Harmonization of data from life course panel surveys
Implementation in the CNB-Young project

[This is work in progress – please do not cite.]
Warsaw, October 2023

1. Introduction

This working paper describes the CNB-Young data layout developed to facilitate cross-national comparative analysis of employment career data from national panel surveys. Given differences in the country- and survey specific methodologies of gathering and organizing biographical information, comparisons of data from these surveys requires a general framework enabling to transform employment career and other life-course data into a common format. The CNB-Young transforms the source data into person-year records with variables for each month of a year reporting whether an event or activity occurred at this time. This layout resembles the calendar format in that it provides information in the presence of a given activity in each time unit covered by the survey. Compared to the spell format, where the same information is coded in variables providing the start and end dates of each spell, this layout consumes more space, but has key advantages. First, it locates the different activities and life events, as well as the survey waves on the same time grid. This creates a coherent basis for analyzing individuals' life course trajectories while taking into account differences in the data quality related to the use of retrospective questions with unequal time distance between the survey and the life events. Second, it allows for simple and straightforward identification of overlapping activities (e.g., periods of multiple job holding or co-occurrence of various economic statuses) without engaging in pairwise comparisons of the dates of each activity spell. Third, it offers a simple way of coding information that is censored. Fourth, this format is convenient and easy to use in event history analysis or various types of sequence analysis, which require extracting a fragment of the respondents' biographies lasting a certain number of years.

2. Rationale

There are two approaches to analyzing work histories in a cross-national framework: the first relies on existing cross-national studies, and the second uses tools of ex-post harmonization to bring together studies from different countries. The latter approach, adopted in CNB-Young, is challenging as it requires comparing data from national panel surveys which define research concepts in different ways and use different tools in the field. However, it is the national panel surveys that currently provide the richest data on the work histories of individuals. The only international panel survey aimed at reconstructing employment trajectories was the European Community Household Panel (ECHP), conducted annually in

1994-2001 in 15 countries. Since 2005, the only international survey that tracks respondents over a period of several years, is the European Survey of Income and Living Conditions (EU-SILC). However, the maximum length of the observation period in the EU-SILC is only 4 years, and the survey itself has not been designed for the purpose of analyzing employment biographies. Cross-national career data spanning longer periods of observation are only available from surveys which collect full retrospective life-history data in a single wave (e.g., the Generations and Gender Survey GGS, Survey of Health, Ageing and Retirement in Europe SHARE). However, such data are the most prone to recall bias due to the length of the period between the event and the interview (Tourangeau, Rips & Rasinski 2000; Mathiowetz & Duncan 1988; Pyy-Martikainen & Rendtel 2009; Pina-Sánchez, Koskinen & Plewis 2014; 2019). The lack of cross-national surveys providing up-to-date and reliable data covering multi-year career histories warrants the development of ex-post harmonization tools suitable for national panel surveys.

Existing projects involving the ex-post harmonization of panel surveys from different countries (e.g., the Cross-National Equivalent File CNEF; Consortium of Household Panels for European Socio-Economic Research CHER, and the Comparative Panel File CPF) limit their attention to household surveys carried out annually. These projects cover a limited set of life-course variables, mostly measured on a yearly basis at the time of the survey and thus not allowing for a full reconstruction of individual employment trajectories (Frick et al. 2007; Turek, Kalmijn & Leopold 2021). By including retrospective data, the CNB-Young project allows for a more detailed analysis of career histories, and opens up the possibility to include surveys which differ in the frequencies of contacts with respondents.

3. The CNB-Young framework: basic principles

This section describes the approach to data harmonization adopted in CNB-Young. Typically, ex post harmonization projects transform data to generate a set of comparable "target variables" that maximize the equivalence of measures across surveys. This is achieved on the one hand by standardizing the response scales and classifications, and on the other hand by flagging discrepancies that cannot be standardized to warn data users that there are differences in the measurement procedures that may bias the comparison results (Granda, Wolf & Hadorn 2010; Słomczyński & Tomescu-Dubrow 2018). CNB-Young does not attempt to minimize cross-national differences in panel survey data by building equivalent target variables. Our goal is to develop and implement an integrated flexible data layout to facilitate cross-country comparisons – leaving the final decisions on how to build equivalent indicators to data users.

3.1. Minimizing information loss

The CNB-Young data format must be flexible enough to preserve all the original information on different types of economic activity as included in the source data, to allow for

maximum freedom for data users in developing their own criteria for the classification of these activities. In our approach, a key part of harmonization lies in proposing a data layout which does not require data transformations that could lead to information loss.

