RUHR

ECONOMIC PAPERS

Sylvi Rzepka

Analyzing Further Training Participation Rates across Waves in the NEPS Data

RUHR UNIVERSITÄT BOCHUM RUB RWI #655

Imprint

Ruhr Economic Papers

Published by

Ruhr-Universität Bochum (RUB), Department of Economics
Universitätsstr. 150, 44801 Bochum, Germany

Technische Universität Dortmund, Department of Economic and Social Sciences
Vogelpothsweg 87, 44227 Dortmund, Germany

Universität Duisburg-Essen, Department of Economics
Universitätsstr. 12, 45117 Essen, Germany

RWI Leibniz-Institut für Wirtschaftsforschung
Hohenzollernstr. 1-3, 45128 Essen, Germany

Editors

Prof. Dr. Thomas K. Bauer
RUB, Department of Economics, Empirical Economics
Phone: +49 (0) 234/3 22 83 41, e-mail: thomas.bauer@rub.de

Prof. Dr. Wolfgang Leininger
Technische Universität Dortmund, Department of Economic and Social Sciences
Economics – Microeconomics
Phone: +49 (0) 231/7 55-3297, e-mail: W.Leininger@tu-dortmund.de

Prof. Dr. Volker Clausen
University of Duisburg-Essen, Department of Economics
International Economics
Phone: +49 (0) 201/1 83-3655, e-mail: vclausen@vwl.uni-due.de

Prof. Dr. Roland Döhrn, Prof. Dr. Manuel Frondel, Prof. Dr. Jochen Kluve
RWI, Phone: +49 (0) 201/81 49-213, e-mail: presse@rwi-essen.de

Editorial Office

Sabine Weiler
RWI, Phone: +49 (0) 201/81 49-213, e-mail: sabine.weiler@rwi-essen.de

Ruhr Economic Papers #655

Responsible Editor: Jochen Kluve

ISSN 1864-4872 (online) – ISBN 978-3-86788-761-8

Ruhr Economic Papers #655

Sylvi Rzepka

Analyzing Further Training Participation Rates across Waves in the NEPS Data

RUHR UNIVERSITÄT BOCHUM RUB

RWI

Bibliografische Informationen der Deutschen Nationalbibliothek

Die Deutsche Bibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über: *http://dnb.d-nb.de* abrufbar.

Das RWI wird vom Bund und vom Land Nordrhein-Westfalen gefördert.

http://dx.doi.org/10.4419/86788761
ISSN 1864-4872 (online)
ISBN 978-3-86788-761-8

Sylvi Rzepka[1]

Analyzing Further Training Participation Rates across Waves in the NEPS Data

Abstract

The spell-based nature of the National Educational Panel Study poses some challenges for analyzing training participation rates across waves. Raw training participation rates of each wave using courses compiled in the SpCourses data set and in the SpEmp data set differ by up to 75 percent across waves. Such differences do not prevail in other data sets for the same time period. In this technical paper we argue that the differences arise due to different time spans for which individuals indicate their training participation. Controlling for each individual's reference period reduces the differences in training participation rates across waves to plausible levels.

JEL Classification: J24, I21, C81

Keywords: Training participation; National Educational Panel Study; Starting Cohort 6

October 2016

1 *Sylvi Rzepka, RWI and RUB.* *- This paper uses data from the National Educational Panel Study (NEPS): Starting Cohort 6 – Adults (Adult Education and Lifelong Learning), doi:10.5157/NEPS:SC6:5.1.0. The NEPS data collection is part of the Framework Programme for the Promotion of Empirical Educational Research, funded by the German Federal Ministry of Education and Research and supported by the Federal States. The author acknowledges financial support from the Deutsche Forschungsgemeinschaft (TA 829/2-2). I thank Paula Schneider and Karim Diebold for excellent research assistance. Thanks to Katja Görlitz, Merlin Penny, and Marcus Tamm for helpful comments and suggestions. I am also thankful for the support provided by the LIfBi Data Center.* *– All correspondence to: Sylvi Rzepka, RWI, Hohenzollernstr. 1-3, 45128 Essen, Germany, e-mail: sylvi.rzepka@rwi-essen.de*

1. Introduction

The economics literature dedicated to training strongly relies on measuring training participation rates within a certain time span (see for instance Pischke, 2001 and Bassanini et al., 2007). The adult cohort of the new National Educational Panel Study[2] (NEPS) is a promising data set to shed light on specific aspects of determinants and returns to training, since the panel study focuses on educational, occupational, and family formation processes (Blossfeld, 2011). It covers detailed life course information from birth to adult life for 17,137 individuals born between 1944 and 1986 and now consists of five panel waves that record current activities in the field of education, occupation and family.

However, analyzing participation in further training across the NEPS panel waves is not straight forward. As table 1 shows, participation rates based on data provided in the SpCourses and SpEmp data set differ and especially training participation calculated using employment spell information varies greatly across different waves. Most strikingly, the training participation rate drops by 25 percent (SpCourse) or 76 percent (SpEmp) from the first NEPS wave in 2009/10 to the second in 2010/11. This does not mirror the development in further training participation documented in other data sets. For comparison, Table 2 shows that training participation rates recorded by the Adult Education Survey (AES) for 2007-2014 fluctuate by 6 to 12.5 percent between two or three-year intervals. While the AES may measure training participation differently than the NEPS (Eisermann et al., 2014), the trend across time should remain comparable across these data sets. Hence it is an open question why we observe such stark differences in training participation rates across NEPS waves. This technical paper explores selective panel attrition and different reference periods as potential causes and proposes a solution how to correct for the differences in empirical work.

Table 1 Average training participation rates

	SpCourse			**SpEmp**		
Wave	Training participation	Obs.	percentage change	Training participation	Obs.	percentage change
2009/2010	0.2755	11419		0.3773	11157	
2010/2011	0.2200	9166	-25.23%	0.2145	7959	-75.90%
2011/2012	0.2539	13961	13.35%	0.3117	12768	31.18%
2012/2013	0.2253	11626	-12.69%	0.2365	9833	-31.80%

Notes: Table 1 presents average training participation rates calculated using the SpCourses data set (employment, unemployment, military service, gap year and parental leave spells) and using the SpEmp data set. We use the cross-sectional, calibrated weights provided in the NEPS.

