A Dynamic Multinomial Model of Self-employment in the Netherlands[*]

Elisabeth Beusch[†] Arthur van Soest[‡]

Version date: 1st March 2019

Abstract

This paper presents a dynamic multinomial logit model to explain the transitions into and out of self-employment using Dutch micro-panel data, the LISS panel. Based on the estimates we simulate employment paths for benchmark individuals. These are used to illustrate the limitations of the common assumption in wealth and pension income modeling, that individuals remain in their observed labour state until retirement. In particular, we find that although one year transition probabilities out of self-employment are not more than 10%, the chances that individuals who are self-employed remain self-employed for the majority of the next ten years can be much smaller, and vary substantially with individual characteristics such as education level and personality.

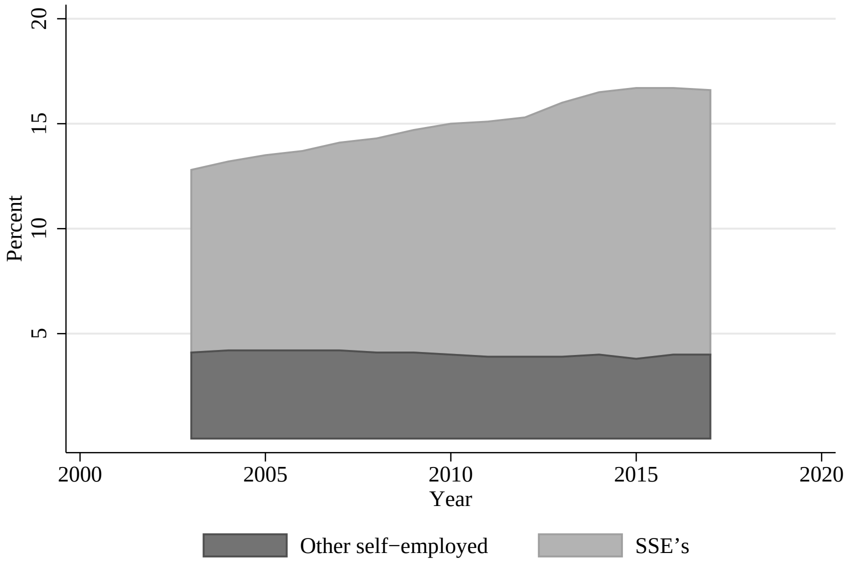
[*] This paper uses data of the LISS (Longitudinal Internet Studies for the Social sciences) panel administered by CentERdata (Tilburg University, The Netherlands). Elisabeth Beusch's research was funded by Instituut Gak through Netspar. We are grateful to Theo Nijman for his guidance and constructive comments. We also thank the participants of the Structural Econometric Group at Tilburg University for valuable comments and discussions.

[†] Tilburg University, e-mail: e.beusch@uvt.nl

[‡] Netspar, Tilburg University

1. Introduction

In recent years the number of self-employed in the Netherlands has grown substantially, leading to an increase of almost 30% in their share in the working population: from 12.8% in 2003 to 16.6% in 2017.[1] The main driver behind this growth have been the so called solo self-employed (SSE; in Dutch “zzp'ers”= zelfstandigen zonder personeel). As can be seen in Figure 1, the share of SSE in the working population increased from 8.1% in 2003 to 12.3% in 2015 and has remained rather stable since then. The share of other self-employed has, on the other hand, seen a slight decline since the financial crisis.



Data source: CBS Arbeidsdeelname; kerncijfers

Figure 1: Cummulative share of self-employment in the working population

Because of the growth in their numbers, Dutch policy makers have become more interested in the effects the self-employed (or SSE in particular) may have on the labour market, social security systems, or government finances. Accordingly, several recent policy papers describe the trend in self-employment and the characteristics of the self-employed, or analyze the performance of (groups of) self-employed and their role in the future; see, e.g., Bosch et al. (2012), Bosch (2014), CBS (2014) or Bolhaar et al. (2016). One key concern of the Dutch policy makers with respect to the self-employed and the social security system are the pension savings of the self-employed; see, e.g., Mastrogiacomo and Alessie (2015) or Knoef et al. (2016). While the pay-as-you-go pension (the so called AOW, making up the first pillar of the Dutch pension-system) covers all individuals who have always lived in the Netherlands, contributions to a fully funded pension plan (the second pillar) are, unlike for the

[1] All numbers are based on the CBS Statline *Arbeidsdeelname; kerncijfers* database downloaded on 4 May 2018.

large majority of employees, neither mandatory nor available or accessible for most of the self-employed.[2] Instead, the self-employed are expected to save themselves through (voluntary) savings (the third pillar). Such pension savings are tax-favoured for everyone with an "incomplete" second-pillar pension, in order to stimulate that individuals indeed save enough for their pension.[3]. This then raises the question whether the self-employed save enough for their retirement in the third pillar.

It turns out that the policy makers' concerns have some basis. Mastrogiacomo (2016) shows that while the self-employed have the same savings ambitions as the employed, they are on average not more likely to save in the third pillar. Only one third of the self-employed contributes to the third pillar, which indicates that the majority will fall short on their savings. In line with this finding, earlier research by de Bresser and Knoef (2015), and Knoef et al. (2016) found that the self-employed are less likely to meet their retirement expenditure or saving goals. Zwinkels et al. (2017) focus on the solo self-employed and estimate that about one fourth of all SSE fall short on their savings if a target replacement rate of 70% of earnings is used.

One simplifying assumption made in the pension wealth projections by Zwinkels et al. (2017), but also e.g. by de Bresser and Knoef (2015), and Knoef et al. (2016), is that the observed individuals remain in the labour state in which they were at the point in time when the data were collected. To our knowledge this assumption ("static micro-simulation") is standard in the pension literature and its consequences have not been discussed so far. However, given that the savings in the second (and third) pillar – and thus a large share of most individuals' pension wealth – are linked directly to the individuals' labour state, it may be worthwhile to study the validity and consequences of this assumption. This paper asks the following questions regarding this assumption of stationarity: what are the dynamics in the Dutch labour market when we consider the different labour states? How likely is it that somebody who is e.g. observed in self-employment will remain as such, and do the transition probabilities depend on observable personal characteristics?

To answer these questions we will use data from the LISS (Longitudinal Internet Studies for the Social sciences) panel, a representative sample of adult individuals in the Netherlands. The LISS panel is administered by CentERdata (affiliated with Tilburg University, The Netherlands). It is based upon a random sample of Dutch households drawn by Statistics Netherlands. Individuals of age 16 and older in the participating households are invited to answer survey questions on a monthly basis. The surveys cover domains such as work, education and income, but also a wide range of other topics, among which individual health

[2] The majority of second pillar pensions are built up via an employer based pension plan. Some industries have industry specific pension funds. In total about 90% of all employees are required to contribute to a pension plan (see https://www.rijksoverheid.nl/onderwerpen/pensioen/opbouw-pensioenstelsel) The Dutch system also knows industry specific pension funds, and self-employed individuals in such industries are also required to pay contributions to a second pillar pension. This includes, e.g., painters and doctors. The majority of the self-employed is however not active in such sectors.

[3] Recently, a pension fund for the self-employed has been opened but from a technical perspective it should be considered as a third pillar annuity

and personality. As such, the LISS panel offers a rich set of information on which we can build our analysis. It also allows us to distinguish between employees, SSE and other self-employed. Because we only observe a small number for some of the transitions once we split the self-employed in two groups, the main analysis is done without a distinction in self-employment types, even though such a distinction may seem desirable given the specific interest in the SSE in the Dutch policy debate.[4]

In addition to a set of personal and household characteristics that are common to most of the studies cited above, we also include personality traits and a (lagged) health index. Recent work on the economic importance of personality traits (see e.g. Borghans et al., 2008) has shown that personality traits matter for different labour market outcomes and, in particularly, the decision to become self-employed (see e.g. Beugelsdijk and Noorderhaven, 2005) This has also been studied in sociology, particularly in the research concerned with career counseling: Obschonka et al. (2013) for example construct a Entrepreneurship-Prone Big-Five Profile (EP) Distance measure and find for the U.S. that the EP distance's geographical distribution corresponds to observed entrepreneurial activity. We therefore also include the EP distance in our analysis. Good health has been identified as a factor that increases the probability to become self-employed for older workers in the US (Rietveld et al., 2015). The rich nature of the LISS data allow us to construct a health index and study the role of health for self-employment transitions in the Netherlands.

We first model self-employment in a static multinomial choice panel data framework with unobserved heterogeneity. We then extend our model to include dynamics to illustrate the importance of state dependence for someone's labour market status. We not only consider self-employment and wage employment, but also account for transitions into and out of paid work. The dynamic multinomial logit model we use for this is similar to that of e.g. Gong et al. (2004), who model the choice between not working, informal work, and formal sector work in Mexico, or Buddelmeyer and Wooden (2011) who model dynamics between casual and other types of employment in Australia. Oguzoglu (2016) follows Gong et al. (2004) to model the influence of disability on the decision of employment state, and Zucchelli et al. (2012) consider self-employment as an alternative to part-time employment for the elderly under possible ill-health. Another case in point is Prowse (2012) who includes self-employment when modelling the labour participation of women. Finally, Been and Knoef (2013) also use a dynamic multinomial logit model to explain self-employment decisions in the Netherlands, focusing on older workers of ages 50 and above and using administrative data. We consider all individuals of working age and use survey data, which has the advantage of providing rich background information such as personality or health indicators, as already emphasized above.

Our models incorporate unobserved heterogeneity, by allowing for correlated random effects following Train (2009). In the static model this will account for both unobserved heterogeneity

[4] As illustrated in Figure 1 most of the recent dynamics in self-employment seem to be driven by the SSE. Hence, looking at the self-employed as one group will still allow us to estimate the dynamics.

in the individuals as well as the omitted dynamic factors. In the dynamic model this allows us to differentiate between what Heckman (1981b) calls spurious and true state dependence. Note that this distinction is important if we want to understand the dynamics in the data. We solve the problem of initial conditions that arises in dynamic models following Wooldridge (2005) and Albarrán et al. (2015).

The paper continues as follows. Section 2 discusses the LISS panel and our sample selection process. Our model is presented in section 3 and the corresponding estimation results in section 4. Section 5 presents the simulation results based on the estimations. Section 6 concludes the paper.

2. Data

In this paper we make use of the LISS (Longitudinal Internet Studies for the Social sciences) panel. The LISS panel consists of monthly Internet surveys to a representative sample of households drawn from the Dutch population register.[5] Among the monthly surveys there are ten core studies with a longitudinal nature. Additionally, individuals are asked to fill in a basic survey, the household box, about the most important general characteristics of their household and its members such as age, gender, education, marital status, as well as their primary occupation and gross income. Individuals, or the contact person of the household — if there is more than one member of the household participating in the LISS panel — are asked to fill in the household box at the beginning when joining the panel and then prompted every month before each survey to fill in changes if such have occurred.

2.1. Self-employment identification and definition in the LISS panel

Among all longitudinal surveys, there are three instances within the LISS panel through which we can identify self-employed individuals. First, information about an individual's labour market status is stored in the household box. When asked about their primary occupation the survey responder is prompted with fourteen options. Of these, the third, "Autonomous professional, free-lancer, or self-employed", helps us to identify the self-employed. There are three problems with using this information to define an individual's labour state. First, it does not enable us to distinguish between SSEs and other self-employed. Second, the questionnaire gives no instruction on how "primary" is defined. Hence individuals with several occupations may rank them by hours, or by how much they identify themselves with each of them.[6]

[5] Households that don't own a computer and Internet connection are provided with such so that they can participate nevertheless.

[6] Even if the two coincide, the outcome may be different from how the person would be recorded in labour statistics. A case in point would be an individual that works part-time for less than 50% and takes care of the household for the rest of the week. This individual may well answer that the primary occupation is taking care of the household, and we would define the individual as "out of the labour force" based on this answer. However, according to the official statistics produced by Statistics Netherlands (CBS), anyone working for more than one hour per week in paid employment would count as part of the working population.

Lastly, this is further compounded by the fact that the survey is filled out by the household's contact person.[7] This can potentially lead to conflicting answers compared to if the individual had answered themselves, both for e.g. part-time (self-)employed and in particular also for DGAs[8].

Second, we can identify self-employed individuals using either one of two annual core studies: the *Work and Schooling* and the *Economic Situation: Income* study. Currently there are eleven waves available for both, covering the years 2008-2018. There are usually slightly fewer individuals who answer the income survey compared to the work and schooling survey.[9] Overall, however, the two samples are comparable in size and the majority of individuals answer both.

We will base our analysis on the income study for two main reasons. First, the income study allows us to identify solo-self-employed (SSE) while the work and schooling study does not allow for a distinction between SSE and other self-employed. Second, as will be discussed in more detail in section 2.2, the income study based sample suffers less from selection or attrition bias than the work and schooling study. The definition of labour status that we apply using the two studies are in general comparable.

In the work survey we can classify individuals according to their primary occupation based on hours worked.[10] This implies that the working population consists of all individuals who indicate that they have any paid work. We then split these individuals into employees and self-employed, based on follow-up questions on their primary occupation. The self-employed are therefore all individuals who answer that they are either *self-employed/freelancer*, or *independent professional* or *DGA*[11]. As an exception to the primary occupation rule, we also define those individuals as self-employed who indicate that they have their own business or they have a partnership as a secondary occupation next to being employed.[12] Going back to the definition by primary occupation, anyone who is not part of the working population and receives unemployment benefits or is looking for a job is classified as unemployed; the remainder of individuals is classified as not in the labour force. Because participants are asked about their primary occupation at the time the survey is taken, this classification is

[7] In most households with more than one adult household member, more than one of the adults participate in the LISS panel.

[8] DGAs ("directeur grootaandeelhouder" in Dutch — majority shareholder director) are individuals who work for an incorporated firm (either an NV or BV, i.e. a Ltd. or private Ltd. company in the British context) in a relatively high administrative position while holding a large part (or the majority) of the firm's shares. DGAS are treated as employees from the perspective of the Dutch tax authority while they may see themselves, in case of being (one of) the company owner(s), as self-employed.

[9] The numbers and response rates differ by period. E.g. in 2014, 7746 individuals were selected to answer the income survey and 7957 individuals for the work and schooling survey, with response rates of 78.9% and 82.6% respectively. In 2017, 6673 and 7256 individuals were selected with response rates of 80.3% and 80.4% respectively. See survey waves' individual codebooks for other years' numbers.

[10] If individuals work the same hours in two different jobs, they are asked to indicate the one that they consider more important. The survey does not indicate whether this importance should be attached according to e.g. own preferences or earnings.

[11] See Appendix B.1 for a discussion of the inclusion of DGAs.

[12] This increases the total number of self-employed in the sample by approximately 21%, compared to a definition by primary occupation only.

a snapshot of the individuals' situation at the moment of data collection (which takes place around April and May of each calendar year).

The timing in the income survey is different: unlike the work and schooling survey, the income survey asks individuals in period t not about their current situation, but about the different income sources that they have had over the whole calendar year $t-1$. That is, the 2008 survey asks about all income received in 2007. We classify all individuals that report receiving only income from employment as employees over the whole year. Individuals with income from both employment and self-employment are classified as self-employed, together with those individuals who report only income from self-employment.[13] Self-employment in turn is defined as indicating at least one type of entrepreneurial work activity. The activities that the income survey covers are (part-time) work as an entrepreneur or freelancer, work as an SSE, owning a company (including a private limited liability company or a limited partnership), or participating in a partnership (either a so called *maatschap* or *vennootschap onder firma, VOF*) and lastly whether one is making a profit (or loss) through enterprise in some way, except as spouse or partner cooperating in the business. Next, we classify all individuals as unemployed who report receiving some type of unemployment benefits and no other source of income, ignoring other social benefits. Because this will only classify those individuals as unemployed who are so for a whole calendar year, the unemployment definition is therefore stricter than in the work sample and is expected to cover a smaller number of individuals. Lastly, those individuals with no income from any of these sources are classified as not in the labour force.

Despite the difference in timing, the income based classification should, overall, lead to a similar outcome to the one based on individuals' occupation. We expect small differences to arise because individuals generally do not switch occupations in January only. But since we have data over several years we will still pick-up the changes in labour states in both data sets. Table 1 shows a comparison of the assigned labour state according to the income based classification ("Income"), compared with the work and schooling based classification ("Work and schooling") as well as the classification based upon the household box with background variables ("Background variables").[14] The rows show the share of matches with labour states in the "work and schooling" and "background variables" classifications for each labour state in the income based classification. As expected, the match for the unemployed is indeed not that good. Because our main concern are the self-employed however, we are not particularly concerned about the match in this category. More importantly, we can see that a fifth of the individuals who report income from self-employment activities are not captured

[13] In the income based classification, about 40% of the self-employed overall have income from both employment and self-employment. In the work based classification, these individuals only make up 25% of the self-employed; 32% of these are individuals who report self-employment as primary and wage-employment as a secondary occupation. As the income survey neither asks which of the two is their primary occupation, nor about the hours worked, it is hard to reconcile these differences. In addition to this, it also not clear where DGAs fall. See Appendix B.2 for a discussion of this issue.

[14] The comparison is made for all observations that fall in the corrected income sample as discussed in section 2.3, excluding the corrected breaks.

Table 1: Matching of labour states across definitions (in % of income definition)

	Work and schooling				Background variables			
Income	0	1	2	3	0	1	2	3
0: Employee	95.36	1.22	1.64	1.79	91.81	0.76	1.88	5.55
1: Self-employed	20.30	71.02	2.74	5.95	26.26	63.77	2.71	7.26
2: Unemployed	9.18	1.39	34.21	55.22	8.94	1.59	39.53	49.94
3: Not in LF	15.69	2.66	10.18	71.48	14.31	2.61	8.21	74.87
Total	71.86	8.69	4.13	15.33	69.75	7.77	4.07	18.41

Source: LISS panel, own calcuations.

as such in the work and schooling survey. Lastly, the average match with the "background variables" classification appears to be worse within the two categories that we care most about: employment and self-employment. This will further motivate our choice of data set, as discussed later in this section.

2.2. Sample selection

Each wave of the LISS panel generally has a response rate ranging from approximately 75 to 80%. Thus, every year we have answers from around 5000-6500 individuals, of which a few are incomplete. Individuals who leave the panel (i.e. stop answering the surveys all together) are replaced in later waves with refreshment samples. Most of the refreshment samples are stratified in order to improve the representativeness of the panel, and aim at oversampling difficult to reach groups with below-average response. For the two studies that we use, there are in total more than 14,000 individuals across the 11 waves and some 81,000 observations. Of all these observations we have information on the labour status for some 60,000 observations if we use the income based classification, and for almost 66,000 if we use the work and schooling data. Overall, we have overlapping information on approximately 45,000 observations.[15]

We restrict the sample to individuals from age 25 up to and including age 60. With these boundaries on age we limit ourselves to individuals' prime working years. We choose the lower bound at 25 years for the following reasons. First, the minimum wage in the Netherlands is a function of the worker's age until 23 in some sectors. As a result of this it seems that young workers in these sectors may have a higher risk of becoming unemployed close to their birthdays (Kabátek, 2015). Second, students who are finishing their education are also harder to classify in the work and schooling survey. Students may hold a (side) job, while studying and, given that April/May is relatively close to the end of the school year, they can also be considered first time job seekers. By age 25 most individuals should no longer be students

[15] Note that there are almost 11,000 combined for the years 2007 and 2018 that we cannot use because of the different timing of the two studies.

Table 2: Sample size across specifications

	Unconditional	Conditional on		Corrected		
Sample		Covariates	Sequence	Sequence	Pairs	Individuals
Income	60404	32649	28203	32529	26510	6019
Work & Schooling	65905	38307	31001	34293	28015	6278
Overlapping	45267	25737		25770		5346

Source: LISS panel, own calcuations.

and we are therefore more confident in attaining comparable classifications.