The same principle is adopted when dealing with overlapping activities (multiple jobs or non-work statuses) as well as inconsistencies within the dataset which arise especially when working with retrospective data (e.g., when the same activity spell is characterized differently in two panel waves, or when two different activities are reported to have taken place in the same period). Data cleansing has led researchers to disregard certain information, e.g., spells of unemployment reported during periods of work (Halpin 1998; Maré 2006; Wright 2020; see also Turek, Kalmijn & Leopold 2021). However, such interference is to some extent arbitrary and may preclude certain types of research questions (Köhler & Thomsen 2009). For example, individuals may well consider themselves unemployed even while performing a temporary job – in such situations, seemingly contradictory declarations carry substantively meaningful information – although it is also possible that such overlaps are a result of recall errors. Rather than correcting inconsistencies by selecting a value deemed as "more reliable", we minimize interference with the original (source) data – allowing users to make their own decisions with regard to ways of addressing the inconsistencies. The only exception to this rule is described in section 3.2 below.

The principle of minimizing information loss involves, in particular:

- retaining the country- and survey-specific classifications of types of activities covered in the source datasets,
- including in the proposed layout information on all successive and overlapping spells of work or non-work activities reported by the respondents in each survey wave,
- preserving the information on the source of data on a given spell/ activity in order to allow for linkage with further information on these activities in the source data.

3.2. Temporal proximity as the basis of data selection

While we acknowledge the importance of preserving information on economic activity spells from the source data, we also aim to maintain a certain level of simplicity and clarity in the final data layout. In case of panel surveys that collect working life histories in such a way that the same activity spell in the same time unit can be covered by more than one wave (as, for example, when a respondent who is in the same job in two consecutive waves reports the starting date of this job in both waves), we prioritize the information gathered at the nearest point in time. This is driven by the assumption – confirmed in methodological studies – that the time distance between the event (the start of the job in our example) and its reporting is likely

to increase recall bias. The selection of information makes the dataset more manageable to users and allows to avoid double counting of spells in a given period of time.

Specifically, in CNB-Young the period of data coverage for each survey wave in which the information on a specific type of event or activity is collected starts in the month following the previous survey wave collecting this information and ends in the month when the data for this wave was collected. If the reported spell starts before the previous wave, the activity status of the first months following this wave is coded as a "censored" starting month of the spells (see section 5.3 for more detailed information on the activity status codes), which alerts users to the fact that the actual coverage of the source data extends beyond the period reported in the CNB-Young transformed dataset. The latter information is also retained by the inclusion of the reported starting date of each spell among the variables characterizing each activity spell. As indicated before, the dataset also includes flags that allow to link each activity spell to the original (source) variables and datasets. That way, researchers interested in studying specific research problems that require using data on past activities recorded in the more recent waves, can easily retrieve and add the necessary information to the dataset; thus, reversing the information loss caused by selecting data based on temporal proximity.

3.3. Linking activities at the seams

A major challenge in reconstructing employment and labor market activity histories based on panel data is associated with linking activity spells which occur at the seams of the surveys (i.e., in cases when an ongoing spell is recorded in one panel wave and continued in the period covered by the next wave). In the initial stage of data preparation, as described in more detail below, CNB-Young makes no attempt to link spells in such cases but records them as separate. This approach is illustrated in Figure 1: for a respondent who participated in 3 waves of a panel survey and in reality had two jobs in the period covered by the survey (job 1 and job 2), four separate job spells appear in the transformed dataset: job 1a, 1b, 2a (partially overlapping with 1b), and 2b. This is consistent with the principle of minimizing interference with the data and using data from the next nearest wave (as described in sections 3.1 and 3.2). At the same time, information that the activity spell was left-censored (e.g., jobs 1b and 2b in Figure 1) is retained in the dataset, together with information on the right-censoring of spells that were ongoing at the time of data collection. Such an approach is convenient as it allows to link each (censored) activity spell to a single data source. This facilitates recording any changes in important activity characteristics (e.g., income from work) for spells spanning two or more successive waves.

