Learning from Mistakes: Occupation Miscoding and Task Distance

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

Saman Darougheh
Danmarks Nationalbank,

Ludo Visschers
The University of Edinburgh,
U. Carlos III de Madrid, CESifo, IZA

February 2024
Preliminary, please do not cite or distribute without permission

Abstract

We use the redesign of the Current Population Survey in 1994 to estimate occupational miscoding. We then build a measure of occupational distance on this miscoding propensity (frequent miscoding between occupation pairs implying a low distance). This measure contains different information than the conventional task-based measures. We investigate its properties relative to those measures. While miscoding and conventional measures correlate positively, they also disagree often about the distance of a given occupational pair. This appears to leave a scope to learn from occupational coding mistakes, adding information to conventional measures, to observe the geography of the occupational space more clearly in the labour market. This can have implications for the rationalization of labour market flows, and the returns we observe (to occupational tenure, e.g.).

1 Introduction

In the most recent decades, attention has become increasingly focused on the role of workers' types of tasks as fundamental for our understanding of the labour market and how it evolves. A comparison of portfolios of usual tasks allows the construction of continuous distance measures. Changes of task portfolios with larger distances are associated with larger changes in economic outcomes, such as wages or unemployment duration (e.g. Gathmann and Schönberg, 2010). Sticking with similar task portfolios, on the other hand, may allow the worker to accumulate task-specific human capital (Lise and Postel-Vinay, 2020;

Guvenen et al., 2020, others) and shape the choices and payoffs which a worker face in the rest of his working life.

It is important that these distances are measured accurately: noisy measurement of distances between jobs could e.g. rather downplay the role of tasks and task differentials shaping economic outcomes. In practice, the measurement of task distances between two jobs typically faces two bottlenecks: (1) worker surveys typically report an occupation to summarize the worker's task portfolio, which is known to occur with mistakes; (2) when translating occupations back into tasks or task dimensions, this is based on information that is necessarily summarized and compressed. [1] In this paper, we propose a new source of information for distance measurement: the propensity for an occupation to be miscoded into another occupation.

This has two advantages: (1) we can incorporate the uncertainty about a worker's true occupation when we see his reported occupation; (2) miscoding yields an intuitive distance notion by itself, that is based on the (dis)similarity of open-text responses across occupations. With this in mind, we want to improve distance measurement between task portfolios, and re-evaluate the importance of tasks and occupational tenure on earnings outcomes.

Miscoding-based distance adds another type of occupational distance measurement to the literature, in addition to two types of measures. First, those based on job/occupations definitions that are then quantified and summarized along a number of dimensions (in the US e.g. using O*NET, or beforehand, the Dictionary of Occupational Titles). Second, distance measures that involve the sizes of observed worker flows between occupation, in the spirit of Nimczik (2017). Miscoding is different in that e.g. we, as researchers, can remain completely agnostic on which elements of descriptions are more important than others (we don't have to look at actual activity descriptions at any stage), while still we do not have to take into account equilibrium behavior of workers (worker flows may respond to economic incentives different from distance alone). On the other hand, we also share something with both approaches: we use flow information to estimate distance, while the fundamental information that gives rise to our measure is a description of a worker's usual activities.

To understand the intuition of the link behind miscoding and distance, consider how occupation codes are assigned in surveys like the CPS, SIPP, PSID, UKLHS and many others. First, the respondent is typically asked to describe his usual work activities (or directly, his occupation), which is written down by the interviewer.[2] This open response is then taken to a professional occupation coder, who will consider which coded occupation fits best the description of the respondent. At each step there is room for ambiguity and

[1] Moscarini and Thomsson (2007) document the effects of these miscodings on estimated occupational transitions and propose an approach to reduce their impact.

[2] Presumably in response to otherwise too-generic occupation answers, CPS interviewers are pressed to inquire further for specifics regarding one's usual activities/occupation.

randomness, which leads to randomness about which occupation code is assigned (see e.g. Kambourov and Manovskii (2008), and Carrillo-Tudela and Visschers (2023)). The premise of our miscoding distance measure is that if two occupations are similar in its associated activities, then survey descriptions of these two occupations will more frequently lead to mistakenly associate one occupation to a description of the other (true) occupation (and vice versa).

This said, a key and necessary ingredient of using miscoding for these purposes, is a good way to estimate it in the data. Fortunately, we have a way of inferring it from the data, by using the redesign of the CPS in 1994.[3] Before the 1994 redesign in the CPS, the occupation description was solicited and coded each interview (independent coding). Afterwards, the worker would be asked if his employment or work activities had changed, and only if so, the occupation description was solicited and coded anew. Assuming that true occupational mobility is similar just before and after the redesign, this allows us to isolate the observed occupation transition matrix of the subset of workers who is neither changing employer or work activities. By implication, this subset of workers contains only true occupational stayers across adjacent months, yet we observe that they coded into one occupation in the first month, another in the second month. Under a set of reasonable assumptions, this allows us to identify the propensity that a worker who is (truly) in occupation A is observed in occupation B (and vice versa). We then use these probabilities to define a distance measure: the less likely miscoding, the more distinct occupations is understood to be, and the larger the distance between them is defined to be.