The age limit at 60 years on the other hand stems from the idea that older individuals face a different consideration set than younger workers and their choices might thus not be comparable. For one, an exit from the labour force includes (early) retirement for them, which is not a general option for e.g. a 40 year old worker. There is also concern that individuals may choose self-employment out of necessity once they become unemployed at an older age and cannot find a new job. (See e.g. Been and Knoef, 2013)

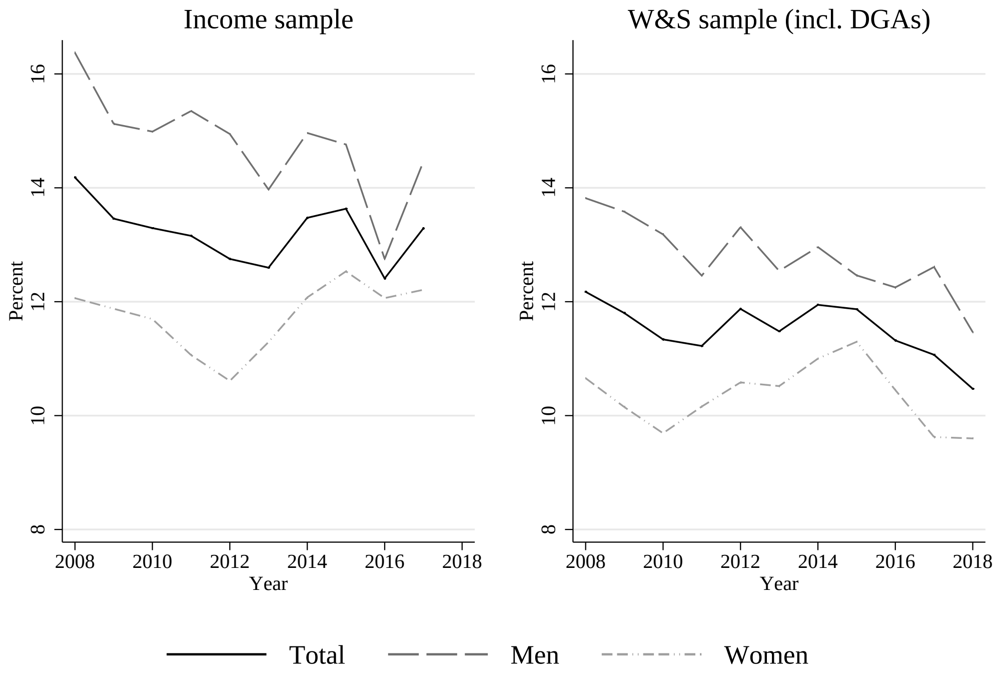
Furthermore, we limit ourselves to those individuals for whom information regarding the basic covariates, such as age, gender, household status, and education, is available. Table 2 shows the change in the sample size due to these constraints in the column titled "Conditional on Covariates" — we lose a bit less than half of the sample.[16] The corresponding self-employment shares, calculated as the number of self-employed individuals per year as a fraction of the respective number of individuals in the working population (i.e. the sum of employees and self-employed), in the two samples after these restrictions are shown in Figure 2.

There are two observations that we can make based on this figure. First, the share of self-employed in the working population is lower than in the actual population shown in Figure 1. While the share in the income based classification looks more or less equal in 2008, we have to take into consideration that CBS classifies individuals with multiple occupations based on the one in which they work most hours. Hence, as we also include part-time self-employed, the share in the LISS panel, if it were representative, should actually be higher than the population share.[17] Second, we observe an upward trend in self-employment shares according to none of the classifications in the LISS samples. These two observations hint that we may have to deal with both initial selection (lower initial shares) and attrition bias (lack of upward trend) in the LISS panel. We will discuss these issues in more detail below.

The last sample restriction that we have to make is model based. In the dynamic models,

[16] The loss of observations is mostly due to the age constraints or missing information on the labour state, and much less because of missing covariates.

[17] Differences should not stem from the denominator — CBS includes anyone in the working population who works at least one hour per week. Our definitions for the working population should therefore be comparable.



Data source: LISS Panel − Work & Schooling, Income, Background Variables

Figure 2: Comparison of self-employment shares in working population of 25-60 year olds across samples.

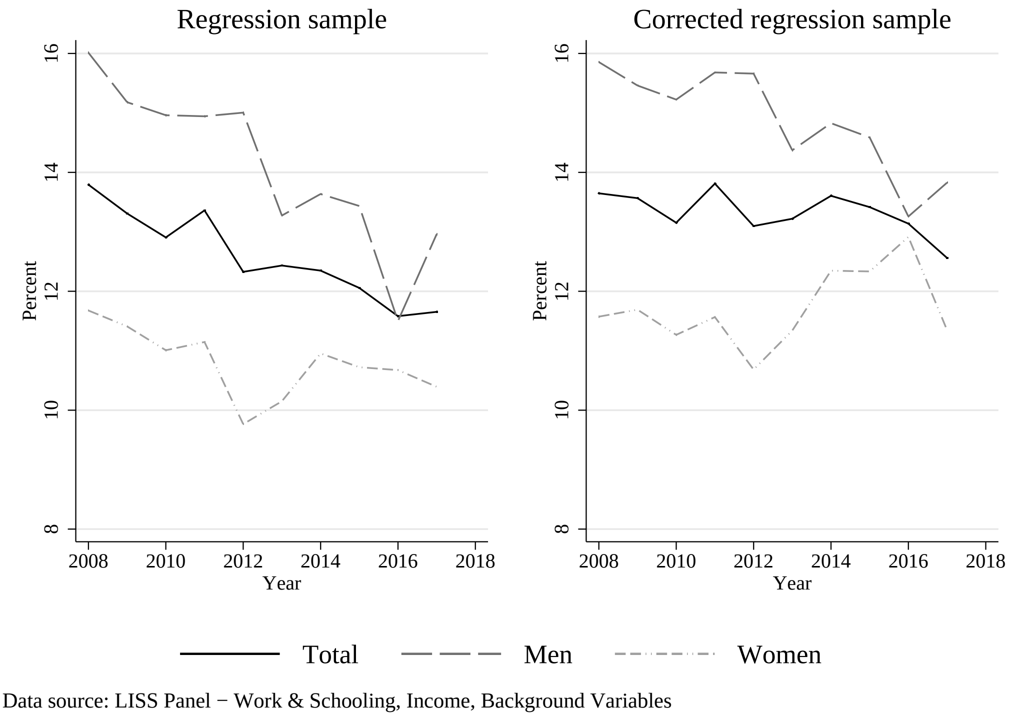
we want to model labour state outcomes based on individuals' past labour state. Hence we can only use those individuals for whom we have at least two consecutive observations. Furthermore, we have to discard any observations that are made after an individual skipped answering the surveys for one or more years.[18] These two restrictions make us lose approximately 15% of the observations, as can be seen in the column titled "Conditional on Sequence" in Table 2.

2.3. Filling in the gaps, selection, and attrition

Unfortunately, the sequence restriction does not only lead to a loss of data but also to a change in the evolution of employment shares over time in the new sample. In both the income and work based sample we see a clear downward trend after discarding the individuals with no sequence as well as any observations following a break in a sequence. The left panel of Figure 3, in comparison with the left panel in Figure 2 illustrates this for the income survey based classification. Thus the restriction of the sample to sequences at "best" worsens and at worst creates attrition bias.

A closer look at the observations that are discarded because of the sequence restriction

[18] I.e. if an individual answers the income survey in the years 2008-2012 and again from 2014-2018, we do not include the 2014-2018 block in the sample.

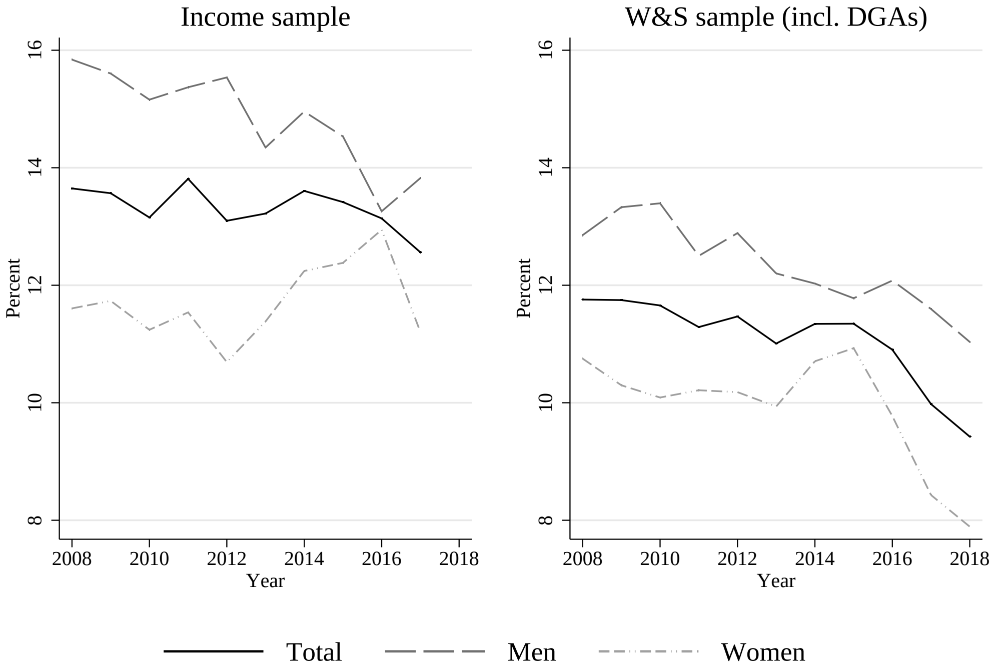


Data source: LISS Panel − Work & Schooling, Income, Background Variables

Figure 3: Comparison of self-employment shares based on income before/after correction

reveals that only a third of these observations belong to individuals who only participate in one wave of the income survey. That is, two thirds of the observations could be retained if we can correct for the break in the corresponding sequence. Furthermore, these are the observations (rather than the single wave answers) that are driving the downward trend shown in the left panel of Figure 3.

We take a parsimonious approach to filling in the breaks in sequences. That is, we only fill in one-year breaks, which account for the majority of all the breaks in the sequences. We however do not limit the correction to one break per individual but may fill in several breaks as long as these are only one period long. Nor do we extend a series beyond what we observe — that is, if an individual answers the work and schooling survey for more periods than the income survey, we do not use these other periods from the work and schooling survey to top up the income survey series. The procedure is as follows: we only consider the labour status information from the work based definition to fill in gaps in the income sample and vice versa, ignoring the information that is available in the background variables. We choose to do this because the number of correct matches between our definitions and a classification based on the background variables (as shown in Table 1) is quite low and we want to avoid generating false labour state transitions due to filled in gaps. In the income survey sample, we additionally make use of the question asked to self-employed individuals whether they were also self-employed in $t-2$. We only use this information to adjust for self-employment



Data source: LISS Panel − Work & Schooling, Income, Background Variables

Figure 4: Comparison of self-employment shares in corrected income and work panel

because we know that the work and schooling based sample misses 20% of the individuals that we categorise as self-employed in the income classification. To avoid selection bias, we do not use this question to fill in gaps for individuals without an answer in the work and schooling survey, because the question is only asked to individuals who are self-employed in $t-1$.

Based on this strategy we end up filling up at least one gap for more than one thousand individuals in the income survey sample. This in turn helps us to retain more than three thousand observations that would otherwise have to be discarded. More importantly, the downward trend in the self-employment shares is less pronounced after the correction. Figure 3 illustrates this for the income based classification. The impact of the correction on the work based classification is not that strong. This is likely because the number of additional observations that can be retained is rather low compared to the number of observations that is affected by the sequence restriction.

Figure 4 shows the corrected regression samples for both definitions. We can see that both samples display a slightly stronger downward trend in self-employment shares compared to the raw data in Figure 2. This negative trend also seems to worsen during later years and the effect is stronger in the work based classification. Because of this, and also due to the lower self-employment shares, we choose the income based classification for our analysis. The results discussed below, unless otherwise mentioned, are therefore for the income based

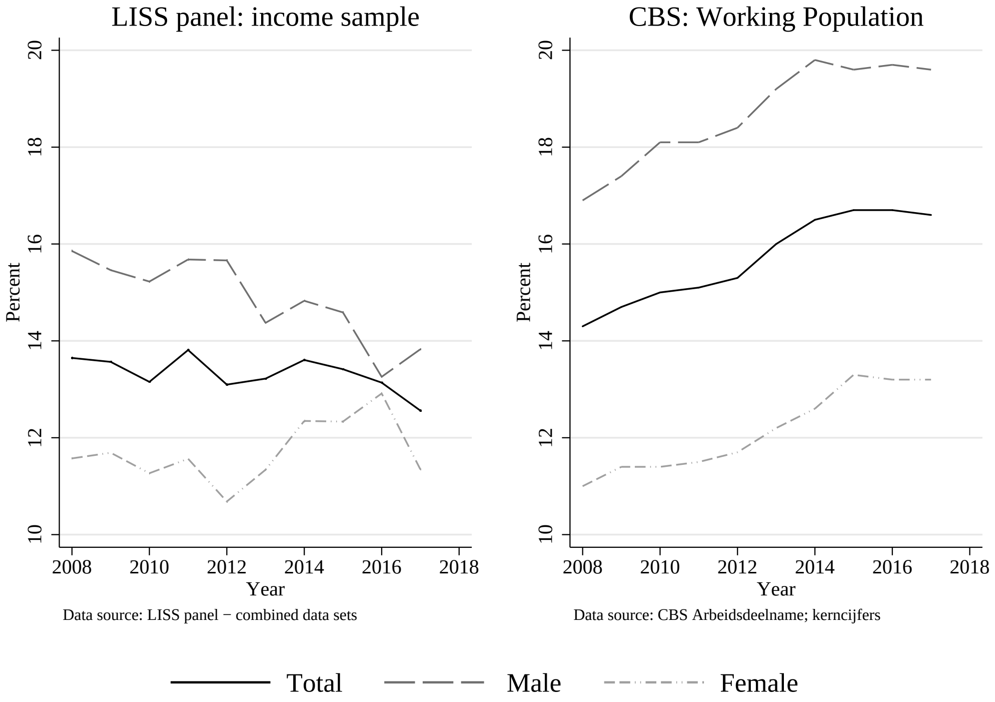


Figure 5: Comparison of self-employment shares in regression sample and population

classification (and the sample corresponding to it).

How does the income based sample compare to the population data that we have for the Netherlands? Figure 5 shows the same left panel as in Figure 4, but this time against the population data already seen in Figure 1 with the shares by gender added. As already mentioned earlier, the LISS panel fails to replicate the patterns seen in the population. Neither the magnitude nor the trend over time for the share of self-employment is matched.

This difference in patterns can be due to selection bias, i.e. self-employed individuals are more likely to not respond when invited to participate in the LISS panel, and to attrition bias, i.e. self-employed individuals are more likely to stop responding to the survey after having agreed to participate in the past. The observation that the numbers in the LISS panel already do not correspond to the CBS numbers from the beginning suggests that some selection bias exists. The positive trend, on the other hand, can be absent in the LISS panel due to either of the two biases. If all individuals are equally likely to leave the LISS panel and refreshment samples (which are stratified by observables such as age, gender, ethnicity) suffer from the same selection bias, we would still not see a positive trend over time. Neither would we if refreshment samples suffered from little selection bias, but the self-employed are instead more likely to leave. Most likely though, the problem is a mixture of both. In order to have a better understanding of what leads to the under representation of self-employed in the LISS panel, we examined the evolution of self-employment shares over time by recruitment waves. We found that they all display similar patterns. They start with relatively higher shares at recruitment, followed by a drop, and ending with a flat or slight downward trend. The observation that shares are initially closer to population figures therefore suggests to us that the trend is more affected by attrition bias than by initial selection.

Both selection and attrition bias are of less concern if they are driven by observables like age, gender, or eduction. In such a case one can correct for the bias by weighing the observations accordingly. We therefore tried weighing observations using weights based upon the observable characteristics that enter our model.[19] Weighing only leads to an increase of one percentage point in self-employment shares overall and does not change the trend. Furthermore, Wooldridge (2007, p. 1293) cautions about using weighting in panels. We therefore decided to not correct for selection and attrition by weighting on observables. If we want to correct for selection on unobservables we would have to impose a structure on the selection process. As this would require strong assumptions, and because the self-employment shares are initially not that different from population shares, we refrain from correcting for selection bias and instead focus on attrition bias only. Correcting for attrition bias requires weaker assumptions, especially if we can make use of an exclusion restriction. We test for attrition bias in section 3.4, where we also discuss how to correct for it.

2.4. Descriptive statistics

Most of the covariates that enter our model are based on information from the background variables. Before discussing these, we will first describe our measures for personality traits and some other indices that we use as controls.

2.4.1. Personality traits based on the Big-Five

The *Personality* core study of the LISS panel focusses on respondents' "personality and characteristics". Its questions are based on established questionnaires from the field of psychology that each have a different focus. One of these questionnaires is the short 50 question set for the Big-Five factor markers by Goldberg (1992). Individuals are asked to answer the questions on a 1 to 5 scale. We code their answers according to the corresponding International Personality Item Pool key.[20] For each of the five factors we then sum up the points on an individual basis and standardise these values with the mean and standard deviation of the complete LISS sample for each year, allowing us to interpret coefficients of the factors in terms of changes relative to the standard deviation. The Entrepreneurship-Prone Big-Five Personality Profile Distance (EP distance) measure is calculated using the non-standardised factor values following Obschonka et al. (2013).

To reduce the number of questions asked to individuals, the LISS panel only poses the Big-Five questions to its participants every second year. In the other years the Big-Five

[19] The weights are determined using population data downloaded from CBS Statline.

[20] https://ipip.ori.org/newBigFive5broadKey.htm retrieved on July 6, 2018.

Table 3: Big-Five factor markers and health index in the regression sample

	Sample as is...				...with gaps filled			
	Mean	Standard Deviation			Mean	Standard Deviation		
	overall		between	within	overall		between	within
Big-Five F1	-0.0150	1.0142	0.9745	0.3561	-0.0120	0.9804	0.9592	0.2651
— F2	-0.0015	1.0091	0.9388	0.4260	-0.0076	0.9626	0.9184	0.3171
— F3	0.0846	0.9620	0.9024	0.4005	0.0715	0.9181	0.8866	0.2975
— F4	-0.0051	1.0097	0.9418	0.4043	-0.0102	0.9665	0.9292	0.3007
— F5	0.0721	0.9898	0.9268	0.3946	0.0723	0.9467	0.9103	0.2927
N_{Big5}	16772				32316			
Health index	-0.2997	1.2143	1.1128	0.5174	-0.2967	1.2562	1.1934	0.5247
N_{health}	25446				32185			

Source: LISS panel, own calcuations.

related questions are only asked of new entrants. Furthermore, the personality survey was not asked to participants in 2016. In order to be able to use personality measures in our set of covariates we however need values on an annual basis. We therefore assume that the personality traits remain relatively stable across time. This assumption is supported by the data. As can be seen in the upper left hand panel of Table 3, the within variation is smaller than the between variation for all factor markers. Further support for this assumption is also provided by the study by Cobb-Clark and Schurer (2012) who have found that personality traits are stable for working-age adults over a four-year period. We fill in the gaps in Big-Five factor markers and EP distance by computing individual means over all observations available and substituting missing values in gap years with those means.[21] The results are shown in the upper right hand panel of Table 3.

2.4.2. Health

In order to explore whether an individual's general health has an impact on their labour status, we want to construct an objective measure of health. The survey-based LISS panel provides us with a subjective measure of overall health: Individuals are asked to asses their own health status in the *Health* core study. The answer-options provided are on a five-step scale ranging from poor to excellent. Studies have shown, however, that own health assessment in surveys can be biased; see, e.g., Jürges (2007). Hence we do not use the self-assessed health status directly. Instead we make use of the richness of the health survey and generate a linear index based on an ordered probit regression with random effects: We regress the survey answers on dummy indicators for the individual's perceived change in health relative to the last year, BMI based indicators for under- and overweight, as well as

[21] In case of the EP distance measure, which is a sum of squares, we fill in the gaps for the underlying Big-Five scores, and calculate the EP distance based on those in the gap years.

a set of indicators for self-assessed difficulties with daily tasks, regularly taken medication, health problems and hospital visits in the past year, etc. We run the regression with a sample consisting only of individuals in our chosen age range that are also in the final regression sample. As a sensitivity check, we repeat this including observation of those without a sequence (but still within the age range). The results are robust to the sample chosen, and we choose to use the health index based on the larger sample. See Appendix C for the regression results.

The health survey was not asked in 2014. We also miss some years for some individuals in the sample. We fill in these gaps but take a different approach than with the factor markers. We assume that health follows a dynamic process. We regress the health index, as an AR(1) process, on its lag and control for unobserved heterogeneity. We then take the estimated coefficients to calculate missing values. The summary statistics before and after corrections are shown in the lower part of Table 3. Because health may be endogenous to labour state choices, we will use the one period lag of the health index in our model.

2.4.3. Other covariates

In addition to the personality and health variables we will also make use of the following explanatory variables for the estimations in section 4. First, we include the individual characteristics age, gender, and education; for the latter we use dummy variables for medium level education (VMBO, VWO, or MBO diploma) and higher education (university (WO) and applied science university (HBO) degrees), which have been shown to have some correlation with the choice to be self-employed.[22]. We also include household specific variables: dummy variables controlling for whether an individual lives with a partner and/or has children, as well as the size of the household. These variables also have been found to have explanatory power in regressions explaining the decision to be self-employed; see, e.g., the overview of research on entrepreneurship by Blanchflower (2000).