Table 2 Training participation from Adult Education Survey

Year	Training participation during the year	Percentage change
2007	0.5200	
2010	0.4900	-6.12%
2012	0.5600	12.50%
2014	0.5800	3.45%

Source: BMBF (2015), Table 5, p. 26.

[2] doi:10.5157/NEPS:SC6:5.1.0. Please note that we made use of the correction routine provided by Künster (2015).

2. Definitions

We explore two different training participation rates which can be constructed using the NEPS Starting Cohort 6. First, we rely on Variable t271000 "Number of attended training programs / courses" (LIfBi, 2015) provided in the SpCourses data set. This data set includes the first three courses which took place during a specific employment, unemployment, parental leave, military or civilian service, or gap spell that was ongoing during the 12 months prior to the interview (Skopek, 2013)[3]. Yet, the reference period is not identical for each individual since the question on training participation is asked with respect to all spells that fall into the 12-month time window. For example, an individual may have job spell that lasted for three years and ends in month six of the 12-month window after which a new spell starts. The reference period for this individual's training participation therefore amounts to 42 months in total. Others that started a new job two months prior to the 12-month-time window have a reference period of 14 months. To account for all combinations we determine the reference period for courses from the SpCourses data set by computing the time span from the earliest to the latest spell that falls into the 12-month time window. By calculating the reference period in this way, we also deal with the fact that individuals may have several spells during this 12-month time window and do not need to identify the most important one of them.

Second, we use the variable ts23235 "Attendance of training programs or courses" (LIfBi, 2015) recorded in the SpEmp data set. This type of training is of particular relevance because most of the training literature focuses on training participation that takes place during employment. ts23235 varies by employment spell and wave and also includes entries on training participation for spells that ended earlier than 12 months prior to the interview. The concrete question that generates this variable reads: "Did you participate in training programs or courses during your occupation as (…) lasting from (…) until today, that you have not yet mentioned?"[4] for new spells and „Did you participate in training programs or courses during your job since your last interview (…) till today that you have not yet mentioned?"[5] (LIfBi, 2013) for ongoing spells. For simplicity reasons we identify a main employment spell for every individual and wave.[6] For this we use the following procedure in order to identify the main spell among parallel spells in one wave:

1) We use the spell with the longest duration.
2) If there are still several spells per wave we refer to the spell with the longest work hours.
3) For observations that still have several spells per wave we pick the first spell mentioned as the main spell per wave[7].

[3] SpCourses also records courses done during a vocational preparatory year and vocational training; however, these training episodes are likely to belong to an initial vocational training program, therefore we exclude them for our analysis of further training.

[4] The German original reads: „Haben Sie während Ihrer Tätigkeit als <26109> von <26122> bis heute Lehrgänge oder Kurse besucht, von denen Sie bisher noch nicht berichtet haben?"(LIfBi, 2013).

[5] The German original reads: „Haben Sie während Ihrer Tätigkeit seit unserem letzten Interview im <20101P3(intmPRE /intjPRE)> bis heute Lehrgänge oder Kurse besucht, von denen Sie bisher noch nicht berichtet haben?" (LIfBi, 2013).

[6] In contrast to the 12-month time window used in SpCourses, this approach makes the reference period spell-specific. This means that the individual indicates training participation for the time period of a job, instead of training participation within a certain time frame. Both approaches are valid options, which is why we rely on both in the analysis of this technical paper.

[7] This is necessary for 109 observations.

3. Reasons for large differences in training participation across waves

As illustrated in the introduction it is unlikely that the stark differences between the different NEPS waves (see Table 1) reveal only true differences in training participation, since other data sets show much lower divergences for the same time period (see Table 2). Therefore, we explore more technical issues that may bring about these differences in participation rates across NEPS waves. More concretely, we discuss whether it may be due to selective panel attrition or different reference periods. We begin with a descriptive analysis before turning to a regression set up.

Table 3 Training participation in the balanced sample

	SpCourse			SpEmp		
Wave	Training Participation	Obs.	Percentage change	Training Participation	Obs.	Percentage change
2009/2010	0.2913	6,663		0.4034	5444	
2010/2011	0.2268	6,663	-28.44%	0.23	5444	-75.39%
2011/2012	0.237	6,663	4.30%	0.2478	5444	7.18%
2012/2013	0.224	6,663	-5.80%	0.2338	5444	-5.99%

Notes: Table 3 presents average training participation rates calculated using the SpCourses data set (employment, unemployment, military service, gap year and parental leave spells) and using the SpEmp data set in a balanced sample. We use the cross-sectional, calibrated weights provided in the NEPS.

3.1 Descriptive analysis

At first sight, selective panel attrition does not explain the differences in training participation across the different waves. As Table 3 documents, the differences remain large even when we consider a balanced sample of individuals that participate in all waves.

Besides panel selectivity the reference period for the training participation can play a role since individuals may refer to a somewhat different time period when answering the question generating variables ts23235 and t271000. In the first interview, individuals indicate their training participation for the duration of all employment spells (spEmp) or spells that were ongoing during the past 12-months (spCourses), which we discussed above. In addition, after the first wave the reference periods may also differ. For example, if the individual continues the job he/ she reported in the last wave the reference period for ts23235 is the time between the interview months. If he/ she starts a new job in between interviews the question refers to the spell length up to the current interview date. Some individuals temporarily drop out of the sample for one or several waves but reenter the NEPS later; for these individuals the time between interviews may span over 24 months or more, depending on the time span of the drop out. Hence both the spell lengths and the period between interviews can vary individually.

Figure 1 and Figure 2 depict the distributions of the reference period (spell length and months between the interviews) for each wave and for training episodes mentioned in the spCourses and in the spEmp data set separately. While the reference period between interviews bunches around 12 months for waves 2010/11 to 2012/2013 and around 28 months for wave 2; the reference periods for training participation mentioned in the first interview are more spread out and vary between 0 and 600 months.