Some panel surveys allow to directly connect jobs at the seams of survey waves by using dependent interviewing (Perales, 2014; Jäckle & Eckman, 2019). Information on how the career data were collected (via dependent interviewing or questionnaire modules which are only

completed by those who declared that their labor market situation has changed since the previous interview) are retained in the CNB-Young dataset as important control variables that are likely to affect the number of non-reported economic activity spells (İsaoğlu 2010). To allow for lining jobs across wave, we provide a cross-wave ID variable, as well as additional methodological variables, such as the total number of survey waves in which a given job was recorded, as well as the first and last of these waves. In surveys that do not use dependent interviewing, but record information on a given job in more than one wave, inconsistencies may arise. In such cases, merging job spells across waves may require additional assumptions with regard to the most reliable criteria for linking the data. CNB-Young aims to provide recommendations with regard to merging job information to facilitate using the data by those who are interested in substantive rather than methodological issues. However, users should be aware that merging job spells cannot be done equally well for all surveys and individuals included in the comparison, given the differences in survey methodology and in the accuracy of retrospective information reported by the respondents. Inconsistencies in information provided in different waves are also flagged in the CNB-Young dataset.

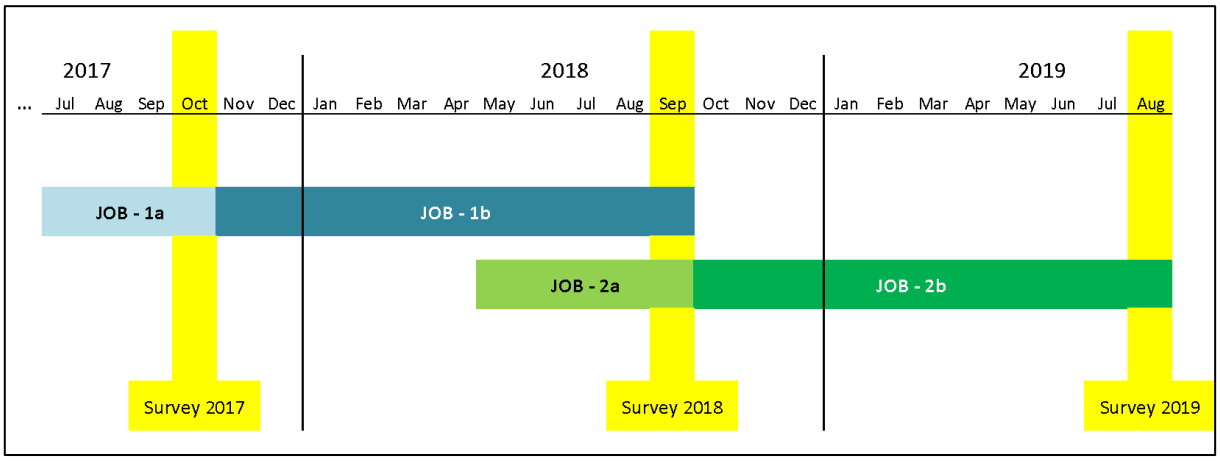

Figure 1. Illustration of a case when two job spells are reported in 2 survey waves (out of 3).

4. The CNB-Young dataset

We attach data records to respondent-years (long data format), with sets of monthly variables for each activity spell, along with other characteristics of the activities reported in a given year. While it is feasible to organize data in a long format with person-months as units (which also allows to deal easily with overlapping states), we find the yearly format preferable for a number of practical reasons. The yearly grid may be more easily adaptable to various survey designs, given that many household panel surveys conducted around the globe are implemented on an annual basis. In panel surveys carried out less frequently, retrospective data can be assigned to the years in which there was no interview. In addition, many surveys do not

collect detailed information on the timing of important biographical events (e.g., the month of marriage or childbirth), labor market transitions (e.g., moving to a different employer), or data on wages or other job characteristics for each month of each year. Even if detailed monthly data is collected, it is more likely to be affected by non-response. With the yearly format, information on the occurrence of an event can be attached to a given year in a straightforward way, even if the exact month the event occurred remains unknown. Also, it allows for minimizing the loss of cases due to missing data (i.e., on the exact month of a job termination). However, it should be noted that the CNB-Young layout can easily be reshaped into a long monthly dataset according to the users' needs.

The variables can be divided into four groups: (1) respondent characteristics; (2) source data characteristics; (3) characteristics of each activity; and (4) yearly summary characteristics. In addition to data coming from the source surveys, methodological control variables (M) are added to each group. They help to account for cross-country differences in the research procedures. The specific variables characterized below (and listed in Appendix 1) were proposed for use with the four surveys included in the CNB-Young project: the German Socio-Economic Panel G-SOEP (Giesselmann et al. 2019); the UK Household Longitudinal Survey and Understanding Society UKHLS, preceded by the BHPS up to 2008 (Platt et al., 2020); two cohort studies of the U.S. National Longitudinal Surveys of Youth: NLSY79 Young Adults (YA) and NLSY97 (Cooksey 2018), and the Polish Panel Survey POLPAN (Tomescu-Dubrow et al. 2021). The variable list can be further adapted to allow the incorporation of other survey projects by future researchers.