Note that, to be clear, the measurement of miscoding distance is itself subject to measurement error, more so at the level of detailed occupations. One of the goals of this paper is to evaluate what we can learn from taking into account miscoding, and

In this version, we first compare our miscoding measure to more conventional task distances. We find that our miscoding measure is positively correlated with conventional task distances: on average, miscoding-based and task-based distances agree. This average correlation masks significant heterogeneity: very often these two types of distance measures do *not* agree. This raises the question when the relative advantages of one measure dominate the other measure, but also whether the best of both approaches can be combined. In this version, we test both classes of distance measures using economic theory using wage growth regressions of job-stayers in the NLSY (see also Gathmann and Schönberg, 2010; Guvenen et al., 2015; Macaluso, 2017; Baley, Figueiredo, and Ulbricht, 2018).

The remainder of this paper is as follows. In section 2, we introduce the data used in this paper. In particular, we provide details on how the CPS measures occupations before and

[3] In Carrillo-Tudela and Visschers (2023), we use the redesign of the SIPP after the 1985 panel similarly to estimate miscoding at the major occupational group (coarsest) level of the occupation classification.

after 1994, and how the redesign in 1994 reduced the number of miscodings in occupational transitions. We describe in section 3 how we estimate occupational distance based on these miscodings. Section 4 reports our findings and compares our estimated miscoding-based distance with task-based distances. Section 5 concludes.

2 Measurement of Miscoding

In this section, we use the CPS and the SIPP to establish the degree of miscoding that takes between occupation pairs. That is the probabilities that an occupation o instead appears as occupation o' in the survey in question.

There a number of ways that we can distill this information from these surveys, comparing independently and dependently coded occupational information.[4] The CPS actually provides a number of different ways to implement this: (1) by using the redesign of the CPS in 1994; (2) by looking at a worker's occupation at interview 5, compare it to the occupation at interview 1, and use the job tenure supplement to isolate those observations that report no change of activities; (3) by inferring how many workers would be changing activity or firm between interview 4 and 5; by (4) by looking at the those recalled after being on layoff; (5) by looking at the SIPP redesign between the 1985 and 1986 panels.

2.1 CPS Redesign in 1994

We use the CPS redesign in 1994 to estimate our miscoding-based measure of occupational distance. In order to compare our distance measure to conventional measures, we use the O*NET database to build several task-based measures of occupational distance. We use the NLSY74 to show that all measures of distance perform similarly in predicting wage growth following job-to-job transitions. Finally, we use the CPS from its redesign in 1994 to the current date in order to estimate miscoding using recall hires.

2.2 Occupational groups

We need a set of occupations that is consistent over time and comparable across all our data sources. The IPUMS project provides the *occ1990* variable, an occupational code that harmonizes occupational groups over time, centered around the year 1990. This variable exists for our the census-based surveys. We follow **Guvenen2020** in aggregating *occ1990* in a manner that allows us to link it to the O*NET data.[5]

[4] (In)Dependent coding definitions.

[5] We refer to Appendix **??** for details.

2.3 CPS

We measure workers' occupational transitions using the Current Population Survey (CPS). One of the purposes of the CPS is to provide information on the United States labor force; individuals are surveyed 8 months about their employment status and labor market characteristics. To this end, employed workers are asked to describe their current occupation: "*What kind of work do you do; that is, what is your occupation?*". The pollster will then note down the information provided. Census employees in Jeffersonville use the information provided to assign the worker to an occupation. Transcribing the provided information to an occupation is often difficult. The CPS handbook reminds the pollster to elicit more information if necessary: "Information about usual activities or duties is very important for assigning an accurate occupation code. This information permits more accurate coding of occupation, especially when a simple job title does not provide enough information to code.".

The CPS underwent a methodological redesign in the year 1994. Previously, employed individuals were asked in each month to provide information on their current occupation, and a CPS employee had to encode the information provided in each months independently. Mistakes in the provision of information or its encoding opened up the possibility that workers had been assigned different occupations in consecutive months even if their actual occupation had not changed. Consequently, an incorrect occupational coding in either the previous or the current month could lead researchers to estimate an occupational transition where none had happened.