Figure 1 Kernel density estimation of distribution of reference period (spCourses)

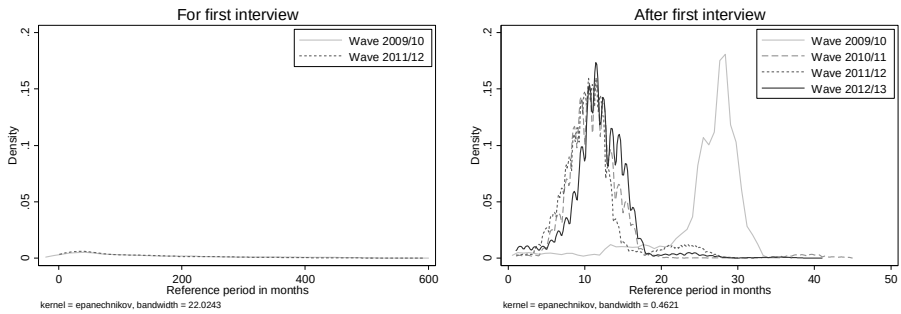
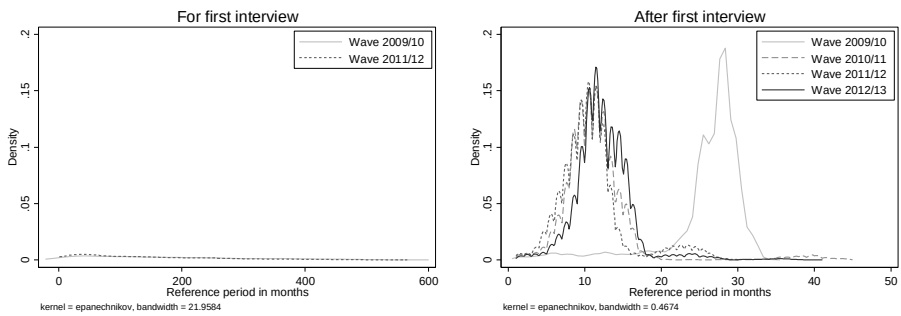


Figure 2 Kernel density estimation of distribution of reference period (spEmp)



3.2. Multivariate regression analysis

We now analyze the large variation of training participation across waves in a regression set up. The first regression controls for wave indicators only and detects significant and large differences in participation rates across waves (Table 4, Panel A, Column 1). For the training participation measured in SpCourses we find a difference between 3.3 to 5.3 percentage points and all point estimates are statistically significantly different from one another. Considering training from the SpEmp data set most wave dummies are also large and statistically significant, with point estimates ranging from 0.0220 to 0.163 (Table 4, Panel B, Column 1). This means that in wave 2009/2010 individuals are 16.3 percentage points more likely to participate in training compared to wave 2010/2011, which serves as reference period. The t-tests confirm that the coefficients on the wave indicators are statistically significantly different from one another.

In a next step we control for panel selection by including an indicator whether an individual drops out of the survey temporarily or permanently and an indicator whether the individual is part of the refresher sample, that was introduced into the NEPS in Wave 2011/12 (column 2). The differences in training participation across waves drop only minimally. Therefore, the regression analysis corroborates the findings of the descriptive analysis that selective panel attrition or composition is not the main driver for the large variations across waves.

Table 4 Regression results

Panel A: using training participation from SpCourses				
	(1)	(2)	(3)	(4)
	Training participation measured in SpCourses			
Wave 2009/10	0.0528***	0.0564***	-0.0290**	-0.0293**
	(6.91)	(7.33)	(-2.79)	(-2.77)
Wave 2011/12	0.0325***	0.0213**	0.0128	0.0163*
	(4.51)	(2.69)	(1.63)	(2.06)
Wave 2012/13	0.00539	-0.00608	-0.00129	-0.00253
	(0.64)	(-0.68)	(-0.15)	(-0.30)
Constant	0.220***	0.227***	0.231***	0.256***
	(39.29)	(38.34)	(35.01)	(17.98)
Observations	46381	46381	46378	46378
wave 2 = wave 4 (p-value)	0.0032	0.0000	0.0000	0.0000
wave 2 = wave 5 (p-value)	0.0000	0.0000	0.0104	0.0142
wave 4 = wave 5 (p-value)	0.0005	0.0004	0.0936	0.0257
Panel B: using training participation from SpEmp				
	(1)	(2)	(3)	(4)
	Training participation measured in SpEmp			
Wave 2009/10	0.163***	0.171***	-0.00814	-0.0113
	(19.65)	(20.41)	(-0.74)	(-1.00)
Wave 2011/12	0.0972***	0.0493***	0.0165*	0.0231**
	(12.52)	(5.77)	(1.97)	(2.75)
Wave 2012/13	0.0220*	-0.0183	0.0121	0.00827
	(2.39)	(-1.89)	(1.31)	(0.90)
Constant	0.214***	0.228***	0.232***	0.251***
	(36.13)	(35.55)	(32.55)	(14.43)
Observations	41693	41693	41682	41682
wave 2 = wave 4 (p-value)	0.0000	0.0000	0.0056	0.0001
wave 2 = wave 5 (p-value)	0.0000	0.0000	0.0864	0.1010
wave 4 = wave 5 (p-value)	0.0000	0.0000	0.6380	0.1120

Notes: In column 1 we control for wave indicators. Column 2 we additionally control for late survey entry and dropout. Column 3 controls for an ad hoc classification to capture the distribution of the reference period for training participation. In column 4 we use 30 quantiles to capture the distribution of the reference period. Wave 3 is the reference category. ALWA is not included in the sample. t-statistics in parentheses. * p<0.05, ** p<0.01, *** p<0.001.

Therefore, we now control for the different reference periods of training participation rates to test whether they bring about the variation across waves. First, we categorize the reference period into 6 reference periods that are in-line with the interview time frame: 1) up to ten months, 2) between

ten and 15 months, 3) between 15 and 18 months, 4) between 18 and 24 months, 5) 24-36 months, and 6) more than 36 months. Using this ad hoc categorization the differences between waves decline vastly. The difference across waves is reduced to 0.1 to 3 percentage points when measuring training participation by the SpCourses data (Table 4, Panel A, Column 3) and between 0.8 and 1.7 percentage points for training participation recorded in SpEmp (Table 4, Panel B, Column 3).

Second, we divide the reference period distribution into 30 quantiles and control for them (Table 4, Column 4). Using this detailed categorization, the variation in participating in training is lower than when only controlling for waves or panel selection (Columns 1 and 2). For instance, the coefficient on wave 2009/10 for participation in training during all spells is -0.0293 which amounts to nearly half of the difference between wave 2009/2010 and wave 2010/11 when only controlling for wave indicators. For training participation measured using the SpEmp data set, the difference decreases even more: while the coefficient on wave 2009/2010 is 0.163 in the simple regression with wave indicators, controlling for the reference periods with quantiles it drops to 0.0113.