4.1. Respondent characteristics

This group includes the respondents' unique ID (as per the source data file) as well as their basic socio-demographic characteristics: date of birth, gender, education, region of residence, marital status, number of children, etc., followed by information on the history of participation in the survey (year the respondent entered the sample, waves in which participated). Another variable provides the number of the survey wave in which the respondent participated in a given year (necessary given that the timing of fieldwork within a single wave need not correspond to calendar years). The most recent survey is always a boundary to the data; it is unknown what happened to the respondent later. To avoid the risk of conclusions based on incomplete data, it is sometimes better to exclude data describing the last year the respondent was surveyed. The decision may depend on whether most of the contacts with respondents took place at the beginning or the end of the year, information which is provided in the second set of variables.

4.2. Variables describing the data source

This group includes the dates of the interviews conducted with the respondents in the current year, previous year, and the following year. Given that for each activity spell we also provide source wave identifiers (see section 4.3), including the interview dates allows data users to easily calculate the depth of retrospection for each activity or event even if information on different activities or events included in the same data record come from different survey waves. To denote the resulting variable (to be computed by data users), we propose the term **retrospective distance scale** RDS, expressed in years or months.

The RDS is a linear scale which allows to control for unequal data quality in the consecutive time units, especially if the retrospective data were collected in more than one survey wave. For example, for life events that took place in the survey year, the yearly RDS value equals 0, 1 for events from the preceding year, etc. In panel surveys, different cycles of contacts with respondents are adopted (in surveys covered by the CNB-Young project, the research cycle ranges from annual to 5 years, in some cases with an even longer period for those who drop out and then reenter the survey), which affects the size of potential recall bias. Harmonizing panel data requires tools to control the effects of unequal survey frequency and depth of retrospection. The linear RDS can be easily transformed to an exponential scale or according to any other function specified by the researcher.

Other variables included in this group flag important methodological characteristics of the data source, e.g., mode of data collection. In this regard, we draw on tools developed in existing data harmonization projects, which use control variables to flag differences in measurement tools and procedures that can bias the results of the comparisons, allowing users to control for the effects of these differences (Saris & Revilla, 2016; Słomczyński & Tomescu-Dubrow, 2018).

4.3. Variables describing activities

This group includes characteristics of each state (labor market activity) available in the source data. A list of activities captured by each survey included in CNB-Young is provided in Appendix 2. To record the timing of each activity, we include a set of 12 variables for each month of the year. For these variables, we use three-digit **Activity Status Codes (ASC)**. The first digit records whether a given activity is present in a given month, with values: "0" – activity not present, "1" – activity present ("-1" is the code for missing data). The second digit flags the beginning and ending months of each activity spell, which, for the first and last month of each year, allows for an easy distinction between months in which the spell is ongoing from those in which it started or ended. In the ASC scheme, "1" marks the first month of the activity, "0" – a month in which the activity is continued, "2" – the last month of the activity, and "3" – the first and last month of the activity (in case of spells that had started and ended in the same month).

Finally, codes with "1" as the last digit correspond to situations where the data comes from the source dataset, and "2" marks censored (truncated) spells. For spells that begin and end in the same month and the last digit is "2", it is important to distinguish whether it is the end or the beginning of this one-month spell that is censored. This depends on the timing of the survey wave from which the data are taken. If the interview took place in this month, the code "132" means that this same month is also the reported as the starting time of an ongoing activity and the data is right censored. If the code "132" appears in the month following the preceding interview, this is the last reported month of the activity and the data is truncated from the left side. It cannot be ruled out that the activity also took place in the preceding months, but has been ignored (or coded as a separate spell) due to the principle of accepting only data from the next closest survey. The third digit of ASC performs a function similar to harmonization control variables, except that methodological information is coded not as a separate variable, but as values of the variable indicating whether the activity occurred in a given time unit.

Apart from the monthly ASC variables, the layout also includes the total number of months of the year in which the activity occurred. This variable is convenient for surveys that record the duration of a spell in months but do not provide the exact dates, thus not allowing for the calculation of ASC values. It also offers more flexibility to retain information on the location of a spell in time when some of the original date variables have missing data. Specifically, we use a special code "-2" to mark spells that took place in this year, but the number of months is unspecified due to non-response. This code carries more information than "-1" (missing data), when we do not know both the years and months of the activity. We also include the reported start and end dates of each activity spell, even if these dates do not fall within the year to which the data record is attached. Although this information is repeated in all data records for each spell, it provides a frame of reference for determining whether this activity should be considered incidental or long-lasting.