Table 4 reports the means by labour status for all covariates. Overall we see that women are slightly over-represented in the sample as they make up 55% of all observations. Approximately half of the individuals in the sample have at least one child and the majority lives with a partner. Unsurprisingly, we find that those not in the labour force are mostly women, and that the majority of them lives with a partner. The two education dummy variables account for 96% of the sample, implying that only 4% of the sample has the lowest education level. Furthermore, we can see that the distribution of the two dummy variables varies between the working and non-working population — the higher educated are much more likely to do paid work. The distribution of the health index is left skewed with the mode at 0.76, and we can see directly that individuals in the working population have a higher health status than those not working. There are also differences in Big-Five factor markers between the working and non-working individuals.[23]

[22] See Bosch et al. (2012), Bosch (2014), CBS (2014) and Bolhaar et al. (2016)

[23] Note that our selected estimation sample is not perfectly representative of the overall LISS sample, as the

Table 4: Means and standard deviations (in brackets) of regression covariates

	Employee		Self-employed		Unemployed		Not in LF		All	
Age	44.08	(9.71)	45.93	(9.25)	47.52	(9.71)	48.31	(9.67)	45.07	(9.79)
Female	0.52	(0.50)	0.45	(0.50)	0.62	(0.48)	0.73	(0.44)	0.55	(0.50)
Lives with partner	0.76	(0.43)	0.78	(0.41)	0.54	(0.50)	0.74	(0.44)	0.75	(0.43)
Has children	0.55	(0.50)	0.57	(0.50)	0.44	(0.50)	0.46	(0.50)	0.53	(0.50)
Female with partner	0.38	(0.49)	0.35	(0.48)	0.33	(0.47)	0.58	(0.49)	0.41	(0.49)
Female with children	0.29	(0.45)	0.27	(0.44)	0.31	(0.46)	0.36	(0.48)	0.30	(0.46)
Medium education	0.58	(0.49)	0.52	(0.50)	0.72	(0.45)	0.72	(0.45)	0.60	(0.49)
Higher education	0.40	(0.49)	0.45	(0.50)	0.21	(0.41)	0.19	(0.39)	0.36	(0.48)
Household size	2.83	(1.33)	3.03	(1.43)	2.46	(1.37)	2.71	(1.36)	2.82	(1.35)
Health index	-0.12	(1.03)	-0.13	(1.04)	-1.18	(1.57)	-1.03	(1.53)	-0.30	(1.21)
F1: extraversion	0.00	(0.98)	0.15	(1.01)	-0.16	(0.94)	-0.18	(0.97)	-0.01	(0.99)
F2: agreeableness	-0.02	(0.95)	-0.09	(1.02)	0.05	(0.99)	0.09	(1.00)	-0.01	(0.97)
F3: conscientiousness	0.12	(0.90)	0.01	(0.97)	-0.03	(0.99)	-0.01	(0.96)	0.08	(0.92)
F4: emotional stability	0.08	(0.95)	0.06	(0.94)	-0.39	(1.05)	-0.34	(1.02)	-0.01	(0.97)
F5: openness	0.08	(0.93)	0.31	(1.01)	0.03	(0.98)	-0.11	(0.96)	0.07	(0.95)
EP distance	18.71	(5.12)	18.14	(5.11)	20.74	(5.95)	20.98	(5.78)	19.08	(5.34)

Source: LISS panel, own calcuations.

Table 5: Observed transition probabilities (in %) by gender

Labour state	Men				Women			
past \ current	0	1	2	3	0	1	2	3
0: employee	93.96	1.65	1.04	3.36	92.23	1.47	1.28	5.02
1: self-employed	8.15	89.35	0.51	1.99	9.34	84.73	0.71	5.22
2: unemployed	24.22	2.34	45.70	27.73	19.05	1.90	42.38	36.67
3: not in labour force	24.26	2.09	7.13	66.52	14.28	1.82	6.06	77.85
LISS total	74.67	13.08	2.51	9.75	66.01	8.82	3.46	21.72
CBS population	71.09	16.88	3.68	8.36	65.84	9.69	4.07	20.39

Based on 12014 and 14496 observation pairs respectively.

Source: LISS Panel and CBS Arbeidsdeelname; kerncijfers, own calculations.

When comparing means between the labour status groups, we find that the means of all variables except for the health index and the fourth personality factor marker are statistically different between employees and the self-employed. This is more or less in line with what the literature reports: men are more likely to be self-employed than women, the self-employed are on average older (compared to the employed), and the self-employed more frequently have a higher level of education. They are also more likely to have children. The argument for the EP distance by Obschonka et al. (2013) predicts that entrepreneur prone individuals are more extraverted, less agreeable, and more open to new experiences. We find all of this reflected in the differences of the means of the three factor markers. However, the theory also argues that entrepreneurs should be more conscientious, and we find the opposite. Still, the EP distance measure has a lower average value for the self-employed as theory would predict.

2.4.4. Labour state dynamics as observed in the data

In Table 5 the observed transition probabilities of the labour states by gender are shown.[24] We would like to point out two observations: first, employees are on average more likely to remain in the same state than the self-employed. This already suggests that the assumption of stationarity in labour states may have limitations. Second, there are substantial flows between these two states. The majority of the self-employed who do not continue as such switch to employment, and while the shares in percent may be misleading, the number of employees is much larger than the numbers in other labor states, and hence the largest contribution in numbers to entrants to self-employment are individuals making the transition from employment.

means of all except for the fourth factor marker, emotional stability, are statistically significantly different from zero. This is, however, not against our intuition: we expect that a more conscientious, or open to new experiences type of individual would be less likely to drop out or not answer when asked to participate in a survey.

[24] The transition matrix is, of course, affected by the imputations for the gaps in the data. The changes are relatively small, however. See Table 18 in Appendix D.

Comparing the transition matrices of men and women, we can see that women are less likely to remain in self-employment than men, and much more likely to remain out of labour force. Because of the large differences between men and women, it seems better to estimate the models separately by gender. Still, overall, we see some similar patterns for both genders. Lastly, it should also be noted that we observe, both in percent as well as in absolute numbers, very few changes from unemployment to self-employment and vice versa. This is likely partially due to our categorisation approach in the income based definition of self-employment.

Last but not least we also note that while our sample, as already discussed, does not replicate the share of self-employed fully, the population figures by the CBS[25] suggest that our classification error is mostly due to not identifying self-employed among employees and less from one of the other two labour states.

3. Model

This section presents the empirical model and the estimation procedure. Both are similar to the econometric specifications used by Gong et al. (2004) and Been and Knoef (2013). Note that the static multinomial model is nested in the dynamic model. We will therefore focus the discussion on the dynamic model, treating the static model as a special case. In the final subsection, we address the issue of attrition bias.

3.1. Dynamic multinomial model of labour states

We model the observed labour market state of an individual as the outcome of a utility maximisation process. Each individual re-evaluates the potential states every period, and chooses the labour state j that maximises utility for that period. In terms of the econometric specification we thus consider a discrete choice model where an individual i derives utility y^*_{ijt} from state j at time t. In other words:

$$y_{ijt} = \begin{cases} 1 & \text{if } y^*_{ijt} > y^*_{ikt} \qquad \text{for } j,k = 0,1,2,3; j \neq k; i = 1,\dots,N; t = 2,\dots,T \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where $y_{it} = (y_{i0t} \dots y_{i3t})$ is a column vector with a 1 in the position that corresponds to individual i's labour market state at time t and zeros everywhere else.[26]

Utility y^*_{ijt} from choosing state j is unobserved. It is assumed to be given by

$$y^*_{ijt} = X_{it}\beta_j + y_{it-1}\gamma_j + \alpha_{ij} + \epsilon_{ijt} \tag{2}$$

[25] The shares presented in Table 5 are calculated for the 25 to 60 year olds, taking the average for 2008–2017.

[26] The notation in column vectors allows reading the code for the estimation more easily. When not otherwise mentioned, vectors are assumed to be column vectors.

Here y_{it-1} is the vector that describes the individual's labour state in the previous period, X_{it} is a vector of k observed strictly exogenous explanatory variables, and the coefficient vectors γ_j and β_j are to be estimated. The variables included in X_{it} are individual as well as household characteristics that may have an influence on the utility y^*_{ijt}. These variables have been discussed in section 2.4.3. Next to these covariates, X_{it} also includes time dummies, to control for fixed time effects.

It should be noted that the model in this paper abstracts from any higher order dynamics. That is, apart from the first lag (the previous labor market state), no further lags enter the model. Because of data limitations, we also do not include a measure of job tenure. The model thus does not take into account that someone who has been self-employed for several years may be more likely to remain self-employed, compared to an individual who has been self-employed for one year only.

The unobserved terms $\alpha_{ij}, j = 0, \ldots, 3$ are treated as random effects that reflect time-invariant unobserved heterogeneity across individuals. They are assumed to be independent of the X_{it}. On the other hand, ϵ_{ijt} is an identically and independently distributed error term. The ϵ_{ijt} are assumed to be independent of X_{it} and α_{ij} and drawn from a Type 1 extreme value distribution. This distributional assumption implies that the state probabilities, given X_{it}, α_{ij} and $y_{i,t-1}$, are the well-known multinomial logit probabilities (see below for details).

Because equation (2) includes the lagged dependent variables, an initial conditions problem as described by Heckman (1981a) arises. First, we only have incomplete measures of individuals' labor state when entering the labor market in the LISS panel and are therefore not able to use them.[27] Our measure of initial conditions is therefore given by the first panel observation that we have of each individual. Second, it is probably wrong to assume that the first choice we observe, y_{i0}, is independent of the unobserved random effects $(\alpha_{i0}, \ldots, \alpha_{i3})$.[28] Here, we solve for the initial conditions following the approach suggested by Wooldridge (2005).

Following Wooldridge (2005), we specify the distribtuion of $\alpha_i = (\alpha_{i0}, \ldots, \alpha_{i3})$ conditional on the initial observation of the dependent variable y_{i0} as

$$\alpha_{ij} = y_{i0}\delta_j + \mu_{ij} \tag{3}$$

where $\mu_i = (\mu_{i0}, \ldots, \mu_{i3})$ is independent of y_{i0} and all X_{it}. This amounts to including the vector of dummies y_{i0} as additional regressors when estimating the model.

This approach to handle the initial conditions problem was constructed for balanced pan-

[27] There is one question (cw123) in the "Work and Schooling" core study that asks individuals in what type of organisation they worked in their very first job. One answer option is "self-employed". Unfortunately the structure of the question does not allow us to identify individuals who were unemployed or chose not to participate in the labor market after their (mandatory) schooling.

[28] Note that even if we were to observe an individual's very first choice of labour state after the end of obligatory schooling, this choice would probably not be independent of α_i either.

els, whereas our estimation is done on an unbalanced panel. Wooldridge (2005, p.44) writes that if the sequence of observations of the dependent variable and the sample selection mechanism are independent conditional on the initial conditions and the exogenous variables, the maximum likelihood estimation using the balanced sub-panel will be consistent. The LISS panel however is not only unbalanced because of sample attrition but also because some individuals only join in later waves (to make up for those that have left). If we were to restrict ourselves to the biggest balanced sub-panel, assuming that the necessary conditions hold, we would lose almost three quarters of our observations, and be left with less than 15% of all individuals. Restricting the estimation to the balanced subpanel thus implies an important loss in efficiency. We will therefore use the unbalanced panel.

Wooldridge (2009) proposes strategies for correlated random effects models with unbalanced panels but restricts himself to static models, so that his approach cannot be applied directly to the setting of a dynamic model. To the best of our knowledge, the only study that considers the problem of estimating dynamic non-linear random effects models with unbalanced panels is by Albarrán et al. (2015). These authors show that even if the selection mechanism is completely at random, unbalancedness can lead to inconsistent estimation in dynamic models and more conditions have to be satisfied for consistent estimation than in the static case. In particular, it is necessary that the process generating y_{i0} is in the steady state, and that the sequence in which periods an individual is observed, is independent from the shocks to the initial conditions. If this condition is not satisfied, the authors suggest that one estimates parameters that are specific to each sub-panel. This approach is however not feasible in a multinomial logit model with choice-invariant variables, as the number of parameters to estimate would become too large. Thus we only include one set of additional dummy variables within the initial conditions, i.e. equation (3), to control for the year in which an individual was first observed, and no interaction terms with these time dummies. δ thus becomes a $(T-2) \times J$ coefficient matrix to be estimated.[29]

3.1.1. Correlation among the random effects

In a last step we allow for correlation in the random effects. Like Gong et al. (2004) and Been and Knoef (2013) we assume that μ_i is drawn from a J-dimensional multivariate normal distribution with mean zero and covariance W.

Following Train (2009, chapter 9), we use a Choleski transformation for the multivariate normals. As Train (2009, p.238) writes, the advantage of using this approach is that “for any pattern of covariance, there is some set of loadings from independent components that reproduces that covariance”. We thus only have to make an assumption concerning the distribution of the unobserved heterogeneity but not of the covariance. Hence,

$$\mu_i = \xi_i^\intercal L^\intercal \tag{4}$$

[29] We have $J-1=3$ states in the initial conditions, as well as $T-1$ periods in which we can observe an individual for the first time, of which we drop one more period in order to identify all the parameters.

where ξ_i is a $J \times 1$ vector of independent standard normal distributed variables, and L is the lower triangular Cholesky factorization of μ_i's covariance matrix W, such that $LL^\intercal = W$.

Note that by allowing μ to be multivariate normal, the independence of irrelevant alternatives (IIA) assumption is no longer imposed. This IIA assumption is often seen as a drawback of the standard multinomial logit model. The estimates of the covariances will give an indication whether individuals who prefer one labour state are also more likely to prefer any particular other labour state. For example, if the covariance for states 1 and 2 (self-employed or unemployed) is positive, we should expect an individual, ceteris paribus, to have a higher probability of choosing self-employment when he or she has a high individual parameter for unemployment (ξ_2).

Substituting (4) and (3) in (2), and writing the utilities in vectorised form we have

$$y^*_{it} = X_{it}\beta + y_{it-1}\gamma + y_{i0}\delta + \xi_i^\intercal L^\intercal + \epsilon_{it} \qquad i = 1, \ldots, N \ ; \ t = 2, \ldots, T \tag{5}$$

where y^*_{it} is a $1 \times J$ vector of indirect utilities for individual i at time t, γ and δ are $J \times J$ matrix of parameters, β is a $k \times J$ matrix of parameters, and L contains the parameters of the covariance structure. All elements of β, γ δ, and L have to be estimated.

3.1.2. Likelihood function

The probability to observe a particular individual choosing labour state j at time t conditional on ξ_i in the multinomial logit model is then given by

$$Prob(y_{it} = j | X_{it}, y_{it-1}, y_{i0}, \xi_i) = \frac{exp(y_{it-1}\gamma_j + y_{i0}\delta_j + X_{it}\beta_j + \xi_i^\intercal L^\intercal)}{\sum_{k=0}^{J} exp(y_{it-1}\gamma_k + y_{i0}\delta_k + X_{it}\beta_k + \xi_i^\intercal L_k^\intercal)} \tag{6}$$

where L_j is the j$^{\text{th}}$ row of L.

It follows that the conditional probability of observing a sequence of choices for individual i is

$$Prob(y_i | X_i, y_{i0}, \xi_i) = \prod_t \prod_j Prob(y_{it} = j | X_{it}, y_{it-1}, y_{i0}, , \xi_i)^{\mathbb{D}_{ijt}} \tag{7}$$

where $\mathbb{D}_{ijt}$ is an indicator function denoting whether state j is chosen by the individual. The unconditional probability, or likelihood function, is then given by

$$Prob(y_i | X_i, y_{i0}) = \int_{\xi_i} Prob(y_i | X_{it}, y_{it-1}, y_{i0}, \xi_i) f(\xi_i) d\xi_i \tag{8}$$

where $f(\xi_i)$ denotes the multivariate normal distribution of ξ_i with means zero. The log likelihood function to be estimated is thus

$$log\mathcal{L} = \sum_{i=1}^{N} logProb(y_i | X_i, y_{i0}) \tag{9}$$

3.2. Static multinomial model of labour states

As already mentioned, the static model is a special case of the dynamic model. That is, in the static model we exclude the past period's labour state and as a consequence also drop the initial conditions. Equations (2) and (3) can then be re-written as

$$y^*_{ijt} = X_{it}\beta^s_j + \alpha^s_i + \epsilon^s_{ijt} \tag{10}$$

with

$$\alpha^s_i = \mu^s_i = \xi^{s\intercal}_i L^\intercal \tag{11}$$

where superscript s indicates that the coefficients are for the static regression. Detailed assumptions and likelihood contributions are then analogous to those for the dynamic model.

3.3. Maximum Simulated Likelihood

The results of this section hold for both the static and the dynamic model. Note that in its current form the multinomial logit model described by equations (6) – (9) is not identified as there are too many parameters. For identification purposes we therefore take $j = 0$, that is wage-employment, as the base category. β_0, γ_0, δ_0 and also the first column in L are normalised to zero, and the parameters for the other alternatives $j = 1, 2, 3$ are estimated relative to the base category.[30] For example, the covariance matrix is given by

$$W = \begin{pmatrix} l^2_{11} & \cdot & \cdot \\ l_{11}l_{21} & l^2_{21} + l^2_{22} & \cdot \\ l_{11}l_{31} & l_{21}l_{31} + l_{22}l_{32} & l^2_{31} + l^2_{32} + l^2_{33} \end{pmatrix}$$

and we estimate $J - 1 = 3$ sets of coefficients for each variable.

Furthermore, the probabilities given by equation (8), i.e. $\int_{\xi_i}$ in particular, have to be simulated. As Train (2009, chapter 10) writes, $log\hat{P}$ is not an unbiased estimator for $logP$ because of the non-linear log operation, even if $\hat{P}$ is an unbiased estimator of P. Thus the bias in the simulator for $logProb(y_i|X_i, y_{i0})$ translates into a bias in the maximum simulated likelihood estimator. This bias however diminishes as more draws are used in the simulation. Using a large number of draws for ξ_i on the other hand increases the computational burden of the estimation. One way to reduce this burden is, instead of using independent random draws, to use an alternative method that provides better coverage of the support of the distribution of the individual and therefore leads to greater accuracy for a given number of draws. This can be achieved using Halton draws. Both Bhat (2001) and Train (2009, chapter 9.3.3) have shown that e.g. 100 Halton draws can provide more precise results than 1000 random draws.

[30] Note that dimensions are now reduced by 1, and all vectors and matrices are now of dimensions $(\cdot \times J - 1)$.

In order to simulate $\int_{\xi_i}$, we take 150 draws for each individual from a $J-1$-dimensional Halton sequence in which we closely follow the method described in Train (2009, chapter 9.3.3). Based on the discussion above, 150 draws for each individual should be sufficient. Furthermore the panel is based on more than six thousand individuals, i.e. large in itself, which should also lower the need for more draws. In addition we randomise the Halton draws following the procedure described by Bhat (2003). As a base of the Halton sequences we use the vector of primes given by $[11, 13, 7]$.

Returning to the model, the probability of observing an individual's observed sequence of labour state choices in the simulation is given by

$$Prob(y_i|X_i, y_{i0}) = \frac{1}{R}\sum_{r=1}^{R}\prod_{t=2}^{T}\left(\prod_{j=1}^{J} Prob(y_i|X_{it}, y_{it-1}, y_{i0}, \xi_i)^{\mathbb{D}_{ijt}}\right)^{\mathbb{S}_{it}}$$

The simulated loglikelihood function is thus given by

$$log\mathcal{SL} = \sum_{i=1}^{N} log\left[\frac{1}{R}\sum_{r=1}^{R}\prod_{t=2}^{T}\left(\prod_{j=1}^{J} Prob(y_i|X_{it}, y_{it-1}, y_{i0}, \xi_i)^{\mathbb{D}_{ijt}}\right)^{\mathbb{S}_{it}}\right] \tag{12}$$

where R is the number of Halton draws taken to simulate $\int_{\xi_i}$, and $\mathbb{S}_{ijt}$ is an indicator function to control for the unbalanced panel, denoting whether an individual's observation enters in the estimation. That is, $\mathbb{S}_{ijt} = 1$ if y_i, y_{it-1}, X_{it} are observed.

All model specifications for the correlated random effects models reported in section 4 are estimated using an own code written in Matlab 2017a. We solve them as an unconstrained minimisation[31] problem using KNITRO as a solver and supplying the gradient as defined in appendix A. As the best guess for the starting values we first estimate a pooled multinomial regression on the same covariates, including the lags, for each specification. The only difference is the non-inclusion of the unobserved individual heterogeneity. For this step we make use of the function "mnrfit" in Matlab's statistics toolbox. If ignoring the unobserved heterogeneity leads to a large bias in the estimates, these starting values may be far off. This in turn could pose a problem if the simulated log likelihood is very flat or "bumpy", in which case the solver may not find a local minimum.

3.4. Attrition Bias

As Verbeek and Nijman (1992, p. 681) note it is well known since Heckman (1976, 1979) published his seminal papers that "inferences based on either the balanced sub-panel or the unbalanced panel without correcting for selectivity bias, may be subject to bias if the nonresponse is endogenously determined". Within the context of this paper we are thus first interested in whether self-employed individuals are more likely to leave the LISS panel, and

[31] Matlab only solves minimisation problems and therefore we estimate the negative log-likelihood function.

thus contribute to the unbalancedness of the panel, and second, if they do so, whether this leads to biased estimates for the model above or not.