Both specifications show that controlling for the individual reference period for training participation reduces the differences in training participation rates across waves to levels that also prevail in other data sets. Hence the remaining difference is more likely to pick up true variation in training participation rates across waves. The difference in the two different specifications is due to the different aggregation of the reference period distribution, one being derived from the survey time frame and the other being data-driven. Yet, both specifications produce differences across waves that are plausible when comparing them to other data sources such as the AES (see Table 2). Applications can choose the level of detail for the control variables depending on the degrees of freedom available in a specific estimation. From our point of view, controlling for the ad hoc classification or a smaller number of quantiles is likely to be sufficient in most cases.

4. Conclusion

The Starting Cohort 6 of the National Educational Panel Study represents a rich data source to analyze education and further training behavior. Yet, comparing further training participation across different waves is not straightforward and at first sight training participation rates, a measure that the economics training literature is most interested in, vary vastly across waves. This variation exceeds a plausible range which other German data sets record for the same time period. We explore panel attrition and different references periods for which training participation is recorded as possible reasons for this excess variation in the training participation rates. While panel attrition does not sufficiently explain the excess variation, controlling for the distribution of the reference period reduces the variation in training participation across NEPS waves to a level that also exists in other German datasets.

Therefore, we recommend to control for the reference period when analyzing the training participation rate across waves. More concretely this means to control for the spell length for first-time survey participants and the time between interviews for follow-up waves. The level of detail used for these control variables depends on the degrees of freedom available in the application, in most cases a classification that is derived from the survey time frame or a small number of quantile dummies will be sufficient. To facilitate replication, we provide a do-file that documents the data preparation and the analysis done in this technical paper in the appendix.

5. References

Bassanini, A., Booth A. L., Brunello G., De Paola M., & Leuven E. (2007). Workplace training in Europe. In A. Bassanini, A. L. Booth, G. Brunello, M. de Paola, E. Leuven, P. Garibaldi, & E. Wasmer (Eds.), Workplace Training in Europe // Education and training in Europe (pp. 143–178). Oxford: Oxford University Press.

BMBF. (2015). Weiterbildungsverhalten in Deutschland 2014: Ergebnisse des Adult Education Survey – AES Trendbericht. Bonn: Bundesministerium für Bildung und Forschung. Retrieved from https://www.bmbf.de/pub/Weiterbildungsverhalten_in_Deutschland_2014.pdf

Eisermann, M.; Janik, F.; Kruppe, T. (2014). Weiterbildungsbeteiligung: Ursachen unterschiedlicher Teilnahmequoten in verschiedenen Datenquellen. In: Zeitschrift für Erziehungswissenschaft, Vol. 17, No. 3, S. 473-495.

Künster, R. (2015). Startkohorte 6: Erwachsene (SC6) Datenversion 5.1.0 Technical Report: Korrektur der Lebensverlaufsdaten. Bamberg.

LIfBi (2015). Startng Cohort 6: Adults (SC6) SUF Version 5.1.0 Codebook (en). Available online: https://www.neps-data.de/Portals/0/NEPS/Datenzentrum/Forschungsdaten/SC6/5-1-0/SC6_5-1-0_Codebook_en.pdf (accessed: 11/02/2016).

LIfBi (2013). Startkohorte 6: Erwachsene (SC6) Wellen 4 und 5 Erhebungsinstrumente (Feldversion). Available online: https://www.neps-data.de/de-de/datenzentrum/datenunddokumentation/startkohorteerwachsene/dokumentation.aspx (accessed: 31/3/2015).

Pischke, J.-S. (2001). Continuous Training in Germany, Journal of Population Economics 14, 523-548.

Skopek, J., (2013). Data Manual, Starting Cohort 6, Adult Education and Lifelong Learning, Release 3.0.1, Bamberg.