To address this problem, the CPS redesign in 1994 introduced dependent coding, where some of the questions were dependent on previous responses of the survey participant. A participant that had been surveyed in the previous month was first asked whether they still had the same employer as in the previous month. If they answered yes, they were asked if they still had the same activities as in the previous month. If they answered yes, the occupation reported in the previous month was assigned to the participant for the current month.

Details In order to compute occupational transitions on the CPS, we use information from all employed workers in the years 1993 and 1994. Figure 1 shows that the number of observed occupational transitions fell from 15,000 per month to less than 5,000 after the redesign. We argue that this fall is not due to a change in actual occupational transitions. Instead, the high number of transitions prior to 1994 is due to miscoding of occupations which was reduced by dependent coding. We will use the fall in observed transitions to compute the distance between occupations. The number of transitions are stable between

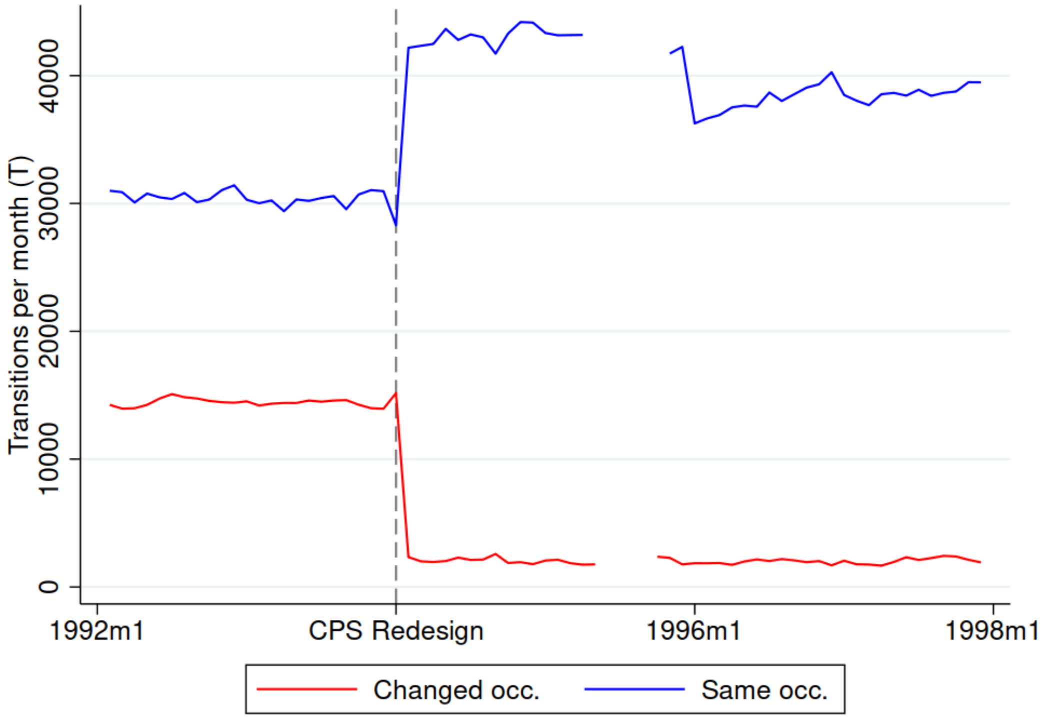

Figure 1: Dependent coding reduces occupational transitions

1993 and June 1995 – except for the sharp decline with the methodological change.[6] This allows us to extend the before period to the January 1994, and the after period up to May 1995, in order to estimate the occupational distances more precisely.

3 Occupational distance

In this section, we first introduce our miscoding-based distance measure. We will validate it by comparing it against conventional task-based measures which we define in section 3.2.

3.1 A miscoding-based distance measure

We use observed occupational transitions in the CPS to estimate the occupational distance. For this, we proceed in three steps: First, we identify "false transitions" in the CPS – transitions that are purely due to the methodological change in 1994. Second, we use these transitions to estimate the probability that a given observation's occupation is miscoded. Finally, we translate the miscoding probabilities to a set of occupational distances.

[6] Household records cannot be linked for May - October 1995, making it impossible to compute occupational transitions for that period.

Identification of false transitions Denote the matrix with the *number* of observed worker transitions from occupation o to o' in adjacent months in the basic CPS, during a given time window t, as $F(o, o', t)$. Call $F^d(o, o', t)$ the matrix with the number of observed worker transitions after dependent interviewing is triggered. We compute the miscoding between occupations o and o' as $f(o, o')$:

$$f(o, o') = F(o, o', 1993) - \frac{N(1994)}{N(1993)} F^d(o, o', 1994),$$
$$N(t) \equiv \sum_{o,o'} F(o, o', t).$$

Intuitively, if true mobility flows have stayed constant across the time windows, when removing $F^d(.)$, we removed all true flows. Even if $F^d(.)$ includes some miscoded true stayers, this does not matter: what matters is that the workers in $f(o, o')$ are all true stayers, and the observed flows are due to miscoding.