In order to test for attrition bias we use a variation of the variable addition test from Verbeek and Nijman (1992). They consider three possible variables that can be included in the regression: the number of waves and individual participates in the panel, an indicator whether the individual participated in all waves, and an indicator whether an individual was observed in the previous period. Because we also want to make use of the refreshment samples, the first and third suggested variable are not applicable, and the second only with a difference in interpretation. Instead we construct a variable that measures the ratio of periods in which an individual participated over the maximum amount of periods they could have participated. This is still a function of the response indicator and thus follows the idea of the variable addition test. If attrition was independent of the unobservables in the model, this additional variable should not enter the model significantly under the null of no attrition bias.

We find e.g. in the dynamic model with Big-Five factor markers that the share of waves during which an individual answers the LISS panel enters the multinomial model statistically significantly in the pooled model as well as in the men's sample: There is evidence that the model suffers from attrition bias.[32] We therefore estimate an extension of the model adding a Heckman correction term (estimated in a first stage). Formally the attrition model extends equations (1) and (5) as follows:

$$A_{it} = \begin{cases} 1 & \text{if } A_{it}^* > 0 \qquad \text{for } i = 1, \dots, N; t = 2, \dots, T \\ 0 & \text{otherwise} \end{cases} \tag{13}$$

$$A_{it}^* = X_{it-1}\beta^H + y_{it-1}\gamma^H + z_{it-1}\theta + \nu_i + \psi_{1,it} \qquad i = 1, \dots, N \; ; \; t = 2, \dots, T \tag{14}$$

where A_{it}^* is latent and A_{it} indicates whether an individual is observed in the LISS panel at time t or not.[33] Together, equations (13) and (14) model the attrition process in the first stage of estimation. X_{it-1} contains the same vector of regressors as X_{it} in equation (5) but one period lagged. y_{it-1} is a vector of dummies indicating the labour state of the individual one period past. Lastly, z_{it-1} is a vector of variables entering the attrition equation but not the other equations in the model (the "exclusion restrictions"). For the exclusion restriction we follow Cheng and Trivedi (2015) and use the number of days that individuals took to answer the last wave of the *Economic Situation: Income* core study. We further include a

[32] We can reject the null hypothesis of the coefficients of the shares being jointly equal to zero at the 5%-level in the pooled regression, and at 0.1% for men. For women the p-value is 0.4132 and we fail to reject the null hypothesis.

[33] We exclude individuals from this step if they leave our sample because they are turning 61. The first stage only tries to correct for an individual's own choice to (not) answer the LISS panel survey; it does not correct for the sample selection made by our exclusion of some groups as described in section 2.

dummy that controls for whether the individual answered within the deadline of the first call to participate in the survey, or only after the reminder, as well as an interaction term of the two. The error term ν_i is assumed to be a time invariant random effect while $\psi_{1,it}$ is assumed to be iid standard normally distributed.

Stage one is estimated in Stata as a panel probit model with random effects. The second stage is then given by

$$y_{ijt} = \begin{cases} 1 & \text{if } y^*_{ijt} > y^*_{ikt} \qquad \text{for } j,k = 0,1,2,3; j \neq k; i = 1,\dots,N; t = 2,\dots,T \\ 0 & \text{otherwise} \end{cases} \tag{15}$$

$$y^*_{it} = X_{it}\beta + y_{it-1}\gamma + y_{i0}\delta + \xi'_i L' + \lambda(\cdot)\tilde{\sigma}_{12} + \psi_{2,it} \qquad i = 1,\dots,N \; ; \; t = 2,\dots,T \tag{16}$$

where we now have $\epsilon_{it} = \lambda(\cdot)\tilde{\sigma}_{12} + \psi_{2,it}$ and $\lambda(\cdot) = \frac{\phi(\cdot)}{\Phi(\cdot)}$ is the inverse Mills ratio based on the estimates of the panel probit model.[34] Note that because we have to make the assumption that $\psi_{2,it}$ follows a Type 1 extreme value distribution we differ from the original Heckman correction. Consequently the coefficient vector $\tilde{\sigma}_{12}$ cannot be interpreted as the covariance vector of $\psi_{1,it}$ with each element in the vector $\psi_{2,it}$.

4. Estimation results

We estimate four different models: the baseline model with the basic personal and household characteristics, and three other models in which we add the (lagged) health index, the EP distance measure, and the Big-Five factor markers, respectively. Because the first stage of the Heckman correction is the same for both the static and dynamic model, we will discuss these first stage results first. In our description of the estimation results and the effects we can derive, all statements are meant as partial effects, keeping all other observed and unobserved characteristics constant, and averaging over the complete sample.

We will focus on the results for the separate regressions for men and women, and do not discuss the pooled regression results in detail — the pooled models do not lead to different conclusions with respect to the direction of the effects that we find. The estimated coefficients differ at times between men and women, which is why we prefer to look at the genders separately. The pooled regression results can be found in Appendix D.

[34] In theory one could allow for time varying regressors across different panel waves by estimating the attrition equation separately for each wave. Here we however assume constant coefficients.

4.1. First stage — Heckman correction

In the first stage regression the sample also includes all individuals, within the selected age range, for whom we have only one observation in the LISS panel.[35] The results for women are presentend in Table 6 and those for men in Table 7.[36]

When looking at Table 6 for women, we find rather weak effects for the labour market state indicators for self-employment and not participating in the labour force on the chances to remain in the sample. The same effects are statistically more significant and of larger magnitude for men. We find the same signs for both genders though. That is, compared to employees, self-employed individuals are less likely to be observed in the following period, and individuals not in the labour force are more likely to be observed again. In particular for men, and in combination with the findings of the variable addition test, this supports our argument of the need to correct for attrition bias: attrition is at least part of the explanation why the self-employed are underrepresented in later waves of the panel.

Furthermore, we see that the first stage estimates for the basic personal and household characteristics are robust across the different model specifications. Only age has a statistically significant positive coefficient, implying that older individuals are more likely than younger individuals to continue participating in the LISS panel. For women, we do not find a statistically significant effect for any other of the basic covariates. This is in contrast to men, where the dummy variable for high education is statistically significant at the 1%-level, and the medium education dummy variable is statistically significant at the 5% level. Together the coefficients on these dummies show that men are more likely to continue participating in the LISS panel the higher their education.

The variables that are excluded from the main model ("the exclusion restrictions") have the expected signs in the first stage. The more days individuals take to answer after the invitation to participate in the survey has been issued, the less likely they are to return in the following year. For women, this effect is particularly strong for those who answer after the first call for participation in a survey[37] (as shown by the significant interaction terms).

In terms of the variables driving the different model specifications, we do not find statistically significant effects of the respondents' contemporary health status on the probability to be observed in the next period. This holds for both men and women. The EP distance measure is not statistically significant for women either, but it is significantly negative at the 5% level for men. Recall that the EP distance's interpretation is that the lower its value is, the more likely it should be that an individual is self-employed. The results show that conditional on employment or self-employment status, less entrepreneurial individuals are less likely to stay in the sample, perhaps because they are more pressed with time.

[35] This does not include new entrants to the income survey from the last year of data collection as we do not know yet whether they will return or leave in the next wave.

[36] The pooled regression results are in Appendix D, Table 21.

[37] For each of the surveys the LISS panel collects data in two calendar months. A reminder is sent to all those panel members that did not complete the questionnaire during the first month.

Table 6: First stage Heckman regression results, women

Model	Baseline	Health index	EP distance	Big5 Factors
Self-employed in $t-1$	-0.14**	-0.11*	-0.13**	-0.11*
	(0.048)	(0.047)	(0.047)	(0.047)
Unemployed in $t-1$	0.10	0.12	0.10	0.10
	(0.088)	(0.088)	(0.087)	(0.087)
Not in labour force in $t-1$	0.07	0.08*	0.07	0.07*
	(0.037)	(0.037)	(0.037)	(0.037)
Age	0.02***	0.02***	0.02***	0.01***
	(0.002)	(0.001)	(0.001)	(0.002)
Has partner	-0.02	-0.03	-0.02	-0.05
	(0.041)	(0.040)	(0.040)	(0.040)
Has child	-0.04	-0.06	-0.03	-0.03
	(0.050)	(0.049)	(0.049)	(0.049)
Middle eduction	0.07	0.05	0.03	0.04
	(0.081)	(0.079)	(0.080)	(0.080)
High education	0.08	0.05	0.03	0.06
	(0.084)	(0.081)	(0.082)	(0.084)
Household size	-0.02	0.00	-0.01	-0.00
	(0.021)	(0.021)	(0.020)	(0.020)
Days until answered	-0.01***	-0.01***	-0.01***	-0.01***
(within call)	(0.002)	(0.002)	(0.002)	(0.002)
Answered in first call	0.32***	0.31***	0.31***	0.30***
	(0.048)	(0.047)	(0.047)	(0.047)
Interaction term	-0.01***	-0.01**	-0.01**	-0.01**
days x first call	(0.003)	(0.003)	(0.003)	(0.003)
F1: extraversion				-0.01
				(0.016)
F2: agreeableness				-0.01
				(0.018)
F3: conscientiousness				0.12***
				(0.016)
F4: emotional stability				-0.04**
				(0.016)
F5: openness for experience				-0.04*
				(0.018)
Health index		0.01		
		(0.013)		
EP distance			0.00	
			(0.003)	
Constant	0.39***	0.46***	0.46***	0.47***
	(0.117)	(0.113)	(0.128)	(0.116)
Observations	16,868	16,605	16,679	16,679
Number of nomem_encr	4,045	3,848	3,869	3,869
ρ	0.145	0.0870	0.102	0.0979
σ_u	0.412	0.309	0.336	0.329
LL	-6559	-6252	-6336	-6307

Dependent variable: Indicator variable for labour state observed in t.
Standard errors in parentheses; *** p<0.001, ** p<0.01, * p<0.05

Table 7: First stage Heckman regression results, men

Model	Baseline	Health index	EP distance	Big5 Factors
Self-employed in $t-1$	-0.23***	-0.19***	-0.21***	-0.19***
	(0.048)	(0.047)	(0.047)	(0.048)
Unemployed in $t-1$	-0.12	-0.05	-0.09	-0.08
	(0.111)	(0.113)	(0.110)	(0.111)
Not in labour force in $t-1$	0.12*	0.13*	0.15*	0.16**
	(0.063)	(0.062)	(0.063)	(0.063)
Age	0.02***	0.02***	0.02***	0.02***
	(0.002)	(0.002)	(0.002)	(0.002)
Has partner	0.05	0.03	0.06	0.05
	(0.051)	(0.050)	(0.050)	(0.050)
Has child	-0.00	-0.01	0.00	0.00
	(0.064)	(0.062)	(0.062)	(0.062)
Middle eduction	0.21*	0.20*	0.22*	0.20*
	(0.089)	(0.086)	(0.087)	(0.087)
High education	0.27**	0.23**	0.26**	0.25**
	(0.091)	(0.089)	(0.089)	(0.091)
Household size	-0.06*	-0.05	-0.05*	-0.05
	(0.026)	(0.025)	(0.025)	(0.025)
Days until answered	-0.01***	-0.01***	-0.01***	-0.01***
(within call)	(0.003)	(0.003)	(0.003)	(0.003)
Answered in first call	0.19***	0.21***	0.17**	0.17**
	(0.056)	(0.055)	(0.056)	(0.056)
Interaction term	-0.00	-0.00	-0.00	-0.00
days x first call	(0.004)	(0.004)	(0.004)	(0.004)
F1: extraversion				-0.01
				(0.019)
F2: agreeableness				-0.01
				(0.019)
F3: conscientiousness				0.06**
				(0.019)
F4: emotional stability				0.02
				(0.019)
F5: openness for experience				-0.01
				(0.021)
Health index		0.03		
		(0.016)		
EP distance			-0.01*	
			(0.003)	
Constant	0.23*	0.33**	0.43***	0.34***
	(0.131)	(0.127)	(0.142)	(0.131)
Observations	13,824	13,578	13,675	13,675
Number of nomem_encr	3,242	3,059	3,104	3,104
ρ	0.203	0.137	0.158	0.160
σ_u	0.505	0.398	0.433	0.436
LL	-5046	-4766	-4865	-4860

Dependent variable: Indicator variable for labour state observed in t.
Standard errors in parentheses; *** p<0.001, ** p<0.01, * p<0.05

For both men and women, more conscientious individuals are statistically significantly more likely to continue participating in the LISS panel, as expected. For men, we do not find any other individually significant effects, but we do find that the Big-Five factor markers are jointly statistically significant (at a 2% confidence level).[38] For women, we find that emotional stability and openness for experience have negative effects on the probability to stay in the sample, which is not what we would have expected.[39]

4.2. Second stage — static estimation results

In order to be able to compare the static and dynamic models, we restrict ourselves to the sample of the dynamic model and estimate the static model excluding those individuals for whom we only have a single observation.

When we look at the static estimation results, we find similar effects of personal and household characteristics in all of the specifications. See e.g. the results for the model with Big-Five factor markers for women in Table 8 and for men in Table 9.[40] We find for women that only few of the personal and household characteristics are individually statistically significant in the self-employment equation. In contrast to this, almost all of them are highly statistically significant (i.e., most at the 1%-, and some at the 5%-level) in the other two equations. In other words, personal characteristics do not help us much to explain the difference between women being self-employed or working as an employee, but they are helpful in explaining the difference between employment and unemployment or not participating in the labour market. The only variables that have a statistically significant coefficient in the self-employment equation are age and household size (both at the 1%-level). We find that the self-employed are on average older, and that the larger the household, the more likely women are to choose for self-employment over wage-employment.

For men the coefficients on personal and household characteristics are also very similar across different models, but they differ from what we found for women. We still find that age is statistically significant at the 1%-level and has a positive sign, implying that also for men, the chances that an individual is self-employed increase with age. However, we do not find a statistically significant coefficient for household size. Unlike for women, we find that men with high education are significantly more likely to be self-employed. And while household size did not matter, we do find for men that having at least one child decreases the probability of being self-employed significantly.

For both genders, we find that the lagged health index does not enter the self-employment equation significantly. Health does have significant effects in the other equations though, and the coefficients are also jointly significant. The negative sign of the coefficients suggests

[38] The results for the corresponding restricted model are available on request.

[39] The Big-Five factor markers are also jointly statistically significant with a p-value of 0.000 in the regression for women.

[40] The pooled regression results for the model with Big-Five factor markers are in Appendix D, Table 22. The results for all other models are available upon request.

Table 8: Static model with Big-Five factor markers, women

	Self-employed	Unemployed	Not in Labour Force
Constant	-17.84***	-5.96***	-4.65***
	(1.2323)	(1.0431)	(0.8464)
Age	0.13***	0.07***	0.1***
	(0.0169)	(0.0135)	(0.012)
Has partner	0.45	-0.53**	0.75***
	(0.305)	(0.2302)	(0.2087)
Has child	-0.64	-0.46*	-0.73***
	(0.3988)	(0.277)	(0.2374)
Middle education	-0.24	-1.28***	-1.51***
	(0.6524)	(0.4414)	(0.3927)
High education	0.39	-2.83***	-2.92***
	(0.6965)	(0.4805)	(0.4226)
Household size	0.47***	0.26**	0.34***
	(0.1511)	(0.1203)	(0.0958)
F1: extraversion	0.13	-0.09	-0.07
	(0.1555)	(0.1128)	(0.0955)
F2: agreeableness	-0.05	0.18	0.19*
	(0.1572)	(0.1118)	(0.0972)
F3: conscientiousness	0.09	-0.58***	-0.65***
	(0.1581)	(0.121)	(0.0998)
F4: emotional stability	-0.24*	-0.47***	-0.32***
	(0.1417)	(0.1059)	(0.0864)
F5: openness for experience	0.36**	0.36***	0.17*
	(0.145)	(0.1097)	(0.0987)
Inverse Mills Ratio	10.49***	-6.63***	-8.06***
	(1.2557)	(1.3516)	(1.0045)
L	6.99***		
	(0.3628)		
	2.48***	3.46***	
	(0.2235)	(0.2033)	
	3.23***	3.65***	1.55***
	(0.2143)	(0.1653)	(0.1167)
W	48.8702	17.3339	22.5515
	17.3339	18.1315	20.6267
	22.5515	20.6267	26.1073
Observations: 14435			
Nr. of Individuals: 3267			
Loglikelihood: -7704.03			

Regression including year fixed effects.
(Non-robust) standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.10

Table 9: Static model with Big-Five factor markers, men

	Self-employed	Unemployed	Not in Labour Force
Constant	-22.07***	-10.28***	-2.5***
	(1.4174)	(1.5229)	(0.9608)
Age	0.2***	0.13***	0.07***
	(0.0181)	(0.019)	(0.0127)
Has partner	0.08	-0.77**	-0.76***
	(0.3333)	(0.3629)	(0.2579)
Has child	-1.65***	-0.38	-0.73**
	(0.348)	(0.4375)	(0.309)
Middle education	0.67	-1.46***	-2.45***
	(0.7667)	(0.5278)	(0.3789)
High education	1.6**	-2.6***	-3.83***
	(0.7987)	(0.5816)	(0.4332)
Household size	0.21	-0.03	0.16
	(0.1608)	(0.1919)	(0.1335)
F1: extraversion	0.39**	-0.07	-0.06
	(0.1608)	(0.1678)	(0.1028)
F2: agreeableness	-0.32**	0.13	0.05
	(0.1311)	(0.1309)	(0.0938)
F3: conscientiousness	0.02	-0.3**	-0.46***
	(0.1423)	(0.1407)	(0.1037)
F4: emotional stability	-0.17	-0.6***	-0.67***
	(0.1417)	(0.1422)	(0.0988)
F5: openness for experience	0.73***	0.34**	0.23**
	(0.1359)	(0.1464)	(0.1009)
Inverse Mills Ratio	21.18***	-0.23	-7.86***
	(1.4209)	(2.2355)	(1.2027)
L	6.37***		
	(0.3519)		
	2.23***	3.15***	
	(0.273)	(0.1969)	
	2.33***	2.44***	1.53***
	(0.2468)	(0.1824)	(0.1421)
W	40.5225	14.2217	14.8488
	14.2217	14.9076	12.8843
	14.8488	12.8843	13.7103
Observations: 11967			
Nr. of Individuals: 2647			
Loglikelihood: -5244.5			

Regression including year fixed effects.

(Non-robust) standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.10

that individuals who had bad health one period earlier are more likely to be observed in unemployment or out of the labour force relative to being employed.[41] Similarly, we do not find a statistically significant effect of the EP distance for women, although the sign is, as we would expect, positive in the other two equations. For men on the other hand, we find a significant effect (p-value < 0.01) also in the self-employment equation. The sign is negative which is in line with our expectations: as the EP distance decreases, the more likely individuals become to be self-employed.

With respect to the Big-Five factor markers we find different effects by gender. For both women and men, individuals who are more open to experiences are also more likely to be self-employed than employees. The effect is however twice as strong in magnitude for men. For women, we find that one additional factor marker, emotional stability, is significant at the 10%-level. The negative sign implies that higher emotional stability reduces the chances to become self-employed. For men on the other hand, we find that high scores for extraversion increase the likelihood of being self-employed, and that high scores for agreeableness reduce it. These effects are in line with our expectations.

In all static models, the coefficient of the inverse Mills ratio is highly statistically significant for the self-employed and for the equation explaining "not in the labour force." It has a positive sign in the self-employment equation and a negative sign in the equation for not in the labour force. For women, the coefficient is also statistically significant and negative in the unemployment equation. This suggests that, keeping observed characteristics constant, attrition is correlated with the unobserved factors driving someone's labor market status, implying that it is important to correct for attrition bias in the model.[42]

Looking at the estimates for the elements driving the unobserved heterogeneity components in the mixed logit model (the matrix L driving the covariance matrix), we see that all coefficients are highly statistically significant. The variance of the unobserved heterogeneity in self-employment is much larger than for the other two states. In particular for men the covariances have about the same magnitude as the variance for unemployment or being out of the labour force. For women, the covariances differ more from each other and we see in particular that self-employed individuals are also estimated to be more likely to also be out of the labour force.

4.3. Second stage — dynamic estimation results

The first observation that we can make unequivocally when comparing the dynamic with the static results, is that the dynamic model is preferable for all four models and samples. The likelihood ratio test rejects the null hypothesis that the dynamic factors play no role, i.e. the

[41] In the pooled regression the coefficient is also statistically significant (-0.25 with p-value < 0.01) in the self-employment equation. It is of smaller absolute magnitude compared to the other two equations (-0.88 and -0.89). In other words, bad health a period earlier will increase the relative risk for an individual in all three categories relative to being employed. But an individual is then more likely to be self-employed compared to the other two states.

