes. For more information regarding data, see Section 3.3.

Figure E.3: Miscoding Distance and Task Principal Components



We compute a task-based distance using the top task-based principal component (PC), "Cognitive". We then scatter plot our miscoding-based distance against this single-PC distance in blue. The solid line indicates the regression coefficient. We then repeat the exercise on the second, third, and fourth principal component in red, green, and purple.

References

Abowd, J. and A. Zellner (1985). “Estimating Gross Labor-Force Flows.” In: *Journal of Business & Economic Statistics* 3.3, pp. 254–283.

Abraham, Katharine G. and James R. Spletzer (2009). “New Evidence on the Returns to Job Skills.” In: *American Economic Review* 99.2, pp. 52–57. ISSN: 0002-8282.

Baley, Isaac, Ana Figueiredo, and Robert Ulbricht (Feb. 25, 2022). “Mismatch cycles.” In: *Journal of Political Economy* 130.11, pp. 2943–2984.

Blackwell, David (1951). “Comparison of experiments.” In: *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*. Vol. 1, pp. 93–102.

Borghans, Lex, Bas Ter Weel, and Bruce A Weinberg (2006). *People people: Social capital and the labor-market outcomes of underrepresented groups*.

Carrillo-Tudela, Carlos, Saman Darougheh, and Ludo Visschers (2025). “Occupational mobility: an international comparison of US survey versus Danish register data.” Mimeo.

Carrillo-Tudela, Carlos, Bart Hobijn, Ludo Visschers, et al. (2014). “Career changes decline during recessions.” In: *FRBSF Economic Letters* 2014-09.

Carrillo-Tudela, Carlos, Fraser Summerfield, and Ludo Visschers (2025). “Workers’ Task and Employer Mobility over the Business Cycle.” In.

Carrillo-Tudela, Carlos and Ludo Visschers (2023a). *Supplement to "Unemployment and Endogenous Reallocation over the Business Cycle"*. Econometrica Supplemental Material. (Econometrica, Vol. 91, No. 3, May 2023, 1119–1153).

— (2023b). *Unemployment and Endogenous Reallocation over the Business Cycle*. Tech. rep. arXiv: 2304.00544.

— (2023c). “Unemployment and endogenous reallocation over the business cycle.” In: *Econometrica* 91.3, pp. 1119–1153.

Cortes, Guido Matias and Giovanni Gallipoli (2018). "The costs of occupational mobility: An aggregate analysis." In: *Journal of the European Economic Association* 16.2, pp. 275–315.

Dvorkin, Maximiliano (2025). "International trade and labor reallocation: misclassification errors, mobility, and switching costs." In: *Review of Economics and Statistics*, pp. 1–45.

Feng, Shuaizhang and Yingyao Hu (2013). "Misclassification errors and the underestimation of the US unemployment rate." In: *American Economic Review* 103.2, pp. 1054–1070.

Flood, Sarah et al. (2023). *IPUMS CPS: Version 11.0 [dataset]*. Minneapolis, MN.

Fredriksson, Peter, Lena Hensvik, and Oskar Nordström Skans (2018). "Mismatch of Talent: Evidence on Match Quality, Entry Wages, and Job Mobility." In: *American Economic Review* 108.11, pp. 3303–3338.

Gathmann, Christina and Uta Schönberg (Jan. 2010). "How General Is Human Capital? A Task-Based Approach." In: *Journal of Labor Economics* 28.1, pp. 1–49. ISSN: 0734-306X.

Guvenen, Fatih et al. (2020). "Multidimensional skill mismatch." In: *American Economic Journal: Macroeconomics* 12.1, pp. 210–44.

Kambourov, Gueorgui and Iourii Manovskii (2008). "Rising Occupational and Industry Mobility in the United States: 1968-97." In: *International Economic Review* 49.1, pp. 41–79. ISSN: 00206598, 14682354.

— (Feb. 2009). "Occupational Specificity of Human Capital." In: *International Economic Review* 50.1, pp. 63–115. ISSN: 00206598.

— (2013). "A Cautionary Note on Using (March) Current Population Survey and Panel Study of Income Dynamics Data To Study Worker Mobility." In: *Macroeconomic Dynamics* 17.01, pp. 172–194. ISSN: 1365-1005.

Keane, Michael P and Kenneth I Wolpin (2001). "The Effect of Parental Transfers and Borrowing Constraints on Educational Attainment." In: *International Economic Review* 42.4, pp. 1051–1103.

Lise, Jeremy and Fabien Postel-Vinay (2020). “Multidimensional Skills, Sorting, and Human Capital Accumulation.” In: *American Economic Review* 110.8, pp. 2328–2376. ISSN: 0002-8282.

Marschak, Jacob and Koichi Miyasawa (1968). “Economic comparability of information systems.” In: *International Economic Review* 9.2, pp. 137–174.

Mathiowetz, N. (1992). “Errors in Reports of Occupations.” In: *Public Opinion Quarterly* 56, pp. 332–135.

Mellow, Wesley and Hal Sider (1983). “Accuracy of response in labor market surveys: Evidence and implications.” In: *Journal of Labor Economics* 1.4, pp. 331–344.

Moscarini, Giuseppe and Kaj Thomsson (2007). “Occupational and Job Mobility in the US.” In: *The Scandinavian Journal of Economics* 109.4, pp. 807–836. ISSN: 03470520, 14679442.

Neal, Derek (1999). “The Complexity of Job Mobility among Young Men.” In: *Journal of Labor Economics* 17.2, pp. 237–261. ISSN: 0734-306X.

Poletaev, Maxim and Chris Robinson (2008). “Human Capital Specificity: Evidencefrom the Dictionary of OccupationalTitles and Displaced Worker Surveys,1984–2000.” In: *Journal of Labour Economics* 26.3.

Poterba, J. and L. Summers (1986). “Reporting Errors and Labor Market Dynamics.” In: *Econometrica* 54.6, pp. 1319–1338.

Roys, N. and C. Taber (2017). “Skill Prices, Occupations and Changes in the Wage Structure.” Mimeo, Department of Economics, Royal Holloway, University of London.

Speer, Jamin D. (2016). “How bad is occupational coding error? A task-based approach.” In: *Economics Letters* 141, pp. 166–168. ISSN: 0165-1765.

Sullivan, P. (2009). “Estimation of an Occupational Choice Model when Occupations are Misclassified.” In: *Journal of Human Resources* 44.2, pp. 495–535.

Vom Lehn, Christian, Cache Ellsworth, and Zachary Kroff (2022). “Reconciling occupational mobility in the current population survey.” In: *Journal of Labor Economics* 40.4, pp. 1005–1051.