From transitions to miscoding First, we use that (by assumption) miscodings are theoretically symmetric by averaging them out:

$$\overline{f}(o, o') = \frac{f(o, o') + f(o', o)}{2}$$

We then construct the transition matrix G associated with flow matrix f, where each element o, o' in G corresponds to $\overline{f}(o, o')/\sum_{\tilde{o}} f(o, \tilde{o})$. Then, by proposition 1 (assuming the assumptions hold) in Carrillo-Tudela and Visschers (2023), we can take the matrix square root to arrive at the matrix that yields the probability of that true occupation o instead is coded as o':

$$\Gamma = G^{\frac{1}{2}}. \tag{1}$$

Compute distance from miscoding We finally define the pseudo-distance metric between two occupations o and o', $d(i, j)$, as:

$$d(i, j) = \sqrt{\frac{\Gamma(i, i) + \Gamma(j, j)}{\Gamma(i, j) + \Gamma(j, i))}} - 1$$

This captures that requirements that distance is symmetric, the distance of one occupation to itself is 0. As miscoding becomes more likely between two occupations, the distance becomes closer to zero.

3.2 Task-based distances

The skill content of each occupation can be characterized by an N-dimensional vector, $q(o) = (q_{o,1} \ldots, q_{o,N})$. We then define three distance measures: The Euklidian, Manhattan, and Angular distance are denoted as d^e, d^m, and d^a and are given by (2) - (4).

$$d^e(o,o') = \left(\sum_{i=1}^{N}(q_{o,i} - q^2_{o',i}\right)^{1/2} \tag{2}$$

$$d^m(o,o') = \sum_{i=1}^{N}|q_{o,i} - q_{o',i}| \tag{3}$$

$$d^a(o,o') = \cos^{-1}\left(\frac{\sum_{i=1}^{N}(q_{o,i} \times q_{o',i})}{\left(\sum_{i=1}^{N} q^2_{o,i}\right)^{1/2} \times \left(\sum_{i=1}^{N} q^2_{o',i}\right)^{1/2}}\right) \tag{4}$$

4 Applications

In this section, we first describe our miscoding-based distance measure and compare it against the conventional task-based measures.

Then, in order to validate the measure, we show in section **??** that it makes predictions regarding wage growth that are in line with economic theory.

Finally, we show the changes in occupational distance throughout the past 30 years.

4.1 Descriptive differences in distance meausures

Throughout this section, we will compare – in the cross section – the miscoding-based distance with the task-based measures of distance. Unlike the task-based measures, the miscoding-based measure will suffer from measurement error. This is because we use the changes in transitions across occupation pairs around 1994 to identify miscoding. For occupations with very few observations, the law of large numbers will not hold, and the occupational distance will be estimated with noise. This measurement error will affect the unweighted comparisons of the measures – where we give each occupation-pair a weight of one. The error will be less relevant in economic applications, since it only appears in occupations with few observations.

Figure 2 compares the cross-sectional distribution of the miscoding-based distance with the task-based measures. The right plot displays the standardized the task-based measures: they have a mean of zero by construction. They also appear quite symmetric. In contrast, the miscoding-based distance is skewed to the right. In red, we exclude, we exclude occupation pairs with less than 50 observations. When we focus on occupation-pairs where the miscoding is estimated based on a larger number of transitions, the distribution shifts to the left: noisiness in the estimates leads - on average - to upwards-biased miscoding distances.