[42] All signs switch in the model with the EP distance.

lagged labour states are jointly significant (p-value of 0.0000 for all) and in most cases also individually significant. See e.g. the regression results of the dynamic Big-Five factor marker model for women in Table 10 and for men in Table 11 in comparison with the static model results in Table 8 and 9 respectively.[43]

Second, we find that the Big-Five factor markers are jointly statistically significant. In models which replace the big five with the EP distance or the health index, we also find that the EP distance or the health index enter significantly (using LR tests, even at a 0.1% level).[44] Hence, adding either personality traits or information on an individual's health improves the model compared to a model with only the core personal and household characteristics. Comparing the two models with personality traits, Akaike's information criterion suggests that we should choose the model with the Big-Five factor markers over the one with the EP distance for both men and women.[45] In the following, we will therefore focus our discussion on the model including the Big-Five. Considering that the lagged health index does not enter with a statistically significant coefficient in the self-employment equation of the dynamic model either, we also prefer the Big-Five model as it offers us a more detailed explanation for which individuals become self-employed.[46]

Interestingly, when we test for the joint significance of the inverse Mills ratio we fail to reject the null hypothesis that the coefficients are jointly equal to zero at any conventional significance level.[47] The results for the regressions without the Heckman correction are shown in Table 12 for women and Table 13 for men.[48] They indeed do not differ much from the results with the Heckman correction. This suggests that correcting for attrition bias is not essential once we estimate the dynamic model. Otherwise we would expect larger changes in the estimated coefficients between the two specifications.

How do the results change from the static models' once we include dynamic effects? For women, age loses its statistical significance in the self-employment equation while it remains significant at the 5%-level for men. In other words, once we condition on women's past labour state, we no longer find an effect of age on the choice between self-employed or employee. Both the coefficient on household size for women, and on the dummy for having at least one child for men are no longer statistically significant in the self-employment equation either.

With respect to the factor markers, we find that only the fifth factor, openness for experience, remains statistically significant (at the 5%-level) for women. The coefficient still has

[43] The dynamic pooled regression results for the Big-Five factor marker model are in Appendix D, Table 23.

[44] The regression results of the basic model, both for the full sample, as well as with the adjusted sample sizes for the likelihood ratio tests against the basic model, are available upon request.

[45] See the corresponding regression results in Appendix D, Table 25 (pooled), 26 (women), and 27 (men). The coefficient on the EP distance is statistically significant in the self-employment equation in none of the three samples.

[46] Note also, that the two models do not differ much in the estimates for all the other coefficients. Hence, the statements we can make on either hold in general also for the other. The results for the model with the lagged health index can be found in Appendix D, Table 28 (pooled), 29 (women), and 30 (men).

[47] The p-value is 0.188 for women and 0.395 for men. In the pooled sample we can weakly reject the null hypothesis with a p-value of 0.065.

[48] The pooled regression results without the Heckman correction are in Appendix D, Table 24.

Table 10: Dynamic model with Big-Five factor markers, women

	Self-employed	Unemployed	Not in Labour Force
Constant	-6.72***	-6.12***	-4.5***
	(1.0592)	(0.9536)	(0.6593)
Age	0.01	0.03***	0.04***
	(0.0111)	(0.0106)	(0.0078)
Has partner	0.3	-0.65***	0.38**
	(0.2423)	(0.1892)	(0.1699)
Has child	-0.34	-0.22	-0.41**
	(0.3001)	(0.2332)	(0.1926)
Middle education	-0.12	-0.81**	-0.92***
	(0.5824)	(0.3518)	(0.2951)
High education	0.18	-1.55***	-1.44***
	(0.6075)	(0.3824)	(0.3139)
Household size	0.13	0.05	0.06
	(0.1203)	(0.0984)	(0.0816)
F1: extraversion	0.03	-0.06	-0.03
	(0.0974)	(0.0899)	(0.0667)
F2: agreeableness	-0.07	0.04	0.01
	(0.1112)	(0.0916)	(0.0701)
F3: conscientiousness	-0.07	-0.18*	-0.23***
	(0.1161)	(0.1021)	(0.0776)
F4: emotional stability	-0.09	-0.34***	-0.19***
	(0.0926)	(0.0853)	(0.0627)
F5: openness for experience	0.26**	0.25***	0.14*
	(0.1032)	(0.0913)	(0.0728)
Last state: self-employed	4.02***	1.11**	1.21***
	(0.2397)	(0.4859)	(0.2955)
Last state: unemployed	0.11	2.92***	1.69***
	(0.5094)	(0.2832)	(0.218)
Last state: not in LF	0.6**	1.54***	2.09***
	(0.2675)	(0.2003)	(0.1068)
Initial state: self-employed	4.93***	1.42**	2.3***
	(0.5339)	(0.5624)	(0.362)
Initial state: unemployed	3.48***	3.35***	3.18***
	(0.6971)	(0.5318)	(0.4812)
Initial state: not in LF	2.85***	3.18***	4.28***
	(0.3806)	(0.2937)	(0.2389)
Inverse Mills Ratio	-0.59	-0.9	-2.05**
	(1.2873)	(1.3931)	(1.002)
L	2.26***		
	(0.2128)		
	1.13***	1.26***	
	(0.2295)	(0.229)	
	1.26***	1.43***	0.63***
	(0.1802)	(0.1649)	(0.1711)
W	5.1287	2.5575	2.8429
	2.5575	2.8603	3.2239
	2.8429	3.2239	4.0307
Observations: 14435			
Nr. of Individuals: 3267			
Loglikelihood: -6086.96			

Regression including year fixed effects.

(Non-robust) standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.10

Table 11: Dynamic model with Big-Five factor markers, men

	Self-employed	Unemployed	Not in Labour Force
Constant	-7***	-7.74***	-5.11***
	(1.2323)	(1.4295)	(0.9316)
Age	0.03*	0.07***	0.06***
	(0.0139)	(0.0166)	(0.0115)
Has partner	-0.08	-0.36	-0.34
	(0.2575)	(0.3142)	(0.2256)
Has child	-0.28	-0.08	-0.12
	(0.3137)	(0.4079)	(0.2754)
Middle education	-0.07	-0.75*	-1.26***
	(0.5538)	(0.4333)	(0.3313)
High education	0.29	-1.53***	-2.15***
	(0.5674)	(0.492)	(0.3739)
Household size	0.18	-0.07	-0.03
	(0.1304)	(0.185)	(0.1285)
F1: extraversion	0.28**	-0.03	-0.03
	(0.1105)	(0.1276)	(0.0855)
F2: agreeableness	-0.15	0.03	-0.04
	(0.093)	(0.1227)	(0.0822)
F3: conscientiousness	-0.19*	-0.16	-0.19**
	(0.1)	(0.1281)	(0.092)
F4: emotional stability	-0.25**	-0.33***	-0.37***
	(0.104)	(0.116)	(0.0827)
F5: openness for experience	0.24**	0.15	0.08
	(0.1036)	(0.1209)	(0.0848)
Last state: self-employed	4.15***	0.48	0.86**
	(0.2447)	(0.7504)	(0.4071)
Last state: unemployed	0.23	2.69***	1.3***
	(0.5791)	(0.3341)	(0.312)
Last state: not in LF	0.57*	0.59*	1.5***
	(0.3311)	(0.3118)	(0.1862)
Initial state: self-employed	4.75***	2.7***	2.29***
	(0.5798)	(0.7324)	(0.4488)
Initial state: unemployed	2.61***	4.14***	3.76***
	(0.9214)	(0.5955)	(0.5501)
Initial state: not in LF	2.1***	4.03***	4.3***
	(0.5243)	(0.4125)	(0.3502)
Inverse Mills Ratio	0.27	-1.91	-2.1
	(1.4686)	(2.2414)	(1.3447)
L	2.1***		
	(0.2382)		
	1.49***	1.25***	
	(0.3318)	(0.3344)	
	1.18***	1.6***	0.24
	(0.2575)	(0.2017)	(0.4127)
W	4.4261	3.1269	2.4758
	3.1269	3.7779	3.7523
	2.4758	3.7523	4.0018
Observations: 11967			
Nr. of Individuals: 2647			
Loglikelihood: -4066.08			

Regression including year fixed effects.

(Non-robust) standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.10

Table 12: Dynamic model with Big-Five factor markers and no Heckman correction, women

	Self-employed	Unemployed	Not in Labour Force
Constant	-6.98***	-6.56***	-5.49***
	(0.8184)	(0.6488)	(0.4526)
Age	0.01	0.03***	0.05***
	(0.0096)	(0.0082)	(0.0063)
Has partner	0.29	-0.67***	0.35**
	(0.2415)	(0.1886)	(0.1692)
Has child	-0.34	-0.23	-0.43**
	(0.299)	(0.2331)	(0.192)
Middle education	-0.11	-0.8**	-0.9***
	(0.5784)	(0.3503)	(0.2933)
High education	0.18	-1.53***	-1.41***
	(0.6016)	(0.3807)	(0.3122)
Household size	0.13	0.04	0.05
	(0.1201)	(0.0977)	(0.0812)
F1: extraversion	0.03	-0.06	-0.04
	(0.0972)	(0.0895)	(0.0665)
F2: agreeableness	-0.07	0.03	0.01
	(0.1107)	(0.0907)	(0.0698)
F3: conscientiousness	-0.05	-0.14*	-0.14**
	(0.099)	(0.0839)	(0.0653)
F4: emotional stability	-0.1	-0.35***	-0.21***
	(0.0894)	(0.0827)	(0.0609)
F5: openness for experience	0.25**	0.23***	0.11
	(0.1004)	(0.0895)	(0.0713)
Last state: self-employed	4***	1.07**	1.13***
	(0.2327)	(0.4764)	(0.2906)
Last state: unemployed	0.13	2.96***	1.76***
	(0.5087)	(0.2809)	(0.2164)
Last state: not in LF	0.61**	1.56***	2.14***
	(0.2611)	(0.1973)	(0.1042)
Initial state: self-employed	4.92***	1.41**	2.3***
	(0.5324)	(0.5596)	(0.36)
Initial state: unemployed	3.47***	3.35***	3.18***
	(0.693)	(0.5308)	(0.4772)
Initial state: not in LF	2.84***	3.17***	4.27***
	(0.3794)	(0.2931)	(0.2381)
L	2.26***		
	(0.2119)		
	1.12***	1.26***	
	(0.2288)	(0.2282)	
	1.25***	1.43***	0.63***
	(0.1803)	(0.1641)	(0.1701)
W	5.0893	2.5314	2.8172
	2.5314	2.8507	3.2082
	2.8172	3.2082	4.002
Observations: 14435			
Nr. of Individuals: 3267			
Loglikelihood: -6089.36			

Regression including year fixed effects.

(Non-robust) standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.10

Table 13: Dynamic model with Big-Five factor markers and no Heckman correction, men

	Self-employed	Unemployed	Not in Labour Force
Constant	-6.82***	-8.77***	-6.25***
	(0.7965)	(0.9056)	(0.5853)
Age	0.02**	0.08***	0.07***
	(0.0099)	(0.0117)	(0.0087)
Has partner	-0.08	-0.34	-0.32
	(0.2537)	(0.3139)	(0.2259)
Has child	-0.28	-0.07	-0.11
	(0.3129)	(0.4071)	(0.2749)
Middle education	-0.09	-0.62	-1.12***
	(0.5254)	(0.4082)	(0.3191)
High education	0.27	-1.38***	-1.99***
	(0.5303)	(0.4576)	(0.3549)
Household size	0.18	-0.1	-0.07
	(0.1278)	(0.1811)	(0.1269)
F1: extraversion	0.28**	-0.03	-0.03
	(0.1097)	(0.1277)	(0.0855)
F2: agreeableness	-0.14	0.02	-0.05
	(0.0927)	(0.1225)	(0.0823)
F3: conscientiousness	-0.19**	-0.12	-0.15*
	(0.0958)	(0.1119)	(0.0884)
F4: emotional stability	-0.25**	-0.33***	-0.36***
	(0.1026)	(0.116)	(0.0825)
F5: openness for experience	0.24**	0.15	0.08
	(0.1035)	(0.1187)	(0.085)
Last state: self-employed	4.17***	0.36	0.74*
	(0.2254)	(0.7462)	(0.3905)
Last state: unemployed	0.24	2.65***	1.26***
	(0.5737)	(0.3338)	(0.3108)
Last state: not in LF	0.57*	0.66**	1.59***
	(0.3281)	(0.2959)	(0.1751)
Initial state: self-employed	4.75***	2.7***	2.28***
	(0.5798)	(0.7191)	(0.4486)
Initial state: unemployed	2.6***	4.13***	3.76***
	(0.9068)	(0.5922)	(0.5508)
Initial state: not in LF	2.07***	4.04***	4.31***
	(0.5227)	(0.4126)	(0.3508)
L	2.1***		
	(0.2383)		
	1.48***	1.27***	
	(0.3285)	(0.3264)	
	1.17***	1.62***	0.23
	(0.2578)	(0.2006)	(0.4156)
W	4.4139	3.1113	2.454
	3.1113	3.7993	3.7779
	2.454	3.7779	4.0285
Observations: 11967			
Nr. of Individuals: 2647			
Loglikelihood: -4067.57			

Regression including year fixed effects.

(Non-robust) standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.10

a positive sign. We find an increase in the probability for a woman to be self-employed over being an employee by a factor of 1.28 if her score on openness for experience increases by one standard deviation. For men, we also find that the coefficients on the factor markers decrease in magnitude, with the largest change occurring in the fifth factor, translating to a 40% smaller increase in the relative probability for self-employment. Its impact is now also of approximately the same size as for women, and the p-value increases to around 0.023. Furthermore, for men we find a change in the other factor markers too. Agreeableness is no longer statistically significant but emotional stability and conscientiousness are significant at the 5%-level. Their signs are negative and thus opposite to what we would expect based on the arguments underlying the EP distance. Regarding conscientiousness this is, however, not entirely surprising considering that we already saw in section 2.4.3 that the self-employed in the LISS panel are on average scoring lower than employees. Last but not least, the coefficient on extraversion remains statistically significant (also at the 5% level) and has, as expected, a positive sign. This implies that for men, an increase in the score for extraversion by one standard deviation increases the relative probability to be self-employed by a factor of 1.32 compared to being employed.

Looking at the dynamic variables we find the following for both genders: Having been in the same labour state one period earlier increases the relative probability for an individual to be in the same labour state compared to being an employee—i.e. the diagonal in the block of coefficients for lags shows the largest values. This is what we would expect given the transition probabilities in Table 5. The coefficient on lagged self-employment in the self-employment equation stands out as the largest of all, implying stronger state dependence in self-employment than in other labor market states. While we do not find a statistically significant coefficient for lagged unemployment in the self-employment equation this is potentially due to there being very few such transitions as our definition covers a lower share of the unemployed. We find that all coefficients for the lags are positive. I.e. given that an individual is observed to not be an employee (past and initial), they are more likely to end up in any of the other three states rather than in employment. It should also be noted that in terms of the relative size of coefficients, the individual and household characteristics, as well as the personality traits have much less influence on the choice probabilities than the lagged values of an individual's labour state.[49] Finally, we also see that the estimated variance in the unobserved heterogeneity becomes substantially smaller once we include the dynamic variables. The variance in the unobserved heterogeneity for the self-employed is still larger than for the other two states but by a much smaller factor. The same holds for the estimated covariances. Looking at the covariances, we find that self-employed women are more likely to be out of the labour force than the unemployed, whereas the opposite holds for men.

[49] Initial values are also strongly significant with substantial coefficients. This indicates that the individual effects are correlated with the initial observation, as expected.

Table 14: Simulated transition probabilities (in %), men

Labour state	All individuals				45 to 60 year old			
past \ current	0	1	2	3	0	1	2	3
0: employee	93.58	1.59	1.29	3.54	92.15	1.53	1.63	4.69
1: self-employed	7.33	90.11	0.82	1.74	5.95	90.81	1.14	2.1
2: unemployed	22.22	2.78	44.7	30.3	22.18	2.33	43.58	31.91
3: not in labour force	25.03	2.51	8.68	63.77	22.62	2.1	9.71	65.57
Total	74.14	12.99	3.06	9.8	69.01	14.32	3.91	12.76

Based on dynamic model with Big-Five factor markers, 15310 observation pairs (n=2694).
Source: LISS Panel, missing values for personal/household characteristics are extrapolated.

5. Simulations

In order to provide a better understanding of the estimates discussed above, this section presents two different types of simulations. We first show the simulated transition probabilities for the complete LISS panel. Then we show simulated employment paths for benchmark individuals. We use these to illustrate the limitations of the stationarity assumption. We only present simulations based on the model specification with the Big-Five factor markers.

5.1. Transition probabilities

When simulating the transition probabilities based on our estimates, to correct for attrition, we make the counterfactual assumption that none of the individuals leave the sample. For those who do in reality, we need to impute the values of the covariates. We assume that, apart from age, other personal and household characteristics remain the same.[50] The missing Big-Five factor marker values are completed with the same mean values used to fill in for the initial gaps as described in section 2.4.1. We then, for each individual i, draw one vector of unobserved heterogeneity components μ_i from the multivariate normal distribution (with mean zero and covariance matrix given by $\hat{L}\hat{L}^{\intercal}$). In each of the time periods for each individual and labour state, we then also draw independent error terms ϵ_{ijt} from a Type 1 extreme value distribution. Taking the first labour state that we observe for an individual as given, we then simulate individuals' labor market state outcomes for the following time periods.

Table 14 shows the simulated transition probabilities for men and Table 15 shows the results for women, based on the dynamic model with Big-Five factor markers and no Heckman correction.[51] The left panels show the results for the whole sample, and the right panels show the results for the 45 to 60 year old individuals in the sample. Comparing the total shares of

[50] This assumption is not too farfetched. All the variables have relatively small within variation.

[51] We choose the model without Heckman correction since we could not reject the null hypothesis that the correction terms have coefficient zero in all equations. Hence we consider the model without correction as the "better" model. This is also reflected when we simulate shares for the model with correction (available upon request) and find that it fares less well at replicating the observed total shares, as well as the CBS population shares.

Table 15: Simulated transition probabilities (in %), women

Labour state	All individuals				45 to 60 year old			
past \ current	0	1	2	3	0	1	2	3
0: employee	91.61	1.57	1.34	5.48	89.98	1.5	1.62	6.9
1: self-employed	8.18	87.54	0.81	3.47	7.12	86.97	1.33	4.58
2: unemployed	18.39	1.61	40.65	39.35	16.9	1.94	38.78	42.38
3: not in labour force	15.72	1.82	6.81	75.66	14.47	1.83	6.22	77.48
Total	66.72	8.87	3.68	20.73	60.63	8.91	4.17	26.29

Based on dynamic model with Big-Five factor markers, 19138 observation pairs (n=3325).
Source: LISS Panel, missing values for personal/household characteristics are extrapolated.

all labour states in our sample with both the initially observed shares and the CBS population shares, as shown in Table 5, we find that the simulations for the whole sample lead to results that are similar to what we observed initially in the data.

Still continuing with the total shares, we also see that our simulations, in particular for women, overestimate the probability that individuals remain self-employed. As a consequence the transitions out of self-employment into other labour states are underestimated but the general pattern seen in the data is nevertheless reproduced by our simulations. Overall, it looks like we fare slightly better for men than for women in terms of replicating the observed transition probabilities from one labour state into other states.

The transition probabilities presented in Table 14 and 15 in the left panels are aggregated values. If we however want to understand whether studies are limited by the assumption of stationarity in labour states, we want to know for which individuals the probability to remain in the same state is lowest—i.e. for whom the stationarity assumption is most likely violated. We can make some observations on this by looking at the right panels with the results for 45 to 60 year old individuals. These show that when we consider older individuals separately, we observe approximately the same share as in the whole sample for individuals who will not remain in self-employment. Hence, even for the older individuals, where one generally assumes that projections are less prone to errors due to the smaller time horizon on which forecasts are made, we find that the stationarity assumption has its limitations.

5.2. Individual simulations

The aim of the following simulations is to help us understand whether and how individuals switch between labour states over the course of time, and how this varies with individual characteristics, in particular with personality. By simulating the employment paths for the chosen benchmark individuals, we can illustrate how the estimates and transition probabilities discussed earlier translate into individual probabilities of remaining in the same labour state. A simple approach would be to take the corresponding probability of remaining in self-employment to the power t and conclude for e.g. a horizon of ten years that 38.14% of the men and 24.76% of the women remain self-employed. Such an approach would however

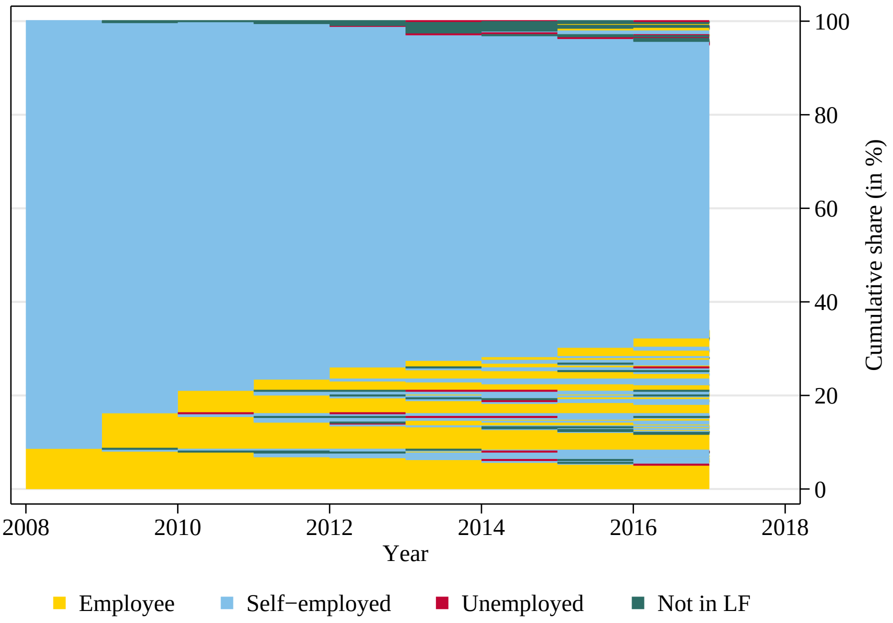


Figure 6: Simulated employment paths: male, self-employed in 2007 (median characteristics)

ignore that the probability of each labour state is affected by the given characteristics of the individual, some of which change over time (such as age).