6. Appendix

```
* Dofile prepares the further education data and relevant spell-data using
SC6 5.1.0
      *using in the SpEmp file
      *using in the SpCourse file
            *prior to this preparation we ran the correction file provided
            by: Künster, R. (2015).
            * Startkohorte 6: Erwachsene (SC6) Datenversion 5.1.0 Technical
            Report: Korrektur
            * der Lebensverlaufsdaten. Bamberg.
            * Files that were changed by this routine are indicated by
            "_V2"
      *it documents the share of training participation over the different
      waves
      *using the two different training rate figures


clear all
capture log close
set more off
set scheme  s1mono

*Data global
global data_dir  "Enter data path here"

*Files global
global main_dir "Enter data file here"

/*The dofile refers to the following sub-files:
Log
Data
Tables
Graphs*/

cd "$main_dir"
log using "$main_dir\log\further_education_across_waves", replace text



*******************************************************************************
      *Preparing Interview dates and periods in between
*******************************************************************************
use  "$data_dir\SC6_Methods_D_5-1-0.dta", clear


      #delimit ;
      keep
      ID_t                          // Target-ID
      wave                          // Panel wave
      intm                          // interview month
      inty                                // interview year
      tx80220                             // temporary dropout
      ;
      #delimit cr

      *Dropping the place holder waves for temporary dropouts
      bys ID_t: egen temp_dropout=max(tx80220)
      recode temp_dropout 1=0 2=1
      label var temp_dropout "Temporary dropouts "
      drop if tx80220==2
```

```
*Interview Number
      bys ID_t: gen int_n=_n


*Interview date
      gen date_interview=ym(inty, intm)
      format   date_interview %tm


      bys ID_t: egen wave_min=min(wave)
      bys ID_t: egen wave_max=max(wave)

      forvalues i=1/5 {
            gen interview`i'_t=date_interview if wave==`i'
            bys ID_t: egen interview`i'=max(interview`i'_t)
            format interview`i' interview`i'_t %tm
            drop interview`i'_t
            label var interview`i' "Interview date of wave `i'"
      }

      /*Period between interviews per wave*/

      gen dis_int=interview2-interview1        if wave==2
      replace dis_int=interview3-interview2    if wave==3
      replace dis_int=interview4-interview3    if wave==4
      replace dis_int=interview5-interview4    if wave==5
      replace dis_int=0                        if wave==wave_min /*first
      interview dis_int=0*/

      *Temporary dropout across different waves:
      replace dis_int=interview3-interview1    if wave==3 & dis_int==.
      replace dis_int=interview4-interview2    if wave==4 & dis_int==.
      replace dis_int=interview5-interview3    if wave==5 & dis_int==.
      replace dis_int=interview5-interview2    if wave==5 & dis_int==.
      label var dis_int "Period in between Interviews"


      gen interview_date=interview1            if wave==1
      replace interview_date=interview2  if wave==2
      replace interview_date=interview3  if wave==3
      replace interview_date=interview4  if wave==4
      replace interview_date=interview5  if wave==5
      format  interview_date %tm

      *Labeling waves
      label def wave 1 "Wave 1" 2 "Wave 2" 3 "Wave 3" 4 "Wave 4" 5 "Wave
      5", modify
      label val wave wave

      keep ID_t wave wave_min interview_date dis_int inty temp_dropout
      sort ID_t wave


      tempfile interview_dates
      save `interview_dates', replace

*************************************************************************
*****
*Further education from SpEmp
*************************************************************************
*****
```

```
use "$data_dir\SC6_spEmp_D_5-1-0_V2.dta", clear
label language en

       #delimit ;
       keep
       ID_t
       splink
       spell
       subspell
       wave

       ts23911           // Type of employee
       ts23215          // Employment in the 2nd job market
       ts23235          // course participation
      ts23219_g1        // actual/ contractual working hours at beginning
                        of contract
       ts23223_g1      // actual working hours currently
       ;

  #delimit cr



      * Merging Biography (for verified spell start and end dates)
      merge m:1 ID_t splink using "$data_dir\SC6_Biography_D_5-1-0.dta"
      drop if _m==2 /*dropping spells other than spEmp */
      drop if _m==1 /*dropping non-verified spEmp spells*/
      drop _m

      gen start_emp=ym(starty, startm)
      gen end_emp=ym(endy, endm)
      format start_emp end_emp %tm

      * Merging the interview dates
      merge m:1 ID_t wave using `interview_dates'
      gen no_sp_emp_spell_t=1 if _merge==2
      bys ID_t: egen no_sp_emp_spell=max(no_sp_emp_spell_t)
      drop no_sp_emp_spell_t
      label var no_sp_emp_spell "Has wave entries that have no emp-spell"
      drop if _m==2 /*interview dates but no sp-emp-spell*/
      drop _m

      * Keeping the latest subspell per wave
      * (this spell provides the relevant info at the time of the spell.)
      bys ID_t splink wave: egen max_sub=max(subspell)
      keep if subspell==max_sub


      gen emp_hours=ts23219_g1
      replace emp_hours=ts23223_g1 if emp_hours<0 | emp_hours==. |
      emp_hours==95
      recode emp_hours -98=. -97=. 95=.
      drop ts23223_g1 ts23219_g1


      *Training variable in SpEmp
      gen training_spemp=ts23235
      recode training_spemp  2=0 -98=. -97=.
      label var training_spemp "Training participation within an SpEmp
      Spell"

      *Dealing with missings in training_spemp
```

```
      bys ID_t wave: egen max_training_spemp=max(training_spemp)
      replace training_spemp=max_training_spemp if training_spemp==.
      replace training_spemp=0 if training_spemp==.


      tab training_spemp wave, m
      tab wave, m



*************************************************************************
******
*Identifying the main emp spell
*************************************************************************
******


*First interview
      gen first_int=0
      replace first_int=1 if wave==wave_min
      label var first_int "First interview"


      *Spell duration
      *For training spells:

      gen spell_dur_m=.
      replace spell_dur_m=(end_emp-start_emp)+1 if end_emp<interview_date &
      end_emp!=. // ends within the wave
      replace spell_dur_m=(interview_date-start_emp)+1 if