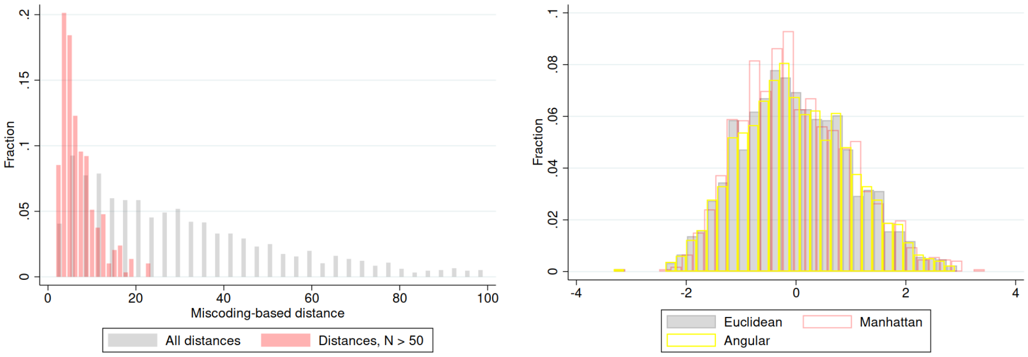

Figure 2: Distribution of miscoding-based and task-based distances

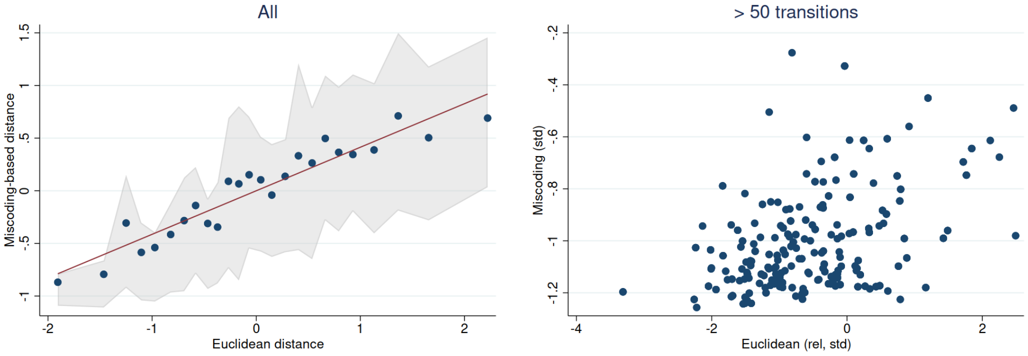

Figure 3: Miscoding distance correlates well with task-based distance

Figure 3 plots a binned scatterplot between the standardized miscoding-based distance and one of the task-based measures. The miscoding-based distance is more dispersed, but the two measures are strongly positively correlated. Yet, there is significant dispersion around the average correlation: the gray-shaded area displays the 25%-75% confidence bands, which are quite wide: occupation-pairs with a zero (standardized) distance in the task-based measure have on average a zero (standardized) distance in the miscoding-based measure, but their 25% confidence bands ranges from -0.5 to $+0.5$.

The second plot in Figure 3 plots the scatterplot, restricting to occupation-pairs with at least 50 transitions in the before period $(T(o, o', 1994) > 50)$. Presumably, the miscoding-based distance for these occupation pairs is estimated with less noise. Yet, there is still significant dispersion left in these occupation pairs across the two measures of distance. The two measures can disagree for four reasons.[7]

Small miscoding distance is correct A small miscoding-based distance is correct, and the large task-based distance is incorrect. This may arise when a worker's "true" occupation is a combination of the two occupational codes a and b. This worker would perform tasks related to both a and b. In different interviewing months, the worker would then be assigned either a or b. If a significant number of workers are in such positions, we would estimate a high likelihood of miscoding, and a small miscoding-based distance. The task-based distance would disagree if a and b were not close in the task-space. Examples are the occupational pairings $\{$Janitor, Personal service$\}$ and $\{$Janitor, Manager$\}$.

Large miscoding distance is correct A large miscoding-based distance is correct for an occupational pair where the task-based distance is small. This would occur when two occupations are actually distant: the actual tasks corresponding to both occupations are very distant. However, the task-based measures are computed based on a reduced task space: O*NET describes each occupation in a 120-dimensional task space. Depending on how this reduction is done, occupations can appear close in this reduced task-space even if they are distant in the actual task space. An example is the occupational pair Nurse attendants and Shipping and receiving clerks. These appear close in the task-space only since both involve managerial tasks.

Large miscoding distance is incorrect A large miscoding-based distance is incorrect. This often occurs when there are clear labels that reduce miscoding disproportionately. For example, nurses and physicians share many tasks and are close on the task-based space.

[7] For clarity of exposition, we will call the miscoding-based distance as "correct" if it is closer to the actual underlying distance.

They have however widely known and recognizable occupation names – when survey respondends are asked about their occupation, they need not describe it, but rather just respond with the actual occupation names. This disproportionately reduces the miscoding in these occupations, and makes it such that we estimate them to be very distant from each other, when in reality they are not

Small miscoding distance is incorrect There may be cases where occupations are incorrectly assigned a small miscoding-based distance. However, we could not find any examples for this case.

4.2 Wage growth

Is our miscoding-based distance measure economically meaningful? In this section, we follow Gathmann and Schönberg (2010) in using occupational distance to predict earnings growth following a job-to-job transition. Occupational pairs with smaller distance have more transferable human capital. Workers' wages in the new job increases with the wages in the old job for multiple reasons: the wage in the old job can be used as an outside option to negotiate a higher wage. It could also be indicative of a more productive worker, who thus should also – ceteris paribus – expect a higher wage in the new job. However, if the job-to-job transition is across a further distant occupational pair, less of that old productivity is transferable to the new job. Consequently, economic theory predicts that the previous wage and the occupational distance of the transition *jointly* have a negative effect on wages in the new job. We test this hypothesis using our various measures of occupational distance, and show that our miscoding-based measure performs similarly to conventional task-based measures.