We therefore simulate the employment paths for a benchmark individual of each gender over ten years starting from 2008. We fix all benchmark individuals' age to 45 years at the start of the simulation. We then choose the other personal and household characteristics such that they reflect the median for each combination of labour state and gender. We then assign these according to the labour state that our benchmark individuals have in 2007. All benchmark individuals have a partner, and one child.[52] For education we choose the level with the highest frequency. This results in all individuals having medium education, except for the self-employed male to whom we assign a high education level. We then take the median value for the Big-Five factor markers within the subsample of individuals observed in the LISS sample that have the same personal characteristics (for age ranging from 45 to 54 years) and labour state in which the benchmark individual starts in. The values we choose are reported in Table 20 in Appendix D.

Next we draw five hundred times the benchmark individual's unobserved heterogeneity vector μ_i, and then proceed in the same way as with the simulations for the transition probabilities – by forward simulating using the same procedure, and drawing once in each of the time periods for each of paths and labour states the ϵ_{ijt} from a Type 1 extreme value distribution. We do this exercise for benchmark individuals of each gender that are in

[52] We keep those values constant accross time for simplicity.

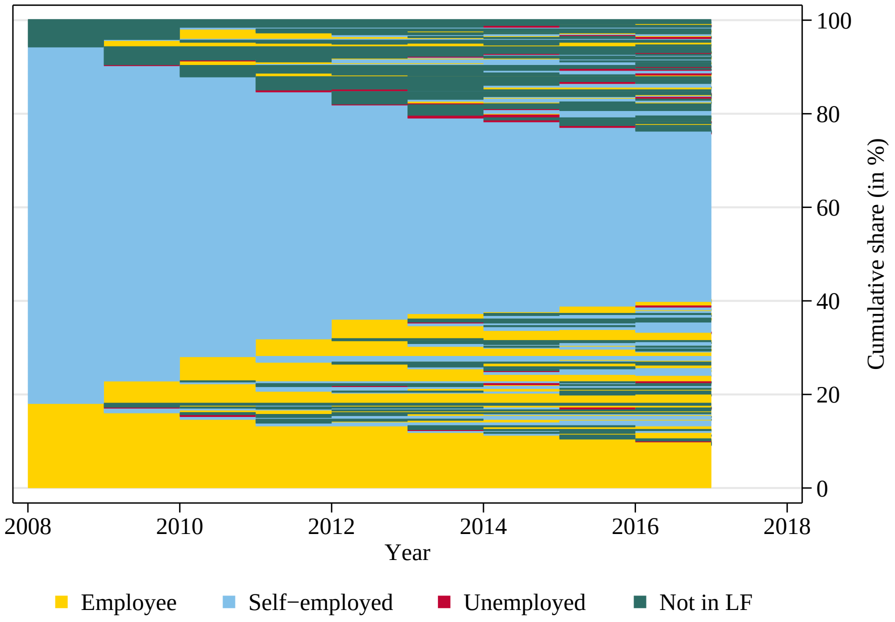


Figure 7: Simulated employment paths: female, self-employed in 2007 (median characteristics)

employment or self-employed in 2007. In the following we will however limit our discussion to those who start of as self-employed.[53] The resulting employment paths for the self-employed benchmark individuals are shown as sequence index plots in Figure 6 for the male benchmark person and in Figure 7 for the female benchmark.

The sequence index plots show all simulated employment paths. The paths are stacked vertically on top of each other and each path is a (thin) horizontal line on which each year is coloured according to the labour state the path takes. We order the paths by labour states starting with the labour state in the first period, followed by the second period, etc. The scale on the y-axis is adjusted to reflect cumulative shares (in percent) instead of the number of paths.

Based on the sequence index plots, we can then directly determine the approximate share of individuals who remain in self-employment for all time periods. Hence, if we were to use this strict definition for stationarity, we would conclude based on Figure 6 that our benchmark male has a chance of approximately 60% of remaining self-employed throughout all the ten years. The outcome is different for the female individual. Figure 7 suggests that she has only about 35% probability of remaining self-employed for all ten years.

Are these the probabilities we need in our comparison with static microsimulation models? We view them as lower bounds, because individuals can also return to self-employment: Looking more closely at the two sequence index plots, we can see employment paths on

[53] The figures for employees are available upon request.

which the individuals first leave and then return to self-employment after they have spent one period or more in a different labour state. Thinking of our initial motivation, if we are concerned about pension savings, which we expect to be lower for the self-employed than for employees, these paths are close to "stationary", because if an individual spends most time periods in self-employment, the occupational pension savings accumulated during a short period in employment will be small. A static microsimulation will then be approximately right in setting them to zero. We therefore also relax our definition of "stationarity" in order to attain an upper bound of the probability to stay in self-employment. For this we count all paths where at least half of the time, i.e. a minimum of five years, is spent in self-employment. We then find a probability 80% for the male benchmark individual and a probability of 58.6% for the female benchmark to remain self-employed.

Two additional observations can be made when comparing the two sequence index plots. First, the probability for self-employment is smaller for the female benchmark individual at any point in time. Second, the probability for self-employment in 2008 is around ten percent smaller than what we would guess if we naively assumed the 86.97% from the right panel in Table 15, whereas it is approximately on point for the male individual. Hence, one may ask whether we chose a female benchmark individual who is not "suited" to self-employment, and whether a more "suited" individual would have led to larger shares.

We can answer these questions partially with the help of our model. While we cannot use the model to find the "most suited" individual, we can use it to look at informed variations of the benchmark individuals. I.e. we can select one or more characteristics that are not common to both individuals and see how the probabilities change as a function of the difference in these characteristics. One obvious choice is to change the education level and study how the share of paths spent in self-employment changes. Considering that the coefficient for high education in the self-employment equation is positive, while the one for medium education is negative, we expect the chances to remain in self-employment of the female individual (where the benchmark has medium education) to become higher and those of the male (where the benchmark has high education) to become lower once we exchange the education levels. This is indeed the case and we find the lower and upper bound of (45.0, 67.8) for the female individual with high education and (48.6, 72.2) for the male with medium education.

Furthermore, we are also interested how personality affects individual probabilities to remain self-employed. We follow the arguments by Obschonka et al. (2013) to define an "entrepreneurial" profile. Such a profile should include high levels in extraversion, conscientiousness, emotional stability, and openness, and low levels in agreeableness. Taking the standardization of the factor markers into account we assign values of 2.5 and -2.5 respectively. In addition, we also define a "non-entrepreneurial" profile that takes the opposite values. Figure 8 shows the sequence index plots for a male benchmark individual with high education and Figure 9 for a female benchmark individual with medium education. (See Table 20 for all variables underlying the plots.)

As expected, we find that the benchmark individuals with "entrepreneurial" Big-Five

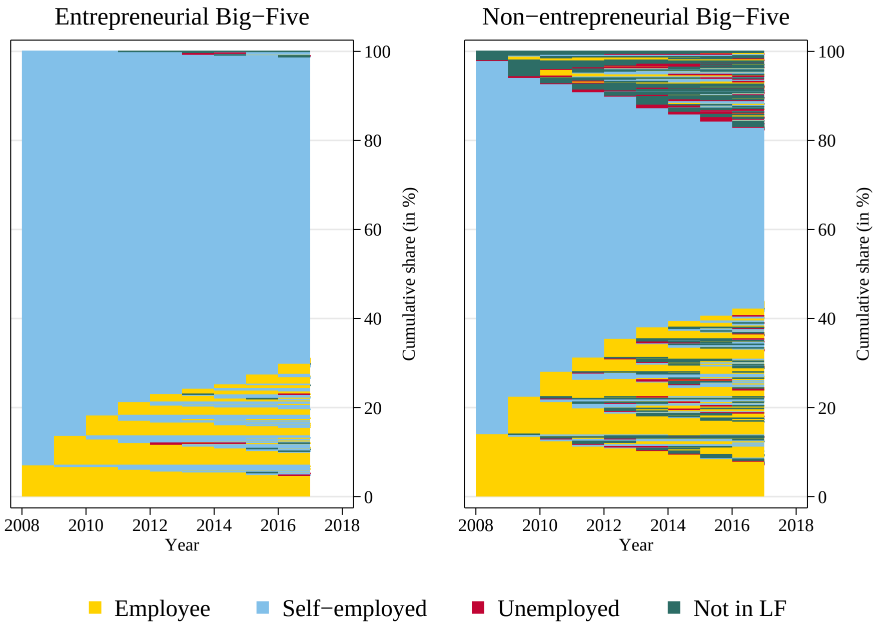


Figure 8: Simulated employment paths: male, self-employed in 2007

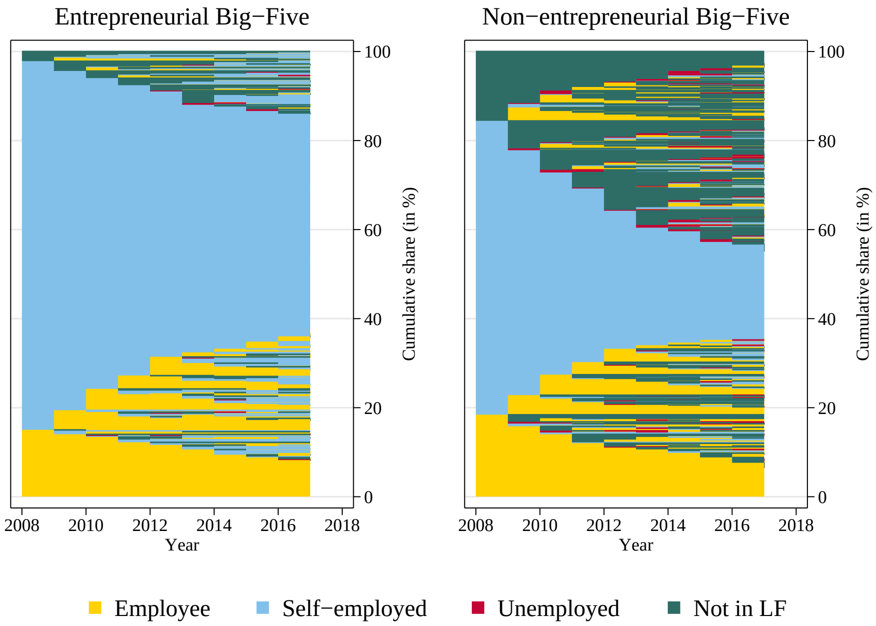


Figure 9: Simulated employment paths: female, self-employed in 2007

Table 16: Probability (in %) that different benchmark individuals who are self-employed in 2007, are self-employed in later years

Big-Five	in 2008	in 2009	in 2012	in 2017	10 years	5+ years
Male: high education						
median	91.4	83.8	77.6	72.0	61.0	80.0
entrepreneurial	93.0	86.8	82.2	77.6	67.6	84.4
non-entrepreneurial	83.8	72.2	60.4	51.4	38.4	62.4
—— medium education						
median	88.2	78.6	69.4	63.4	48.6	72.2
entrepreneurial	90.8	83.0	77.2	70.8	59.4	79.2
non-entrepreneurial	77.8	62.6	43.8	32.0	22.2	47.0
Female: high education						
median	80.4	75.0	59.8	64.0	45.0	67.8
entrepreneurial	86.4	81.8	73.2	74.2	59.2	78.4
non-entrepreneurial	73.6	66.2	45.6	47.0	29.0	54.2
—— medium education						
median	76.2	68.6	51.6	54.0	35.4	58.6
entrepreneurial	82.8	77.2	64.0	69.0	49.2	72.0
non-entrepreneurial	66.0	56.4	33.4	34.0	19.6	39.4

Based on dynamic model with Big-Five factor markers and no Heckman correction (n=500). The benchmark individual is self-employed in 2007. The rightmost column gives the probability that the individual is self-employed in at least 5 of the 10 years.

factor markers have a larger probability to remain self-employed than the median, or "non-entrepreneurial" ones. The latter in turn are less likely to remain in self-employment than the benchmark individuals with median Big-Five factor markers. In addition, we find that the change in the probabilities, both in the lower and upper bound, is larger between the "non-entrepreneurial" individuals in comparison to the median than for the "entrepreneurial". This is in line with the Big-Five factor markers that we chose for the different benchmark individuals as the magnitude of the differences are also larger between the median values and the "non-entrepreneurial" profile. The same relationship can also be found again in the difference between the lower and upper bound.[54] Table 16 summarises the simulation results for each combination of education level and personality profile. In the two rightmost columns, the table shows the lower and upper bounds based on our earlier definitions. The other columns show the share of employment paths in self-employment at a specific point in time.[55]

Looking at the results from the simulations we also find the following general patterns. The probability of self-employment is always lower for a female individual in comparison to a male

[54] The only exception is in the case of the female benchmark with medium education where the difference in the lower and upper probability bound is actually smaller for the "non-entrepreneurial" profile.

[55] See Figure 11 in Appendix E for a graphic representation of the evolution of the shares over time and type of benchmark individual.

benchmark individual. This includes the cases where both have exactly the same characteristics. Furthermore, we find that a shift from high education to medium education increases the difference in probabilities between the "entrepreneurial" and "non-entrepreneurial" case. In addition to this, the difference between the lower and upper bound (the final two columns) within each case also increases. Finally, we also find that the share of paths observed in self-employment in a year stabilizes after the fifth to sixth year in the simulation.

How should these results then be interpreted in relation to a static microsimulation for wealth and pension income modeling? First, it is important to keep in mind that all calculated probabilities are specific to the benchmark individuals and their characteristics, and are not representative for the population at large. Nevertheless, they still inform us on the potential limitations of assuming a static framework. If, for example, we take the most stable of the benchmark individuals, the male with high education and "entrepreneurial" Big-Five factor marker levels as shown in the left panel of Figure 8, then we still find that he has a chance of 12.6% to spend at least seven of the next ten years as an employee.[56] Within the context of the Dutch pension system, such an individual would accumulate pension savings in the second pillar during at least seven of the ten years—and therefore likely end up in a better financial position than what a model under the stationarity assumption would predict. That is, a static microsimulation may over-predict the share of the self-employed who have too low pension savings.

Last, a different source for limitations of the stationarity assumption could also arise if individuals' probabilities to remain self-employed are negatively correlated with their pension savings. If we think again of the reasoning behind the Big-Five factor markers, it seems plausible that a "non-entrepreneurial" individual who is nevertheless self-employed, will probably not be very successful as an entrepreneur. Consequently their savings would be relatively low but they would also have a relative low probability of remaining in self-employment, as shown in the right panel of Figure 8. We therefore think that it is important that future research on projecting pension wealth for a heterogeneous population also looks at how employment paths and their patterns relate to different employment states and (pension) savings.

6. Conclusion

We estimate dynamic multinomial models to explain transitions into and out of self-employment and other labor market states for individuals of working age in the Netherlands. Our sample consists of 25 to 60 year old individuals that participate in the LISS panel, a representative sample of adult individuals in the Netherlands. We find that the inclusion of a lagged health index or personality traits improves the model compared to a baseline that only includes individual characteristics, such as age, gender, or education, and household characteristics,

[56] Because of the different scaling of Figure 6 and 8, it might look like this share is larger in the latter's left panel. It is not though. The median case counts 15.2% of the paths with at least seven periods in employment.

such as household size, the presence of a partner and/or child. Furthermore, we also find that the Big-Five factor markers, in particular for men, help to explain transitions into and out of self-employment. Adding all Big-Five factor markers individually is found to provide a better fit than using the Entrepreneurship-Prone Big Five Profile by Obschonka et al. (2013), a distance measure built from the factors.

We also find that expanding the static model to a dynamic framework reduces the variance in the unobserved heterogeneity substantially. The estimated variance in the unobserved heterogeneity for the self-employed is larger than for the other states as well as the covariances. The results show that self-employed women are more likely to leave the labour force than the unemployed, whereas the opposite holds for men.

Using simulations we then show that the probability to remain self-employed in the next year is on average around 90% for men and slightly lower for women. Simulated employment paths for multiple years for benchmark individuals are used to illustrate the limitations of the stationarity assumption that is common in wealth and pension income modeling. Based on the simulated employment paths we show that we can at best expect that the benchmark self-employed male has an 80% chance to spend the majority of the next ten years in self-employment. This probability falls to 62% if the individual does not have an entrepreneurial personality, on even to 47% if, in addition, he has medium education rather than high education (the benchmark). For the self-employed benchmark woman, the chances to spend the majority of the next ten years in self-employment are only 59%, and this falls to below 40% if the woman does not have an entrepreneurial personality.

With the ongoing pension reforms, there is a lot of recent interest in pension adequacy of a heterogeneous population, with a focus on vulnerable groups such as the self-employed. Our results suggest that future work on projecting pension incomes and pension adequacy should account for the labour market dynamics and the transitions between states in which individuals do or do not (sufficiently) accumulate pension wealth. Combining the type of model and dynamic simulations here with administrative data on how much pension wealth is accumulated in a given labor market state, seems a fruitful avenue for future research.

References

Albarrán, P., R. Carrasco, and J. M. Carro (2015). Estimation of dynamic nonlinear random effects models with unbalanced panels.

Been, J. and M. Knoef (2013). The necessity of self-employment towards retirement: evidence from labor market dynamics and search requirements in unemployment insurance. In *28th Annual Congress of the European Economic Association.*

Beugelsdijk, S. and N. Noorderhaven (2005). Personality characteristics of self-employed; an empirical study. *Small Business Economics 24*(2), 159–167.

Bhat, C. R. (2001). Quasi-random maximum simulated likelihood estimation of the mixed multinomial logit model.

Bhat, C. R. (2003). Simulation estimation of mixed discrete choice models using randomized and scrambled Halton sequences.

Blanchflower, D. G. (2000). Self-employment in OECD countries. *Labour economics 7*(5), 471–505.

Bolhaar, J., A. Brouwers, and B. Scheer (2016). De flexibele schil van de Nederlandse arbeidsmarkt: een analyse op basis van microdata. *CPB Achtergronddocument.*

Borghans, L., A. L. Duckworth, J. J. Heckman, and B. Ter Weel (2008). The economics and psychology of personality traits. *Journal of human Resources 43*(4), 972–1059.

Bosch, N. (2014). Succes als startende zelfstandige. *CPB Achtergronddocument.*

Bosch, N., G. Roelofs, D. Van Vuuren, and M. Wilkens (2012). De huidige en toekomstige groei van het aandeel zzp'ers in de werkzame beroepsbevolking. *CPB Achtergronddocument.*

Buddelmeyer, H. and M. Wooden (2011). Transitions out of casual employment: the Australian experience. *Industrial Relations: A Journal of Economy and Society 50*(1), 109–130.

CBS (2014, December). Achtergrondkenmerken en ontwikkelingen van zzp'ers in Nederland. cbs.nl.

Cheng, T. C. and P. K. Trivedi (2015). Attrition bias in panel data: a sheep in wolf's clothing? a case study based on the Mabel survey. *Health economics 24*(9), 1101–1117.

Cobb-Clark, D. A. and S. Schurer (2012). The stability of Big-Five personality traits. *Economics Letters 115*(1), 11–15.

de Bresser, J. and M. Knoef (2015). Can the Dutch meet their own retirement expenditure goals? *Labour Economics 34*, 100–117.

Goldberg, L. R. (1992). The development of markers for the Big-Five factor structure. *Psychological Assessment 4*(1), 26–42.

Gong, X., A. Van Soest, and E. Villagomez (2004). Mobility in the urban labor market: a panel data analysis for Mexico. *Economic Development and Cultural Change 53*(1), 1–36.

Heckman, J. J. (1976). The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement, Volume 5, number 4*, pp. 475–492. NBER.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica 47*(1), 153–161.

Heckman, J. J. (1981a). The incidental parameters problem and the problem of initial conditions in estimating a discrete time-discrete data stochastic process. In C. F. Manski and D. L. McFadden (Eds.), *Structural Analysis of Discrete Data and Econometric Applications*, pp. 179–195. MIT Press, Cambridge, MA.

Heckman, J. J. (1981b). Statistical models for discrete panel data. In C. F. Manski and D. L. McFadden (Eds.), *Structural Analysis of Discrete Data and Econometric Applications*, pp. 114–178. MIT Press, Cambridge, MA.

Jürges, H. (2007). True health vs response styles: exploring cross-country differences in self-reported health. *Health economics 16*(2), 163–178.

Kabátek, J. (2015). Happy birthday, you're fired! the effects of an age-dependent minimum wage on youth employment flows in the Netherlands.