      end_emp>=interview_date & end_emp!=. // ends in a later wave
      label var spell_dur_m "Spell length in months"
      gen spell_dur_y=round(spell_dur_m/12)
      label var spell_dur_y "Spell length in years"
      bys ID_t wave: egen max_spell_dur=max(spell_dur_m)



*Identifying the main spell, i.e. most relevant spell per wave

      forvalues i=1/5{
            gen empl_int`i'=0
            replace empl_int`i'=1 if end_emp>=interview_date
      }

      gen empl_int=0
      replace empl_int=1 if end_emp>=interview_date
      label var empl_int "Employed during interview month"


      /*Generating a relevance indicator to identify the spell
      the most important per wave:
      1) there is just one per wave or
      2) longest spell or
      3) spell with the most hours, if there are parallel spells
      4) or first mentioned spell of the remaining parallel spells*/

      /*There is only one spell per wave*/
      bys ID_t wave: gen para1b=_N if wave>wave_min
      label var para1b "Parallel spells in waves after first"

      gen main_spell=0
```

```
label var main_spell "Main spell in wave"
replace main_spell=1 if para1==1 & wave>wave_min


*1) For waves that have refreshments, e.g lot of retrospective job
spells,
*only those are relevant that lasted 12 months prior to the interview
gen relevant1=.
replace relevant1=1 if end_emp>=(interview_date-12) & wave==wave_min
& wave==1
replace relevant1=1 if end_emp>=(interview_date-12) & wave==wave_min
& wave==2
replace relevant1=1 if end_emp>=(interview_date-12) & wave==wave_min
& wave==4
label var relevant1 "Spell within a time frame of 12 months prior to
the interview in wave 2 and 4"

bys ID_t wave : egen relevant1_tot=total(relevant1) if
(wave==wave_min & wave==1) | (wave==wave_min & wave==2) |
(wave==wave_min & wave==4)

/*if there is just one relevant1 spell --> that's the main_spell*/
replace main_spell=1 if relevant1_tot==1 & relevant1==1


*2) longest spell
bys ID_t wave: egen m_main1=max(main_spell) /*individual doesn't have
a main spell in this wave yet.*/

*creating new max spell dur var
drop max_spell_dur
bys ID_t wave: egen max_spell_dur=max(spell_dur_m)

gen relevant2=0
replace relevant2=1 if spell_dur_m==max_spell_dur & max_spell_dur!=.
& spell_dur_m!=.
label var relevant2 "Spell is the longest of the parallel spells"
bys ID_t wave: egen relevant2_tot=total(relevant2)

/*replace main_spell with the longest available spell, if there are
no spells of the same length*/
replace main_spell=1 if m_main1!=1 & relevant2_tot==1 & relevant2==1
& max_spell_dur!=. & spell_dur_m!=.


*3) most hours
bys ID_t wave: egen m_main2=max(main_spell) /*individual doesn't have
a main spell in this wave yet.*/

bys ID_t wave: egen max_hours=max(emp_hours)
gen relevant3=0
replace relevant3=1 if emp_hours==max_hours & emp_hours!=.  &
max_hours!=.

bys ID_t wave: egen relevant3_tot=total(relevant3)

/*replace main_spell with the job spell with the most hours, if there
are no spells with the same hours*/
replace main_spell=1 if m_main2!=1 & relevant3_tot==1 & relevant3==1
& emp_hours!=.  & max_hours!=.
```

```
bys ID_t wave: egen m_main3=max(main_spell) /*individual doesn't have
a main spell in this wave yet.*/


*4) first mentioned spell of the remaining parallel spells
*For the remaining parallel spells we just keep one random one
bys ID_t wave main_spell: gen random=_n

replace main_spell=1 if random==1 & m_main3!=1

bys ID_t wave: egen m_main4=max(main_spell) /*individual doesn't have
a main spell in this wave yet.*/

*Testing whether one spell per ID_t and wave worked.
preserve
keep if main_spell==1
bys ID_t wave: gen N=_N
tab N
restore

keep ID_t splink wave_min wave training_spemp main_spell dis_int
spell_dur_m first_int temp_dropout

*adding cross-sectional weights (these are the calibrated weights)
merge m:1 ID_t using "$data_dir\SC6_Weights_D_5-1-0.dta",
keepusing(w_t2_cal w_t3_cal w_t4_cal w_t5_cal)
drop _m

label var w_t2_cal "Weight for TP with participation in wave 2
(calibrated at micro-census 2009)"
label var w_t3_cal "Weight for TP with participation in wave 3
(calibrated at micro-census 2010)"
label var w_t4_cal "Weight for TP with participation in wave 4
(calibrated at micro-census 2011)"
label var w_t5_cal "Weight for TP with participation in wave 5
(calibrated at micro-census 2012)"

gen weight=.
replace weight=w_t2_cal if wave==2
replace weight=w_t3_cal if wave==3
replace weight=w_t4_cal if wave==4
replace weight=w_t5_cal if wave==5
label var weight "calibrated weight"


save "$main_dir\data\spemp_wave.dta", replace



*********************************************************************
*****
*SpCourses from FurtherEducation
*********************************************************************
*****


*Preparing Further education data set
use "$data_dir\SC6_FurtherEducation_D_5-1-0.dta", clear
drop if tx28200!=35 /*keeping only spcourses*/
drop tx28202_R tx28202_g13 tx28204 tx28203
merge m:1 ID_t wave splink using "$data_dir\SC6_spCourses_D_5-1-
0_V2.dta", keepusing(t271000 sptype)
```

```
label lang en

drop _m

gen start_c=ym(tx2821y, tx2821m)
gen end_c=ym(tx2822y, tx2822m)
format start_c end_c %tm


*SpCourses
gen training_spc_t=0
replace training_spc_t=1 if tx28200==35

bys ID_t wave: egen training_spc_t2=total(training_spc_t)

gen training_spc=0
replace training_spc=1 if training_spc_t2>=1 & training_spc_t2!=.
label var training_spc "Training in spCourses"
drop training_spc_t training_spc_t2

tab training_spc wave, m

*one line per ID_t wave & splink
bys ID_t wave splink: gen n=_n
keep if n==1 /*keeping one link per splink*/

keep ID_t wave splink sptype training_spc
tempfile fedu
save `fedu', replace

*Adding begin and end dates to the different spell types

*Keeping the latest subspell per wave
*Military
use "$data_dir\SC6_spMilitary_D_5-1-0_V2.dta", clear

bys ID_t splink wave: egen max_sub=max(subspell)
keep if subspell==max_sub

tempfile milit
save `milit', replace

use "$data_dir\SC6_spEmp_D_5-1-0_V2.dta", clear
bys ID_t splink wave: egen max_sub=max(subspell)
keep if subspell==max_sub

tempfile emp
save `emp', replace

use"$data_dir\SC6_spUnemp_D_5-1-0_V2.dta", clear
bys ID_t splink wave: egen max_sub=max(subspell)
keep if subspell==max_sub

tempfile unemp
save `unemp', replace
use"$data_dir\SC6_spParleave_D_5-1-0_V2.dta", clear
bys ID_t splink wave: egen max_sub=max(subspell)
keep if subspell==max_sub

tempfile pleave
save `pleave', replace
```

```
      use"$data_dir\SC6_spGap_D_5-1-0_V2.dta", clear
      bys ID_t splink wave: egen max_sub=max(subspell)
      keep if subspell==max_sub