For this, we standardize our distance measures and combine them with the NLSY. We then estimate (5), where $w_{i,t}$ refers to individual i's log wage in month t, $o(i,t)$ is the occupation that individual i is in at time t, and $D_{o,o'}$ refers to the distance between occupations o, o' according to some distance measure.

$$w_{i,t} = \alpha w_{i,t-1} + \beta w_{i,t-1} D_{o(i,t),o(i,t-1)} + \gamma D_{o(i,t),o(i,t-1)} + T_t + O_{o(i,t)} + \epsilon_{i,t} \qquad (5)$$

For this, we use all the job-to-job transitions in our NLSY data that coincide with a change in occupation. Table 1 compares the performance of our miscoding measure with the Euclidean measure. Columns 1 and 2 show the basic regressions, where the miscoding measure slightly outperforms in coefficient size, standard error, and R-squared. Next, we test to what extent mismeasurement of miscoding affects our results. We exclude occupational pairs with a bootstrapped coefficient of variation above one.[8] We find that this improves

[8] To compute the bootstrapped coefficient of variation, we first bootstrap standard errors for the miscoding-

	(1) Miscoding	(2) Euclidan	(3) Miscoding	(4) Euclidan	(5) Miscoding	(6) Euclidean
L.wage	0.272***	0.278***	0.313***	0.317***	0.278***	0.375***
	(9.40)	(9.27)	(8.70)	(8.70)	(7.42)	(9.76)
L.Wage x Dist	-0.0933***	-0.0763**	-0.0817**	-0.0878**	-0.141***	-0.0560
	(-3.62)	(-2.86)	(-2.98)	(-2.74)	(-4.31)	(-1.79)
Dist	0.624***	0.497**	0.545**	0.579**	0.931***	0.350
	(3.60)	(2.74)	(2.96)	(2.69)	(4.18)	(1.65)
N	4007	4007	2735	2735	3926	3926
R^2	0.322	0.320	0.350	0.350	0.451	0.451

t statistics in parentheses

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

Table 1: Miscoding-based and task-based measure perform similarly in job-to-job wage growth

the R-squared of both the miscoding-based and the task-based measure, and that the coefficient size and standard errors become more similar across the two. Finally, we run the main specification for the miscoding measure, but weight each observation with the inverse of the bootstrapped standard error of the miscoding distance. This significantly increases the coefficient, standard error and R-squared. This may be for two reasons: (i) when better-estimated, the miscoding-based distance performs better in the wage regressions, or (ii), the miscoding-based distance happens to be better estimated for occupational pairs where the occupational distance matters more for wage growth. To test the second hypothesis, we repeat the same regression on the Euclidean distance, using the same sample and same observation weights. We find that the R-squared is similar, but the coefficient becomes much smaller and statistically insignificant[9].

This exercise allows us to conclude that there is economic content in the miscoding-based distance, and that in the above exercise it performs similarly to the Euclidean distance measure.

5 Value of distances in reemployment

This begs the question of how the two distances would perform in a setting where they would actually be used: when unemployed workers (potentially with the help of case workers) consider which jobs to apply for.

based distance measure, and then divide these by the absolute value of the mean.

[9] We perform a battery of robustness tests with our various specifications task-based occupational distance. The results are in Appendix A.

Consider a situation in which case workers want to recommend occupations to unemployed workers that are 'close'. To get a sense whether and to which extent the closest occupations, as observed using the miscoding vs. Euclidean measure, lead to lower wage losses, we turn to the SIPP. This data is well-suited for this exercise since it covers long panels of workers together with monthly wage information, allowing us to estimate the wage loss associated with each occupational change through a E-U-E transition. We restrict our sample to workers of working age. Due to the availability of wages in early years, and the later redesign of the SIPP, we restrict ourselves to the years 1995 to 2013.

For each occupation, we find the closest x *other* occupations (i.e. excluding occupation-stayers) according to each distance measure. The dummy $\mathbf{1}_{\text{close occ}(o,o',d,x)}$ is encoded to 1 when o' is within the set of close occupations of origin occupation o. Here, d denotes which distance measure was used for the estimation, and x denotes the number of occupations that are being coded as "close". We only keep origin-occupations in our data where both distances are defined, and where we have both close and non-close destination occupations. For each worker, we compute unemployment spells s that have a preceding and subsequent employment observation. For each spell, we compute as w^{previous} the last observed (log) wage from the previous spell, and as $w^{\text{reemployment}}$ the observed reemployment wage.