Knoef, M., J. Been, R. Alessie, K. Caminada, K. Goudswaard, and A. Kalwij (2016). Measuring retirement savings adequacy: developing a multi-pillar approach in the Netherlands. *Journal of Pension Economics & Finance 15*(1), 55–89.

Mastrogiacomo, M. (2016). De pensioenpuzzel van zelfstandigen. *Netspar Brief* (7).

Mastrogiacomo, M. and R. J. Alessie (2015). Where are the retirement savings of self-employed? An analysis of 'unconventional' retirement accounts. (454).

Obschonka, M., E. Schmitt-Rodermund, R. K. Silbereisen, S. D. Gosling, and J. Potter (2013). The regional distribution and correlates of an entrepreneurship-prone personality profile in the United States, Germany, and the United Kingdom: A socioecological perspective. *Journal of Personality and Social Psychology 105*(1), 104.

Oguzoglu, U. (2016). Disability and multi-state labour force choices with state dependence. *Economic Record 92*(296), 28–46.

Prowse, V. (2012). Modeling employment dynamics with state dependence and unobserved heterogeneity. *Journal of Business & Economic Statistics 30*(3), 411–431.

Rietveld, C. A., H. van Kippersluis, and A. R. Thurik (2015). Self-employment and health: Barriers or benefits? *Health economics 24*(10), 1302–1313.

Train, K. E. (2009). *Discrete choice methods with simulation.* Cambridge University Press.

Verbeek, M. and T. Nijman (1992). Testing for selectivity bias in panel data models. *International Economic Review*, 681–703.

Wooldridge, J. M. (2005). Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity. *Journal of applied econometrics 20*(1), 39–54.

Wooldridge, J. M. (2007). Inverse probability weighted estimation for general missing data problems. *Journal of Econometrics 141*(2), 1281–1301.

Wooldridge, J. M. (2009). Correlated random effects models with unbalanced panels.

Zucchelli, E., M. N. Harris, and X. Zhao (2012). Ill-health and transitions to part-time work and self-employment among older workers. *Available at SSRN 2269449*.

Zwinkels, W., M. Knoef, J. Been, C. Caminada, and K. Goudswaard (2017). Zicht op zzp-pensioen. *Netspar Industry Paper Series: Design Paper 91.*

A. First order derivatives

Let for notational ease $\mathsf{P}(y_{ijtr}) \equiv Prob(y_{ijtr}|X_{it}, y_{it-1}, y_{i0}, \xi_i)$ and similarly for all other probabilities. The simulated loglikelihood function from equation (12) is then given by

$$logS\mathcal{L} = \sum_{i=1}^{N} log \left[\frac{1}{R} \sum_{r=1}^{R} \prod_{t=2}^{T} \left(\prod_{j=1}^{J} \mathsf{P}(y_{ijtr})^{\mathbb{D}_{ijt}} \right)^{\mathbb{S}_{it}} \right]$$

Furthermore let vector θ be a vector with κ elements containing the stacked columns of all the coefficient matrices $\beta, \gamma, \delta,$ as well as a vector of all the elements of the Cholesky factorisation $[l_{11}, l_{21}, l_{22}, l_{31}, l_{32}, l_{33}]'$. And lastly, let V_{it} be a vector of the covariates, where X_it, y_{it-1}, y_{i0} are stacked accordingly to the elements in θ. Note that e.g. for a covariate in X_{it} its index κ in V_{it} is jointly defined by k and j. The last rows in V_{it} are given by the vector $[\xi_1, \xi_1, \xi_2, \xi_1, \xi_2, \xi_3]'$.

Then using section 3.1.2 the first order derivative can be written as

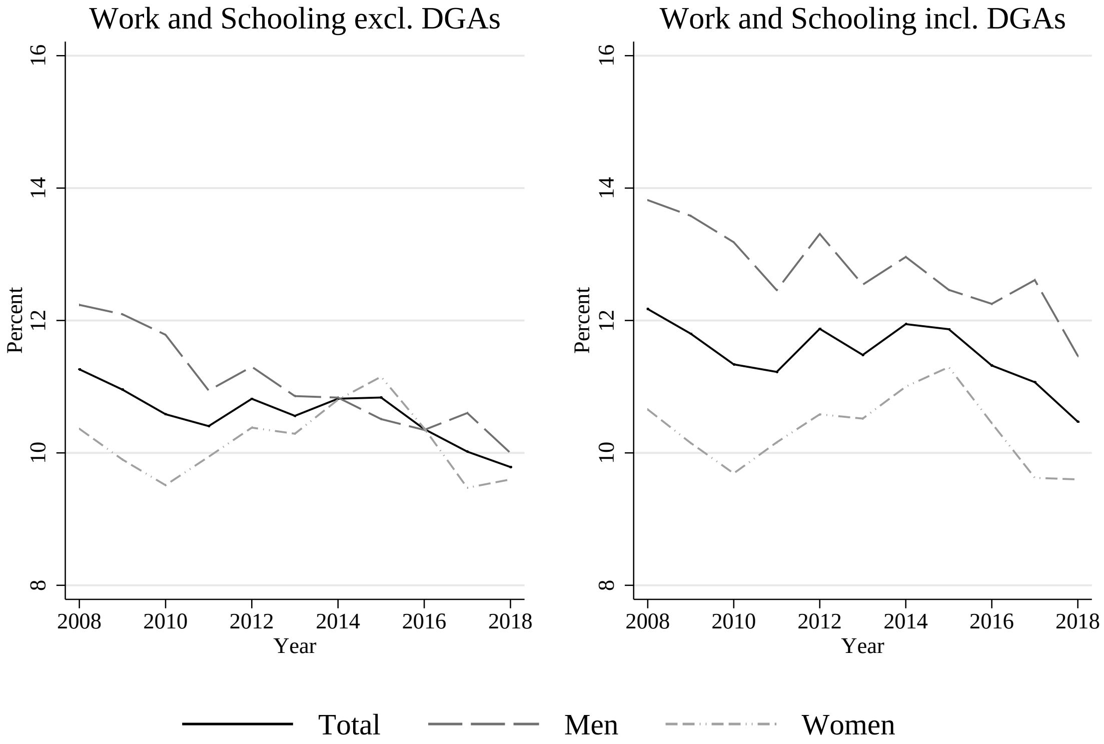
$$\begin{aligned} \frac{\partial logS\mathcal{L}}{\partial \theta_\kappa} &= \sum_{i=1}^{N} \left[\frac{1}{\mathsf{P}(y_i)} \left(\frac{1}{R} \sum_{r=1}^{R} \mathsf{P}(y_{ir}) \left(\sum_{t=1}^{T} \mathbb{S}_{it} \frac{\mathsf{P}(y_{itr})'}{\mathsf{P}(y_{itr})} \right) \right) \right] \\ &= \sum_{i=1}^{N} \left[\frac{1}{\mathsf{P}(y_i)} \left(\frac{1}{R} \sum_{r=1}^{R} \mathsf{P}(y_{ir}) \left(\sum_{t=1}^{T} \mathbb{S}_{it} \left[\mathbb{D}_{i\kappa t} - \mathsf{P}(y_{i\kappa tr}) \right] V_{i\kappa t} \right) \right) \right] \end{aligned}$$

B. DGAs and the self-employed

B.1. DGAs in the work and schooling survey

We use question cwxxx121, where "xxx" denotes the indicator for the year and wave, to classify individuals within the working population. The question's text explains to the survey takers that *[a] director of a limited liability or private limited company (Dutch: NV or BV, respectively) is generally on the payroll of that company. In that case, please enter that you [are an] employee in permanent or temporary employment. A majority shareholder director, also, generally receives an income as an employee. Nevertheless, if this [applies] to you, we request that you indicate that you [are] a (majority shareholder) director.*[57] Hence, self-employed individuals who have incorporated their company and act as its DGA should answer that they are DGAs, whereas e.g. the directors of Shell or Unilever should answer with the option that they are *director of a limited liability or private limited company.* Therefore, including DGAs based on question cwxxx121 in the self-employment definition should not lead to a mistake in our definition of the self-employed.

[57] Taken from the English version codebook of the Work and Schooling core study wave 10.



Data source: LISS Panel − Work & Schooling, Income, Background Variables

Figure 10: Comparison of self-employment shares with and without DGAs in raw sample

Figure 10 illustrates this further. By comparing the left and the right panel we see that the inclusion of DGAs shifts the shares more or less equally across all periods, which is consistent with evidence presented in CBS (2014, p.13, figure 3.2.1). It does so mostly through increasing the share of self-employed among the working men, which is consistent with the finding reported by CBS in the same report that 80% of DGAs were men in 2012.[58]

B.2. DGAs in the income survey

In the income survey, question cixxx008 individuals are asked: *Did you receive income as employee in [t-1]?* including the explanation that *A [DGA] generally receives income as an employee as well and so please answer YES here.* The question block on self-employment on the other hand makes no special mention of DGAs. The section is however vague when asking invididuals *[w]hich work situation as described [. . .] in [$t-1$] applied to you (or for a part of [$t-1$])?* and offering statements such as *work as an entrepreneur or as a freelancer (alongside a job)*, or *a company owner*, or *make profit (or losses) through enterprise in some*

[58] The shares presented in Figure 10 are calculated for all individuals from 25 to 60 years of age. The sample is further restricted to those for whom the covariates of the regression analysis are not missing. It is not conditional on observations being available for at least two periods consecutively.

other way [...] in questions cixxx37 to cixxx044. DGAs are therefore only included in our definition of the self-employed if they self-identify as one of these options, otherwise they will be categorised as employees.

It is theoretically possible to identify DGAs through their wages because there is a legal requirement for a DGA to earn at least 45000 euro (gross). This minimum salary for DGA is substantially higher than the CPB's estimate of model income at 34000 euro for 2017[59] and we could therefore use reported income as an identification strategy. However, many individuals do not report their gross income in the survey, and the income brackets provided are not indicative enough as one bracket runs from 36000 to 48000.

C. Health index

Table 17: Regression results for health index

	Regression sample only	incl. single observations
	Health in general	Health in general
Health compared to $t-1$: poor	-1.555***	-1.536***
	(0.0692)	(0.0674)
— moderate	-0.585***	-0.581***
	(0.0238)	(0.0231)
— very good	-0.0120	-0.0166
	(0.0225)	(0.0217)
— excellent	0.455***	0.421***
	(0.0443)	(0.0421)
Hinder in daily life (index)	-0.290***	-0.282***
	(0.0114)	(0.0110)
Long-standing disease	0.264***	0.279***
	(0.0210)	(0.0203)
Regularly pain in joints	-0.101***	-0.107***
	(0.0200)	(0.0193)
— Hearth problems	-0.185***	-0.194***
	(0.0471)	(0.0456)
— Breathing problems	-0.166***	-0.163***
	(0.0362)	(0.0350)
— Coughing, flu, etc.	-0.208***	-0.194***
	(0.0217)	(0.0210)
— Stomach/intestinal problems	-0.207***	-0.210***
	(0.0249)	(0.0241)

[59] See https://www.cpb.nl/publicatie/macro-economische-verkenning-mev-2019

Table 17: Regression results for health index (continued)

	Regression sample only	incl. single observations
	Health in general	Health in general
— Headaches	-0.107***	-0.113***
	(0.0209)	(0.0202)
— Fatigue	-0.253***	-0.245***
	(0.0198)	(0.0191)
— Sleeping problems	-0.0814***	-0.0864***
	(0.0219)	(0.0213)
Other recurrent complaints	-0.249***	-0.244***
	(0.0259)	(0.0251)
No recurrent complaints	0.119***	0.113***
	(0.0231)	(0.0223)
Taking medicine for:		
— high blood pressure	-0.123***	-0.131***
	(0.0291)	(0.0283)
— diabetes	-0.311***	-0.327***
	(0.0507)	(0.0495)
— joint pain/infection	-0.166***	-0.173***
	(0.0355)	(0.0345)
— hormonal osteoporosis	-0.371***	-0.325**
	(0.137)	(0.135)
Not taking any medicine	0.231***	0.214***
	(0.0252)	(0.0243)
Hospital stay	-0.0646**	-0.0749***
	(0.0256)	(0.0248)
Female	0.0583***	0.0653***
	(0.0155)	(0.0151)
Age	-0.00781***	-0.00756***
	(0.000829)	(0.000798)
Underweight	-0.140**	-0.0958*
	(0.0567)	(0.0542)
Overweight	-0.162***	-0.170***
	(0.0163)	(0.0158)
Obese	-0.343***	-0.342***
	(0.0233)	(0.0226)
Constant cut1	-4.397***	-4.305***
	(0.123)	(0.120)

Table 17: Regression results for health index (continued)

	Regression sample only Health in general	incl. single observations Health in general
Constant cut2	-1.964***	-1.888***
	(0.115)	(0.112)
Constant cut3	0.725***	0.778***
	(0.114)	(0.112)
Constant cut4	1.960***	2.013***
	(0.115)	(0.112)
Depression, anxiety controls	Yes	Yes
Difficulty with actions controls	Yes	Yes
Other medication controls	Yes	Yes
Education controls	Yes	Yes
Year fixed effects	Yes	Yes
Observations	28926	30669

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

D. Additional tables

Table 18: Observed transition probabilities (in %) excluding/including corrections

Labour state	Excluding corrections				Including corrections			
past \ current	0	1	2	3	0	1	2	3
0: employee	93.34	1.35	1.07	4.24	93.06	1.55	1.16	4.22
1: self-employed	7.93	87.76	0.65	3.66	8.68	87.28	0.60	3.44
2: unemployed	18.09	2.30	46.38	33.22	21.01	2.07	43.64	33.28
3: not in labour force	16.80	1.94	6.17	75.09	16.96	1.89	6.34	74.81
overall	69.92	10.30	3.03	16.75	69.94	10.75	3.03	16.29

Based on pooled samples with 24061 and 265510 observation pairs respectively (n=6019).

Source: LISS Panel, own calculations.

Table 19: Simulated transition probabilities (in %), pooled

Labour state	without Heckman correction				with Heckman correction			
past \ current	0	1	2	3	0	1	2	3
0: employee	92.43	1.57	1.40	4.59	91.45	1.56	1.49	5.49
1: self-employed	7.32	88.55	0.91	3.22	6.96	87.96	1.00	4.08
2: unemployed	21.56	1.88	43.13	33.43	20.77	1.64	40.29	37.29
3: not in labour force	16.92	2.06	6.95	74.06	16.3	2.03	6.25	75.42
Total	69.32	10.52	3.49	16.67	66	10.20	3.56	20.24

Based on dynamic model with Big-Five factor markers, 34448 pairs (n=6019).
Source: LISS Panel, missing values for personal/household characteristics are extrapolated.

Table 20: Personal characteristics for individual simulation

	Employee		**Self-employed**					
	Female	Male	Female			Male		
	median	median	median	high	low	median	high	low
Age in 2008	45	45	45			45		
Partner	Yes	Yes	Yes			Yes		
Nr. of children	1	1	1			1		
Medium education	Yes	Yes	Yes			No		
High education	No	No	No			Yes		
F1	0.0018	0.0913	0.3166	2.5	-2.5	0.2073	2.5	-2.5
F2	0.4373	-0.5206	0.2526	-2.5	2.5	-0.4403	-2.5	2.5
F3	0.0217	0.1470	0.9095	2.5	-2.5	-0.1899	2.5	-2.5
F4	0.0722	0.3549	-0.3043	2.5	-2.5	0.5378	2.5	-2.5
F5	-0.3563	-0.2132	-0.3008	2.5	-2.5	1.1217	2.5	-2.5

Note: “high” denotes the “entrepreneurial” and “low” the “non-entrepreneurial” benchmark individual.

Table 21: First stage Heckman regression results, pooled

Model	Baseline	Health index	EP distance	Big5 Factors
Self-employed in $t-1$	-0.18***	-0.15***	-0.16***	-0.15***
	(0.034)	(0.033)	(0.033)	(0.033)
Unemployed in $t-1$	0.02	0.06	0.03	0.04
	(0.069)	(0.069)	(0.068)	(0.068)
Not in labour force in $t-1$	0.08**	0.09**	0.09**	0.10**
	(0.032)	(0.032)	(0.032)	(0.032)
Age	0.02***	0.02***	0.02***	0.02***
	(0.001)	(0.001)	(0.001)	(0.001)
Female	-0.05	-0.08	-0.04	-0.08
	(0.047)	(0.045)	(0.046)	(0.046)
Has partner	0.03	0.00	0.04	0.03
	(0.046)	(0.045)	(0.045)	(0.045)
Has child	-0.05	-0.06	-0.04	-0.05
	(0.049)	(0.048)	(0.048)	(0.048)
Female x partner	-0.04	-0.02	-0.05	-0.05
	(0.057)	(0.054)	(0.055)	(0.055)
Female x has child	0.03	0.04	0.04	0.05
	(0.049)	(0.047)	(0.047)	(0.047)
Middle eduction	0.14*	0.12*	0.12*	0.12*
	(0.060)	(0.058)	(0.059)	(0.059)
High education	0.17**	0.14*	0.14*	0.15*
	(0.061)	(0.060)	(0.060)	(0.061)
Household size	-0.03	-0.02	-0.03	-0.02
	(0.016)	(0.016)	(0.016)	(0.016)
Days until answered	-0.01***	-0.01***	-0.01***	-0.01***
(within call)	(0.002)	(0.002)	(0.002)	(0.002)
Answered in first call	0.26***	0.27***	0.25***	0.25***
	(0.036)	(0.036)	(0.036)	(0.036)
Interaction term	-0.01**	-0.01**	-0.01**	-0.01**
days x first call	(0.002)	(0.002)	(0.002)	(0.002)
F1: extraversion				-0.01
				(0.012)
F2: agreeableness				-0.00
				(0.013)
F3: conscientiousness				0.09***
				(0.013)
F4: emotional stability				-0.02
				(0.012)
F5: openness for experience				-0.03*
				(0.013)
Health index		0.01		
		(0.010)		
EP distance			-0.00	
			(0.002)	
Constant	0.34***	0.44***	0.46***	0.47***
	(0.090)	(0.087)	(0.096)	(0.089)
Observations	30,692	30,183	30,354	30,354
Number of individuals	7,287	6,907	6,973	6,973
ρ	0.170	0.107	0.126	0.124
σ_u	0.452	0.346	0.379	0.376
LL	-11614	-11026	-11212	-11183

Dependent variable: Indicator variable for labour state observed in t.
Standard errors in parentheses; *** p<0.001, ** p<0.01, * p<0.05

Table 22: Static model with Big-Five factor markers, pooled

	Self-employed	Unemployed	Not in Labour Force
Constant	-19.65***	-7.79***	-4.15***
	(0.9777)	(0.8237)	(0.628)
Age	0.17***	0.09***	0.09***
	(0.0124)	(0.0103)	(0.0085)
Female	-1.22***	0.61**	0.63**
	(0.4002)	(0.2832)	(0.2654)
Has partner	0.21	-0.81***	-1.04***
	(0.3082)	(0.2764)	(0.2508)
Has child	-1.55***	-0.96***	-1.06***
	(0.2993)	(0.292)	(0.2435)
Female x partner	-0.39	0.15	1.77***
	(0.3974)	(0.3123)	(0.2928)
Female x has child	1.26***	0.84***	0.49*
	(0.3829)	(0.3041)	(0.2633)
Middle education	0.01	-1.43***	-1.98***
	(0.5194)	(0.3255)	(0.263)
High education	0.95*	-2.8***	-3.43***
	(0.5407)	(0.3519)	(0.2905)
Household size	0.37***	0.16	0.3***
	(0.1122)	(0.1027)	(0.0756)
F1: extraversion	0.26**	-0.09	-0.06
	(0.1113)	(0.0908)	(0.0684)
F2: agreeableness	-0.24**	0.12	0.07
	(0.0995)	(0.0807)	(0.0651)
F3: conscientiousness	0.14	-0.35***	-0.54***
	(0.1048)	(0.0903)	(0.0705)
F4: emotional stability	-0.31***	-0.59***	-0.5***
	(0.0976)	(0.0826)	(0.0628)
F5: openness for experience	0.5***	0.33***	0.21***
	(0.0989)	(0.086)	(0.0691)
Inverse Mills Ratio	15.53***	-4.13***	-8.04***
	(0.9191)	(1.1389)	(0.7769)
L	6.95***		
	(0.2621)		
	2.16***	3.37***	
	(0.1778)	(0.143)	
	2.47***	3.33***	1.69***
	(0.1594)	(0.1156)	(0.0826)
W	48.3517	14.9999	17.1594
	14.9999	15.9799	16.5171
	17.1594	16.5171	20.0146
Observations: 26402			
Nr. of Individuals: 5914			
Loglikelihood: -12989.71			

Regression including year fixed effects.