      tempfile gap
      save `gap', replace

      *Merging all subdatasets that feed into spcourses: military, emp,
      unemp, parleave gap
      use `fedu', clear
      merge 1:1 ID_t splink wave using `milit', keepusing(ID_t wave splink)
      drop _m

      merge 1:1 ID_t splink wave using `emp', keepusing(ID_t wave splink)
      drop _m

      merge 1:1 ID_t splink wave using `unemp', keepusing(ID_t wave splink)
      drop _m

      merge 1:1 ID_t splink wave using `pleave', keepusing(ID_t wave
      splink)
      drop _m

      merge 1:1 ID_t splink wave using `gap', keepusing(ID_t wave splink )
      drop _m

      /*merging the verified begin and end dates from biography*/
      merge m:1 ID_t splink using "$data_dir\SC6_Biography_D_5-1-0.dta",
      keepusing(starty startm endy endm sptype)
      keep if _m==3 /*dropping non-verified spells (_m==1) and _m==2
      because these are sptypes that do not enter into spcourses*/
      drop _m
      merge m:1 ID_t wave using `interview_dates'
      keep if _m==3
      drop _m


*First interview
      gen first_int=0
      replace first_int=1 if wave==wave_min
      label var first_int "First interview"

*Spellduration for spCourses
      gen start=ym(starty, startm)
      gen end=ym(endy, endm)
      format start end %tm

      gen spell_dur_spc=.
      replace spell_dur_spc=end-start+1 if end<interview_date & end!=.
      replace spell_dur_spc=(interview_date-start)+1 if end>=interview_date
      & end!=. // ends in another wave
      label var spell_dur_spc "Spell duration SpCourses (in months)"

/*Generating a relevance indicator to identify the spell
      the most important per wave:
      1) spell must have been ongoing in the past 12 months.
      2) take earliest begin and latest end date of courses that fulfill 1)
      3) maximize to make information on relevant spell lengths available
      to all lines of the individual
      4) gen main_spell=1 for one of the lines of the individual, to
      harmonize with the dataset based on SpEmp*/
```

```
      *1) only those are relevant that lasted 12 months prior to the
      interview
      gen relevant1=.
      replace relevant1=1 if end>=(interview_date-12) & end!=.
      label var relevant1 "Spell within a time frame of 12 months prior to
      the interview in wave 2 and 4"

      bys ID_t wave : egen relevant1_tot=total(relevant1)
      tab relevant1_tot

*2) when there are several ongoing spells we need to add the spells to get
      the reference period
      bys ID_t wave relevant1: egen max_end=max(end) if relevant1==1
      bys ID_t wave relevant1: egen min_start=min(start) if relevant1==1

replace spell_dur_spc= max_end-min_start+1 if relevant1==1 &
relevant1_tot>1 & max_end<interview_date /*spell ends prior to the
interview*/
replace spell_dur_spc= interview_date-min_start+1 if relevant1==1 &
relevant1_tot>1 & max_end>interview_date /*spell ends after interview*/


      *3) longest spell
      *creating new max spell dur var
      bys ID_t wave: egen max_spell_dur=max(spell_dur_spc) if relevant1==1


      *4) generating main spell to harmonize with data set for SpEmp.
      bys ID_t wave: gen n=_n
      gen main_spell=1 if n==1

      *Testing whether one spell per ID_t and wave worked.
      preserve
      keep if main_spell==1
      bys ID_t wave: gen N=_N
      tab N
      restore

      *Recoding missing training_spc variable
      bys ID_t wave: egen training_spc_max=max(training_spc)
      replace training_spc=training_spc_max if training_spc==.
      replace training_spc=0 if training_spc==.

keep ID_t wave splink sptype training_spc temp_dropout wave_min dis_int
interview_date first_int start end spell_dur_spc main_spell temp_dropout

      *adding cross-sectional weights
merge m:1 ID_t using "$data_dir\SC6_Weights_D_5-1-0.dta",
keepusing(w_t2_cal w_t3_cal w_t4_cal w_t5_cal)
drop _m

label var w_t2_cal "Weight for TP with participation in wave 2 (calibrated
at micro-census 2009)"
label var w_t3_cal "Weight for TP with participation in wave 3 (calibrated
at micro-census 2010)"
label var w_t4_cal "Weight for TP with participation in wave 4 (calibrated
at micro-census 2011)"
label var w_t5_cal "Weight for TP with participation in wave 5 (calibrated
at micro-census 2012)"

      gen weight=.
      replace weight=w_t2_cal if wave==2
```

```
      replace weight=w_t3_cal if wave==3
      replace weight=w_t4_cal if wave==4
      replace weight=w_t5_cal if wave==5
      label var weight "calibrated weight"

      save "$main_dir\data\fedu_final.dta", replace

      log close
```

```
*Dofile analyzes the further education data using SC6 5.1.0
      *analysis separate for training participation recorded in spEmp
      (referring to employment spells)
      * and training participation recorded in spCourses (referring to emp,
      unemp, mil, gap spells)
      * uses the following data sets:
                  *"$main_dir\data\fedu_final.dta" (spCourses)
                  *"$main_dir\data\spemp_wave.dta" (spEmp)
      *A) explores the likelihood of selective attrition playing a role
      *B) runs regressions: does reference period play a role? (different
      specifications)

clear all
capture log close
set more off
set scheme  s1mono

*Files global
global main_dir "Enter data file here"

/*The dofile refers to the following sub-files:
Log
Data
Tables
Graphs*/

cd "$main_dir"
log using "$main_dir\log\further_education_analysis2", replace text



*******************************************************************************
****
* Analysis
* A) Selective panel attrition?
*******************************************************************************
****
*training participation recorded in spCourses
      use "$main_dir\data\fedu_final.dta", clear

      tab wave if main_spell==1, gen(wave_t)
      label var wave_t1 "Wave 1"
      label var wave_t2 "Wave 2"
      label var wave_t3 "Wave 3"
      label var wave_t4 "Wave 4"
      label var wave_t5 "Wave 5"

      bys ID_t: egen wave1=max(wave_t1)
      bys ID_t: egen wave2=max(wave_t2)
      bys ID_t: egen wave3=max(wave_t3)
      bys ID_t: egen wave4=max(wave_t4)
      bys ID_t: egen wave5=max(wave_t5)
      gen in_all_wave=0
```

```
replace in_all_wave=1 if wave2==1 & wave3==1 & wave4==1 & wave5==1 &
main_spell==1 /*appears in all waves and has a relevant spell.*/
label var in_all_wave "Participated in all waves"

sum training_spc if wave2==1 & wave3==1 & wave4==1 & wave5==1 &
main_spell==1


*Tabout on the ID_t-wave-level - balanced sample 2-5
tabout wave [aw=weight] using "$main_dir\tables\incidence_AN-