For each worker i and unemployment spell s, we can then estimate

$$w_{i,s}^{\text{reemployment}} = \alpha(d,x)\mathbf{1}_{\text{close occ}(o(i,s),o'(i,s),d,x)}\beta w_{i,s}^{\text{previous}} + X_{i,s} + \epsilon_{i,s}. \tag{6}$$

That is, $\alpha(d,x)$ contains the average (percentage) wage gain associated with finding reemployment in a close occupation[10], when "closeness" is defined according to distance measure d and with a threshold of x closest occupations. We estimate (6) using both distance measures, and for up to the closest 6 occupations.

Figure 4 contains the estimated coefficients. The figure shows that the miscoding distance outperforms the euclidean distance. The difference is larger when we only consider few occupations as close.

[10] While this is naturally subject to endogeneity issues, we are working on estimating a quasi-random variation of this setting.

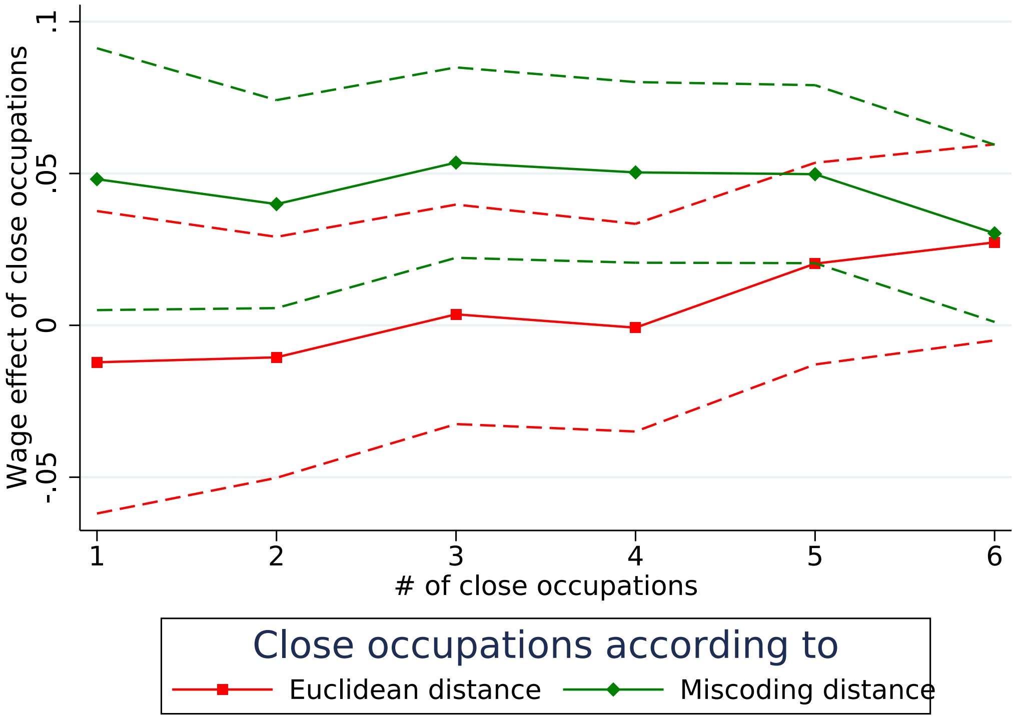

Figure 4: Miscoding distance outperforms euclidean distance in reemployment wage growth

6 Conclusion

We have constructed a measure of occupational distance using miscoding in the CPS. This measure correlates positively with conventional task-based measures of distance. Yet, these two types of measure often disagree. This is because these two measures use different sources of information with different (dis)advantages. When studying wage growth in job-to-job transitions, both types of distance measures perform in line with economic theory. Both measures perform quantitatively similar: when looking at coefficient size, precision, and r-squared, both types of measures have similar predictions. This is despite both measures often disagreeing on the distance for a given occupation pair.

Our findings suggest that both measures on their own are noisy estimates of the underlying occupational distance. Imprecise estimates of occupational distance could lead researchers to underestimate the importance of occupation-specific human capital as a driver of wage growth, and the importance of occupational tenure in career decisions. Combining the different sources of information in the two types of distance measures could lead to improved measure of occupational distance, and better estimates of the returns to occupation-specific human capital (work in progress).

A Additional Data

A.1 O*NET

The US Department of Labor's O*NET project characterizes each occupation by the mix of ability, knowledge and skills that it requires. O*NET provides scores of importance for each occupation across 120 characteristics (task). We normalize the range of these scores to $[0-1]$. We then use three versions of these scores. First, we simply use the absolute score of each characteristic. Second, we normalize the scores for each occupation so that they sum up to one. Third, we use principal component analysis in order to reduce the task space.

The Euclidean and Manhattan measure furthermore can be normalized. In an alternative specification, we first compute the factors before computing our distance measure.