(Non-robust) standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.10

Table 23: Dynamic model with Big-Five factor markers, pooled

	Self-employed	Unemployed	Not in Labour Force
Constant	-6.81***	-7.17***	-4.89***
	(0.7856)	(0.7682)	(0.5285)
Age	0.02*	0.05***	0.05***
	(0.0084)	(0.0087)	(0.0063)
Female	-0.04	0.48**	0.33*
	(0.262)	(0.2256)	(0.1885)
Has partner	0.05	-0.35	-0.37**
	(0.2356)	(0.2375)	(0.1859)
Has child	-0.25	-0.28	-0.23
	(0.2619)	(0.2505)	(0.1949)
Female x partner	0.22	-0.3	0.73***
	(0.3025)	(0.268)	(0.2181)
Female x has child	-0.16	0.19	-0.1
	(0.2704)	(0.2548)	(0.1969)
Middle education	-0.13	-0.79***	-1.12***
	(0.3972)	(0.2647)	(0.2138)
High education	0.21	-1.53***	-1.78***
	(0.4078)	(0.2901)	(0.2322)
Household size	0.17*	0	0.03
	(0.0876)	(0.0846)	(0.0663)
F1: extraversion	0.15**	-0.04	-0.03
	(0.0728)	(0.0714)	(0.0515)
F2: agreeableness	-0.11	0.04	-0.02
	(0.0708)	(0.0695)	(0.0502)
F3: conscientiousness	-0.14*	-0.15*	-0.21***
	(0.077)	(0.0772)	(0.0577)
F4: emotional stability	-0.17***	-0.36***	-0.28***
	(0.0655)	(0.0652)	(0.0481)
F5: openness for experience	0.26***	0.19***	0.11**
	(0.0711)	(0.0716)	(0.0538)
Last state: self-employed	4.06***	0.93**	1.14***
	(0.1651)	(0.3819)	(0.2277)
Last state: unemployed	0.3	2.75***	1.55***
	(0.3483)	(0.2023)	(0.1664)
Last state: not in LF	0.63***	1.19***	1.89***
	(0.1985)	(0.1604)	(0.0885)
Initial state: self-employed	4.95***	1.76***	2.16***
	(0.3848)	(0.3905)	(0.2708)
Initial state: unemployed	2.92***	3.78***	3.44***
	(0.5375)	(0.3869)	(0.3484)
Initial state: not in LF	2.51***	3.53***	4.27***
	(0.2842)	(0.23)	(0.1894)
Inverse Mills Ratio	-0.36	-1.03	-2.06***
	(0.96)	(1.1517)	(0.7955)
L	2.2***		
	(0.1536)		
	1.11***	1.48***	
	(0.1756)	(0.147)	
	1.1***	1.52***	0.65***
	(0.1372)	(0.1216)	(0.1206)
W	4.8371	2.4439	2.4218
	2.4439	3.4189	3.4722
	2.4218	3.4722	3.9505
Observations: 26402			
Nr. of Individuals: 5914			
Loglikelihood: -10201.11			

Regression including year fixed effects.

(Non-robust) standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.10

Table 24: Dynamic model with Big-Five factor markers and no Heckman correction, pooled

	Self-employed	Unemployed	Not in Labour Force
Constant	-6.98***	-7.67***	-5.89***
	(0.5708)	(0.5313)	(0.3637)
Age	0.02**	0.05***	0.06***
	(0.0067)	(0.0066)	(0.0049)
Female	-0.05	0.46**	0.28
	(0.2604)	(0.223)	(0.1864)
Has partner	0.05	-0.34	-0.35*
	(0.2342)	(0.2374)	(0.1855)
Has child	-0.26	-0.29	-0.26
	(0.2604)	(0.2499)	(0.1942)
Female x partner	0.21	-0.32	0.7***
	(0.3007)	(0.2678)	(0.2174)
Female x has child	-0.16	0.2	-0.07
	(0.2689)	(0.2547)	(0.1959)
Middle education	-0.12	-0.75***	-1.04***
	(0.3894)	(0.2615)	(0.2114)
High education	0.23	-1.47***	-1.69***
	(0.3974)	(0.2856)	(0.2292)
Household size	0.17*	-0.01	0.01
	(0.0873)	(0.0832)	(0.0657)
F1: extraversion	0.15**	-0.05	-0.03
	(0.0724)	(0.0713)	(0.0514)
F2: agreeableness	-0.11	0.04	-0.02
	(0.0707)	(0.0692)	(0.0501)
F3: conscientiousness	-0.13*	-0.11*	-0.14***
	(0.0692)	(0.0659)	(0.0511)
F4: emotional stability	-0.17***	-0.37***	-0.29***
	(0.0651)	(0.0647)	(0.0477)
F5: openness for experience	0.26***	0.18***	0.09*
	(0.0704)	(0.0705)	(0.0533)
Last state: self-employed	4.04***	0.86**	1.04***
	(0.1567)	(0.3751)	(0.2224)
Last state: unemployed	0.29	2.77***	1.58***
	(0.3482)	(0.2017)	(0.166)
Last state: not in LF	0.64***	1.22***	1.95***
	(0.1961)	(0.1567)	(0.085)
Initial state: self-employed	4.95***	1.77***	2.16***
	(0.3845)	(0.3913)	(0.2694)
Initial state: unemployed	2.91***	3.77***	3.43***
	(0.5352)	(0.3865)	(0.3468)
Initial state: not in LF	2.5***	3.52***	4.26***
	(0.2837)	(0.2298)	(0.1888)
L	2.2***		
	(0.1535)		
	1.12***	1.47***	
	(0.1759)	(0.1482)	
	1.1***	1.52***	0.66***
	(0.1371)	(0.1221)	(0.1204)
W	4.8245	2.4551	2.4063
	2.4551	3.4135	3.4555
	2.4063	3.4555	3.9316
Observations: 26402			
Nr. of Individuals: 5914			
Loglikelihood: -10204.73			

Regression including year fixed effects.

(Non-robust) standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.10

Table 25: Dynamic model with EP distance, pooled

	Self-employed	Unemployed	Not in Labour Force
Constant	-6.4***	-7.59***	-5.26***
	(0.8093)	(0.7696)	(0.5423)
Age	0.01	0.04***	0.04***
	(0.0084)	(0.0086)	(0.0063)
Female	-0.07	0.42*	0.23
	(0.2531)	(0.2191)	(0.1849)
Has partner	0.02	-0.43*	-0.45**
	(0.2352)	(0.2338)	(0.1843)
Has child	-0.26	-0.25	-0.21
	(0.2613)	(0.2499)	(0.1942)
Female x partner	0.19	-0.24	0.76***
	(0.3007)	(0.2651)	(0.2175)
Female x has child	-0.21	0.16	-0.12
	(0.2694)	(0.253)	(0.1961)
Middle education	-0.15	-0.8***	-1.16***
	(0.3865)	(0.2634)	(0.2141)
High education	0.31	-1.45***	-1.79***
	(0.391)	(0.2839)	(0.2275)
Household size	0.17*	-0.01	0.03
	(0.0869)	(0.0847)	(0.0663)
EP distance	-0.01	0.04***	0.04***
	(0.0107)	(0.01)	(0.0077)
Last state: self-employed	4.05***	0.93**	1.15***
	(0.1656)	(0.3738)	(0.227)
Last state: unemployed	0.32	2.78***	1.54***
	(0.3499)	(0.2008)	(0.1665)
Last state: not in LF	0.63***	1.18***	1.89***
	(0.1961)	(0.1592)	(0.0878)
Initial state: self-employed	5.03***	1.81***	2.21***
	(0.3845)	(0.3856)	(0.2715)
Initial state: unemployed	3.03***	3.87***	3.52***
	(0.5154)	(0.391)	(0.3535)
Initial state: not in LF	2.6***	3.63***	4.35***
	(0.2821)	(0.2321)	(0.191)
Inverse Mills Ratio	-0.16	-1.06	-2.07***
	(0.9515)	(1.1191)	(0.784)
L	2.21***		
	(0.1536)		
	1.14***	1.49***	
	(0.176)	(0.1484)	
	1.13***	1.56***	0.6***
	(0.1338)	(0.1172)	(0.1222)
W	4.9029	2.5242	2.5039
	2.5242	3.5324	3.6228
	2.5039	3.6228	4.0719
Observations:	26402		
Nr. of Individuals:	5914		
Loglikelihood:	-10235.85		

Regression including year fixed effects.
(Non-robust) standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.10

Table 26: Dynamic model with EP distance, women

		Self-employed	Unemployed	Not in Labour Force
Constant		-6.56***	-6.67***	-4.93***
		(1.1146)	(0.955)	(0.6893)
Age		0.01	0.02**	0.04***
		(0.011)	(0.0104)	(0.0077)
Has partner		0.26	-0.68***	0.33**
		(0.2409)	(0.188)	(0.1683)
Has child		-0.36	-0.22	-0.42**
		(0.295)	(0.2315)	(0.1916)
Middle education		-0.1	-0.79**	-0.93***
		(0.5728)	(0.3503)	(0.2922)
High education		0.31	-1.41***	-1.39***
		(0.5916)	(0.3738)	(0.3054)
Household size		0.13	0.02	0.06
		(0.1194)	(0.0984)	(0.0814)
EP distance		-0.01	0.04***	0.03***
		(0.0149)	(0.0121)	(0.0096)
Last state: self-employed		4.05***	1.11**	1.24***
		(0.2398)	(0.4806)	(0.2958)
Last state: unemployed		0.13	2.93***	1.67***
		(0.5064)	(0.2805)	(0.2167)
Last state: not in LF		0.62**	1.53***	2.09***
		(0.2641)	(0.198)	(0.1058)
Initial state: self-employed		4.88***	1.39**	2.29***
		(0.5319)	(0.5559)	(0.3634)
Initial state: unemployed		3.49***	3.45***	3.26***
		(0.6774)	(0.5306)	(0.4821)
Initial state: not in LF		2.84***	3.29***	4.34***
		(0.3742)	(0.2959)	(0.2398)
Inverse Mills Ratio		-0.35	-0.68	-2**
		(1.2709)	(1.3529)	(0.9855)
L		2.25***		
		(0.2135)		
		1.16***	1.31***	
		(0.2288)	(0.2228)	
		1.26***	1.48***	0.58***
		(0.1822)	(0.1639)	(0.1757)
W		5.0511	2.6181	2.8361
		2.6181	3.0796	3.4134
		2.8361	3.4134	4.1256
Observations:	14435			
Nr. of Individuals:	3267			
Loglikelihood:	-6103.72			

Regression including year fixed effects.
(Non-robust) standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.10

Table 27: Dynamic model with EP distance, men

	Self-employed	Unemployed	Not in Labour Force
Constant	-6.41***	-8.03***	-5.5***
	(1.2027)	(1.4117)	(0.9209)
Age	0.02	0.06***	0.05***
	(0.0138)	(0.0163)	(0.0114)
Has partner	-0.14	-0.45	-0.43*
	(0.2534)	(0.309)	(0.2221)
Has child	-0.31	-0.1	-0.1
	(0.308)	(0.4019)	(0.2749)
Middle education	-0.13	-0.79*	-1.34***
	(0.542)	(0.4341)	(0.3337)
High education	0.3	-1.53***	-2.22***
	(0.5467)	(0.485)	(0.3705)
Household size	0.2	-0.05	-0.03
	(0.1243)	(0.1839)	(0.1286)
EP distance	-0.02	0.05**	0.05***
	(0.0164)	(0.0202)	(0.0143)
Last state: self-employed	4.13***	0.52	0.89**
	(0.2476)	(0.7188)	(0.4014)
Last state: unemployed	0.26	2.71***	1.34***
	(0.5869)	(0.3345)	(0.3126)
Last state: not in LF	0.56*	0.59*	1.49***
	(0.3306)	(0.3079)	(0.185)
Initial state: self-employed	4.88***	2.76***	2.37***
	(0.5861)	(0.7407)	(0.4359)
Initial state: unemployed	2.71***	4.22***	3.81***
	(0.8558)	(0.5976)	(0.5513)
Initial state: not in LF	2.24***	4.1***	4.4***
	(0.5159)	(0.4169)	(0.3523)
Inverse Mills Ratio	0.4	-2.39	-2.45*
	(1.4459)	(2.1704)	(1.3444)
L	2.12***		
	(0.2385)		
	1.48***	1.26***	
	(0.3368)	(0.3401)	
	1.2***	1.57***	0.41
	(0.2376)	(0.2028)	(0.3464)
W	4.5066	3.1374	2.5555
	3.1374	3.7761	3.7657
	2.5555	3.7657	4.1003
Observations: 11967			
Nr. of Individuals: 2647			
Loglikelihood: -4086.93			

Regression including year fixed effects.
(Non-robust) standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.10

Table 28: Dynamic model with health index, pooled

	Self-employed	Unemployed	Not in Labour Force
Constant	-6.59***	-7.15***	-4.7***
	(0.7895)	(0.7363)	(0.5164)
Age	0.01	0.04***	0.04***
	(0.0085)	(0.0083)	(0.0062)
Female	-0.18	0.38*	0.2
	(0.2633)	(0.2148)	(0.1829)
Has partner	0.08	-0.33	-0.38**
	(0.2354)	(0.2251)	(0.1789)
Has child	-0.21	-0.35	-0.3
	(0.2619)	(0.239)	(0.1901)
Female x partner	0.1	-0.25	0.74***
	(0.3062)	(0.2561)	(0.212)
Female x has child	-0.11	0.27	-0.03
	(0.2716)	(0.2429)	(0.1915)
Middle education	-0.03	-0.65**	-1.02***
	(0.4027)	(0.256)	(0.2116)
High education	0.48	-1.16***	-1.53***
	(0.4098)	(0.2767)	(0.2248)
Household size	0.17*	0.02	0.06
	(0.0881)	(0.0825)	(0.0665)
Health Index, lagged	-0.09	-0.35***	-0.31***
	(0.0538)	(0.0381)	(0.0333)
Last state: self-employed	4.02***	0.83**	1.14***
	(0.1637)	(0.369)	(0.2253)
Last state: unemployed	0.24	2.87***	1.5***
	(0.3559)	(0.1968)	(0.1669)
Last state: not in LF	0.67***	1.17***	1.92***
	(0.2016)	(0.1588)	(0.0877)
Initial state: self-employed	5.15***	1.95***	2.2***
	(0.3908)	(0.3749)	(0.2668)
Initial state: unemployed	2.88***	3.43***	3.24***
	(0.5244)	(0.3676)	(0.3439)
Initial state: not in LF	2.5***	3.42***	4.16***
	(0.2853)	(0.2266)	(0.1894)
Inverse Mills Ratio	-0.42	-0.47	-1.77**
	(0.9601)	(1.0979)	(0.7858)
L	2.23***		
	(0.1565)		
	1.16***	1.36***	
	(0.1688)	(0.1458)	
	1.06***	1.61***	-0.33*
	(0.14)	(0.1138)	(0.1987)
W	4.9878	2.5883	2.3697
	2.5883	3.181	3.4133
	2.3697	3.4133	3.8278
Observations: 26319			
Nr. of Individuals: 5866			
Loglikelihood: -10143.69			

Regression including year fixed effects.
(Non-robust) standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.10

Table 29: Dynamic model with health index, women

	Self-employed	Unemployed	Not in Labour Force
Constant	-6.72***	-6.4***	-4.58***
	(1.0301)	(0.9)	(0.6442)
Age	0.01	0.02**	0.03***
	(0.0107)	(0.01)	(0.0076)
Has partner	0.28	-0.6***	0.38**
	(0.2386)	(0.1841)	(0.1686)
Has child	-0.27	-0.18	-0.39**
	(0.2992)	(0.2266)	(0.1924)
Middle education	-0.14	-0.63*	-0.79***
	(0.5481)	(0.341)	(0.2925)
High education	0.27	-1.14***	-1.15***
	(0.5643)	(0.3677)	(0.3071)
Household size	0.12	0.06	0.08
	(0.1221)	(0.0966)	(0.0828)
Health Index, lagged	-0.11	-0.37***	-0.3***
	(0.0687)	(0.0492)	(0.0451)
Last state: self-employed	4.03***	1.02**	1.21***
	(0.2416)	(0.477)	(0.2947)
Last state: unemployed	0.02	3.02***	1.63***
	(0.5083)	(0.2589)	(0.2162)
Last state: not in LF	0.56**	1.51***	2.07***
	(0.2706)	(0.1971)	(0.1058)
Initial state: self-employed	4.88***	1.56***	2.34***
	(0.5248)	(0.5445)	(0.3602)
Initial state: unemployed	3.51***	3.19***	3.12***
	(0.6573)	(0.4869)	(0.4818)
Initial state: not in LF	2.86***	3.11***	4.22***
	(0.3674)	(0.2854)	(0.2365)
Inverse Mills Ratio	-0.61	-0.29	-1.91**
	(1.2789)	(1.312)	(0.9636)
L	2.22***		
	(0.2066)		
	1.28***	1.09***	
	(0.204)	(0.2046)	
	1.3***	1.53***	0.26
	(0.1761)	(0.1551)	(0.3168)
W	4.9092	2.8448	2.8775
	2.8448	2.845	3.3432
	2.8775	3.3432	4.1
Observations: 14399			
Nr. of Individuals: 3248			
Loglikelihood: -6054.79			

Regression including year fixed effects.
(Non-robust) standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.10

Table 30: Dynamic model with health index, men

		Self-employed	Unemployed	Not in Labour Force
Constant		-6.43***	-7.44***	-4.86***
		(1.2339)	(1.4121)	(0.9146)
Age		0.01	0.06***	0.05***
		(0.0138)	(0.0163)	(0.0113)
Has partner		-0.04	-0.38	-0.33
		(0.2483)	(0.3068)	(0.2179)
Has child		-0.29	-0.17	-0.16
		(0.3058)	(0.3943)	(0.2729)
Middle education		0.11	-0.73*	-1.26***
		(0.5824)	(0.4305)	(0.322)
High education		0.61	-1.29***	-2.01***
		(0.5893)	(0.4788)	(0.3591)
Household size		0.21*	-0.02	-0.01
		(0.1262)	(0.1811)	(0.1282)
Health Index, lagged		-0.07	-0.3***	-0.31***
		(0.0843)	(0.0664)	(0.0556)
Last state: self-employed		4.18***	0.54	0.95**
		(0.2419)	(0.7238)	(0.4074)
Last state: unemployed		0.42	2.68***	1.35***
		(0.5715)	(0.3352)	(0.311)
Last state: not in LF		0.73**	0.63**	1.55***
		(0.3211)	(0.3019)	(0.1798)
Initial state: self-employed		4.87***	2.5***	2.05***
		(0.5846)	(0.7152)	(0.4329)
Initial state: unemployed		2.32**	3.79***	3.38***
		(0.9402)	(0.5778)	(0.5303)
Initial state: not in LF		1.76***	3.85***	4.21***
		(0.5156)	(0.406)	(0.3476)
Inverse Mills Ratio		-0.3	-1.76	-1.61
		(1.4966)	(2.2136)	(1.3821)
L		2.09***		
		(0.2363)		
		1.23***	1.38***	
		(0.3575)	(0.2794)	
		0.89***	1.66***	0.44
		(0.2452)	(0.1836)	(0.3343)
W		4.3711	2.5745	1.8555
		2.5745	3.4159	3.3856
		1.8555	3.3856	3.7517
Observations:	11920			
Nr. of Individuals:	2618			
Loglikelihood:	-4042.31			

Regression including year fixed effects.

(Non-robust) standard errors in parentheses; *** p<0.01, ** p<0.05, * p<0.10

E. Additional figures

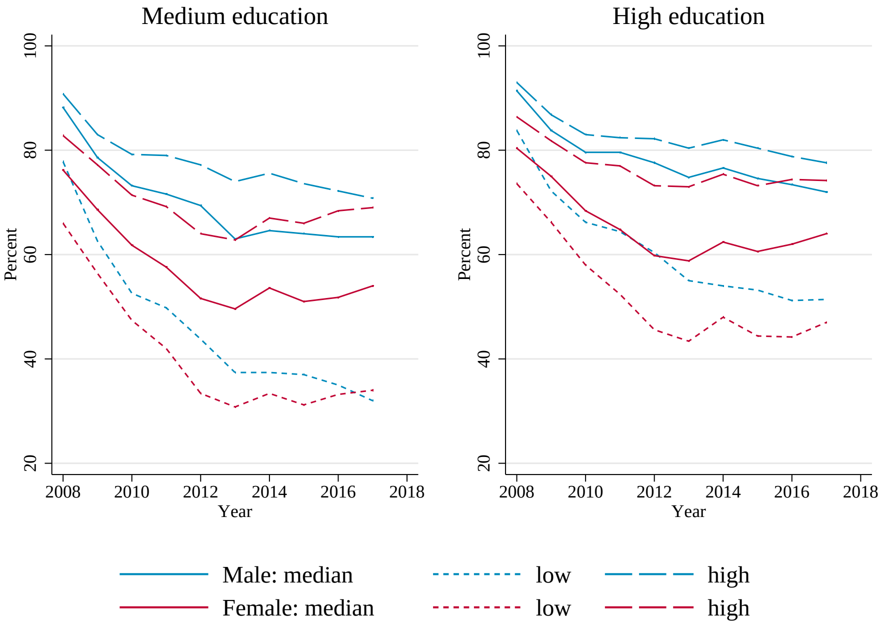


Figure 11: Share of employment paths spent in self-employment for different Big-Five factor markers (benchmark individual starts as self-employed in 2007)