Def_fixedsample2-5_wght.xls" ///
if wave2==1 & wave3==1 & wave4==1 & wave5==1 & wave>1 & main_spell==1
, ///
replace     sum  cells(mean training_spc )

*Reference period (between interviews or spell duration)
rename spell_dur_spc spell_dur_m /*harmonizing names*/
gen ref_period=dis_int
replace ref_period=spell_dur_m if wave==wave_min /*first interview*/
replace ref_period=spell_dur_m if dis_int>spell_dur_m /*reference
Spell shorter than the

            difference between interviews*/
gen late_enroll=0
replace late_enroll=1 if wave_min==4
label var late_enroll "Individual joins NEPS in wave 4"

gen dropout=0
replace dropout=1 if in_all_wave==0
label var dropout "Not in all waves"


tempfile training_spc
save `training_spc', replace

*training participation recorded in spEmp
use "$main_dir\data\spemp_wave.dta", replace

tab wave, gen(wave_t)
label var wave_t1 "Wave 1"
label var wave_t2 "Wave 2"
label var wave_t3 "Wave 3"
label var wave_t4 "Wave 4"
label var wave_t5 "Wave 5"

bys ID_t: egen wave1=max(wave_t1)
bys ID_t: egen wave2=max(wave_t2)
bys ID_t: egen wave3=max(wave_t3)
bys ID_t: egen wave4=max(wave_t4)
bys ID_t: egen wave5=max(wave_t5)
gen in_all_wave=0
replace in_all_wave=1 if wave2==1 & wave3==1 & wave4==1 & wave5==1
label var in_all_wave "Participated in all waves"

sum training_spemp if wave2==1 & wave3==1 & wave4==1 & wave5==1 &
main_spell==1


*Tabout on the ID_t-wave-level - balanced sample 2-5
tabout wave [aw=weight] using "$main_dir\tables\incidence_AN-
Def_fixedsample2-5spEmp_wght.xls" ///
```

```
      if wave2==1 & wave3==1 & wave4==1 & wave5==1 & wave>1 & main_spell==1
      , ///
      replace     sum  cells(mean training_spemp )

      *Reference period (between interviews or spell duration)
      gen ref_period=dis_int
      replace ref_period=spell_dur_m if wave==wave_min /*first interview*/
      replace ref_period=spell_dur_m if dis_int>spell_dur_m /*reference
      Spell shorter than the difference between interviews*/
      gen late_enroll=0
      replace late_enroll=1 if wave_min==4
      label var late_enroll "Individual joins NEPS in wave 4"

      gen dropout=0
      replace dropout=1 if in_all_wave==0
      label var dropout "Not in all waves"

      tempfile training_spemp
      save `training_spemp', replace

*************************************************************************
*******
*B) Regressions: does reference period play a role? (different
specifications)
*************************************************************************
*******

*Distribution of reference period (looping over spCourse and spEmp training
definition)


      foreach var in training_spc training_spemp{
      use ``var'', clear

      label var ref_period "Reference period in months"


      kdensity ref_period if wave==2 & main_spell==1  & first_int==0,
      lcolor(gs12) ///
      addplot(kdensity ref_period if wave==3 & main_spell==1  &
      first_int==0, lcolor(gs9) lpattern(dash) ///
      || kdensity ref_period if wave==4 & main_spell==1 & first_int==0,
      lcolor(gs6) lpattern(shortdash)     ///
      || kdensity ref_period if wave==5 & main_spell==1  & first_int==0,
      lcolor(gs3))   ///
      legend(ring(0) pos(2) label(1 "Wave 2009/10") label(2 "Wave 2010/11")
      ///
      label(3 "Wave 2011/12") label(4 "Wave 2012/13")) title("After first
      interview", nobox) ylabel(0(0.05)0.2)


      graph export
"$main_dir\graphs\dist_refperiod_after_firstint_`var'.emf", replace

      kdensity ref_period if wave==2 & main_spell==1  & first_int==1,
      lcolor(gs12) ///
      addplot(kdensity ref_period if wave==4 & main_spell==1  &
      first_int==1, lcolor(gs6) lpattern(shortdash)) ///
      legend(ring(0) pos(2) label(1 "Wave 2009/10") label(2 "Wave
      2011/12")) ///
      title("For first interview", nobox) ylabel(0(0.05)0.2)
```

```
	graph export "$main_dir\graphs\dist_refperiod_firstint_`var'.emf",
	replace
	}

*looping over spCourse and spEmp training definition
	*Only wave dummies
	foreach var in training_spc training_spemp{
	use ``var'', clear
		reg `var' wave_t2 wave_t4 wave_t5 if wave>1 & main_spell==1
		[pw=weight] /*wave 2 ref*/
		test wave_t2=wave_t4
		test wave_t2=wave_t5
		test wave_t4=wave_t5

		}


*Wave dummies with a selection dummy

	foreach var in training_spc training_spemp{
	use ``var'', clear
	reg `var' wave_t2 wave_t4 wave_t5 dropout late_enroll if wave>1 &
	main_spell==1 [pw=weight]
	test wave_t2=wave_t4
	test wave_t2=wave_t5
	test wave_t4=wave_t5
	}


*30 different quantiles
	foreach var in training_spc training_spemp{
	use ``var'', clear

	xtile ref_period30=ref_period if wave>1, n(30)
	tab ref_period30, gen(ref_period30_)
	forvalues i=1/27{
		label var ref_period30_`i' "Quantile `i'"
	}

	reg `var' ref_period30_1 ref_period30_2 ref_period30_3 ref_period30_4
	///
	ref_period30_5 ref_period30_6 ref_period30_7 ref_period30_8
	ref_period30_9 ///
	ref_period30_10 ref_period30_11 ref_period30_13 ref_period30_14
	ref_period30_15 ///
	ref_period30_16 ref_period30_17 ref_period30_18 ref_period30_19
	ref_period30_20 ///
	ref_period30_21 ref_period30_22 ref_period30_23 ref_period30_24
	ref_period30_25 ///
	ref_period30_26 ref_period30_27  ///
	wave_t2 wave_t4 wave_t5 if wave>1 & main_spell==1 [pw=weight]
	/*ref_period12 Reference - median*/

	test wave_t2=wave_t4
	test wave_t2=wave_t5
	test wave_t4=wave_t5
	}


*Adhoc classification
	foreach var in training_spc training_spemp{
	use ``var'', clear
```

```
        forvalues i=1/6{
        gen ref_period`i'=0
        replace ref_period`i'=. if ref_period==.
        }

        replace ref_period1=1 if ref_period<=10 & ref_period>0
        replace ref_period2=1 if ref_period>10 & ref_period<=15
        replace ref_period3=1 if ref_period>15 & ref_period<=18
        replace ref_period4=1 if ref_period>18 & ref_period<=24
        replace ref_period5=1 if ref_period>24 & ref_period<=36
        replace ref_period6=1 if ref_period>36 & ref_period!=.

        label var ref_period1 "Ref period in months 0-10"
        label var ref_period2 "Ref period in months 10-15"
        label var ref_period3 "Ref period in months 15-18"
        label var ref_period4 "Ref period in months 18-24"
        label var ref_period5 "Ref period in months 24-36"
        label var ref_period6 "Ref period in months 36+"

        reg `var' ref_period1 ref_period3 ref_period4 ///
        ref_period5 ref_period6 ///
        wave_t2 wave_t4 wave_t5 if wave>1 & main_spell==1 [pw=weight]
        /*ref_period2 Reference*/

        test wave_t2=wave_t4
        test wave_t2=wave_t5
        test wave_t4=wave_t5
        }

log close
```