A.2 NLSY

The NLSY79 is a representative sample of individuals who were in the ages 14-22 in the United States in the year 1979. Our sample consists of 1992 workers and 44655 person-year observations. We follow Guvenen et al. (2020) in our sample selection and focus on male workers from the years 1978 to 2010.

B Collected Comments

Important: the CPS actually allows a dependent/independent interviewing comparison post 1994 because intvw 5 (month 13) is again independently coded, so can do intvw4 intvw 5 (or similar) comparison for independent interviewing, and three times applied transition matrix between month 1 and 4 (or 13 and 16) to get dependent interview transitions.

References

Baley, Isaac, Ana Figueiredo, and Robert Ulbricht (2018). "Mismatch Cycles." In: *2018 Meeting Papers*.

Carrillo-Tudela, Carlos and Ludo Visschers (2023). *Unemployment and Endogenous Reallocation over the Business Cycle*. Institute of Labor Economics (IZA).

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. (2015). "Multidimensional Skill Mismatch." In.

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.

Lise, Jeremy and Fabien Postel-Vinay (Aug. 1, 2020). “Multidimensional Skills, Sorting, and Human Capital Accumulation.” In: *American Economic Review* 110.8, pp. 2328–2376. ISSN: 0002-8282.

Macaluso, Claudia (Jan. 2017). *Skill Remoteness and Post-Layoff Labor Market Outcomes*. Society for Economic Dynamics.

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.

Nimczik, Jan Sebastian (Jan. 2017). *Job Mobility Networks and Endogenous Labor*. Verein für Socialpolitik / German Economic Association.

A Tables

	(1)	(2)	(3)	(4)
	Miscoding	Euclidean (abs, std)	Euclidean (rel, std)	Euclidean (factor, std)
L.wage	0.272***	0.282***	0.278***	0.278***
	(9.40)	(9.40)	(9.27)	(9.70)
L.Wage x Dist	-0.0933***	-0.0660*	-0.0763**	-0.0814**
	(-3.62)	(-2.53)	(-2.86)	(-3.08)
Dist	0.624***	0.436*	0.497**	0.535**
	(3.60)	(2.46)	(2.74)	(2.95)
N	4007	4007	4007	4007
R^2	0.322	0.319	0.320	0.321

t statistics in parentheses

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

Table 2: Other Euclidean specifications

	(1)	(2)	(3)	(4)
	Miscoding	Manhattan (abs, std)	Manhattan (rel, std)	Manhattan (factor, std)
L.wage	0.272***	0.282***	0.278***	0.279***
	(9.40)	(9.35)	(9.21)	(9.62)
L.Wage x Dist	-0.0933***	-0.0598*	-0.0678*	-0.0708**
	(-3.62)	(-2.28)	(-2.56)	(-2.74)
Dist	0.624***	0.393*	0.438*	0.466**
	(3.60)	(2.22)	(2.44)	(2.61)
N	4007	4007	4007	4007
R^2	0.322	0.318	0.319	0.319

t statistics in parentheses

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

Table 3: Other Manhattan specifications

	(1)	(2)
	Miscoding	Angular (relative, std)
L.wage	0.272***	0.277***
	(9.40)	(9.26)
L.Wage x Dist	-0.0933***	-0.0748**
	(-3.62)	(-2.80)
Dist	0.624***	0.486**
	(3.60)	(2.67)
N	4007	4007
R^2	0.322	0.320

t statistics in parentheses

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

Table 4: Other Angular specifications

	(1)	(2)	(3)	(4)
	Miscoding	(std)”	Manhattan (rel, std)	Euclidean (rel, std)
L.wage	0.228***	0.292***	0.282***	0.282***
	(5.70)	(7.43)	(7.61)	(7.59)
L.Wage x Dist	-0.0545	-0.0738	-0.104	-0.100
	(-1.31)	(-1.50)	(-1.94)	(-1.84)
L.Wage x Dist^2	0.0721*	-0.0193	-0.00933	-0.00873
	(2.20)	(-0.60)	(-0.36)	(-0.34)
L.Wage x Dist^3	-0.0300	0.00650	0.0117	0.0110
	(-1.96)	(0.32)	(0.58)	(0.53)
Dist	0.373	0.490	0.699	0.672
	(1.34)	(1.47)	(1.93)	(1.83)
Dist^2	-0.485*	0.143	0.0637	0.0611
	(-2.16)	(0.64)	(0.36)	(0.35)
Dist^3	0.199	-0.0506	-0.0856	-0.0811
	(1.95)	(-0.37)	(-0.64)	(-0.58)
N	4007	4007	4007	4007
R^2	0.324	0.320	0.321	0.321

t statistics in parentheses

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

Table 5: Polynomial terms