The number of characteristics of each activity spell included in CNB-Young is larger for jobs. For each job spell, we provide occupational and industry codes (both the original codes used in each survey and international classifications, e.g., ISCO), information on the type of employment contract, number of working hours per week, and income. Certain items are available only for a subset of countries but are included due to their importance in assessing labor market performance of individuals in the countries where the data are gathered (e.g., reasons for job terminations, or access to employee benefits in the U.S.). For the non-work activities, such as unemployment or caring for family members, the number of included additional characteristics is more limited (see Appendix 1). The amount of information available in the source datasets regarding different activities is sometimes large and need not be included in the CNB-Young data file. Instead, to facilitate the addition of further activity characteristics from source data by users according to their research needs, we include ID

variables that link each activity spell to the source data, both in terms of survey wave and specific source variables or survey items. We also provide additional methodological information which allows to examine jobs which were reported in more than one wave.

4.4. Summary Indicators

The set of variables referring to the whole year usually contains some measures that can be derived from monthly data by the data users themselves. However, this requires a lot of work and carries the risk of errors. More importantly, some summary statistics for years may be retrieved from direct questions on occurrence and duration of specific activities which sometimes are asked in a manner unrelated to "activity history reconstruction" questions. The CNB-Young datasets include certain frequently used annual summary variables, both constructed and taken directly from the source data, if available. These are, among others, the total number of months in each type of activity; the number of spells of each type of activity (for seam years, spells that are right-censored and not followed by corresponding left-censored spells are not counted); the number of job terminations; or measures of annual income from work. We also provide cumulative indicators such as the total years of work experience up to the year of the data record, used, inter alia, for estimating earnings, as well as dummy variables to establish whether the respondent had ever experienced activities or states typically considered in labor market research, such as work, unemployment, parental leave, in the years preceding the year of the data record.

Finally, it is convenient for data users to have variables that characterize labor market flows, or transfers between different states. However, computing them requires assumptions with regard to the main activity in a given month in case of overlaps. Given our focus on employment careers, in CNB-Young priority is given to job spells (i.e., if someone is working and at the same time reports another type of activity, work is treated as the main status). Constructed variables based on such assumptions may seem questionable from the point of view of certain substantive research questions, but nonetheless provide a rough idea of the occurrence of various states and the frequency of transfers between them. To retain transparency and document our decisions with respect to establishing the hierarchy of spells, we include in the dataset 12 variables that identify what we define as the main activity spells in each month of the year. This is also important as in the case of some panel surveys (notably, the BHPS / UKHLS), the event history modules cover only the main economic activity, according to the respondents' self-definition.

5. Concluding remarks

The CNB-Young proposal seeks to maximize the flexibility of the data layout to make it adaptable to the different methodological design and scope of national panel surveys. At the same time, clarity and ease of use are major aspects. Providing transformed variables in a

unified format - while minimizing information loss - makes complex life-course data more accessible to users, reducing the work required to prepare the data for cross-country analysis. However, users should keep in mind that, given the differences between the surveys and the country-specific context which affects what is captured by survey items, direct comparisons of some of the CNB-Young variables should be approached with caution. CNB-Young will provide additional documentation describing these differences and their implications for cross-national comparability.

A final remark concerns the necessity of disaggregating all the non-work statuses or activities into separate spells. Choosing this as a general rule to follow while preparing career history data is intended to standardize the data layout (by using the same approach regardless of activity type) while allowing us to store the maximum available information. Indeed, some activities or statuses possess important differentiating characteristics, be it by their nature (e.g. paid or unpaid maternity leave), links to other spells (e.g., interruptions such as prolonged unpaid leave associated with a specific job), or the survey, from which data on them is derived (e.g., when data on different spells of the same activity in one year come from different panel waves). However, if these differentiating characteristics are not present with regard to specific types of activities recorded in a given survey, disaggregation into spells that increases the size of the dataset without adding any meaningful information need not be pursued. In such cases, only information on the type of activity, source dataset and questions, and ASC variables for each month of the year can be provided to avoid unnecessary fragmentation of information. CNB-Young offers a "blueprint" for data transformations, but specific implementation may to some extent differ to reflect the survey characteristics.

References

Cooksey, E. C. (2018). Using the National Longitudinal Surveys of Youth (NLSY) to Conduct Life Course Analyses. In N. Halfon, C. B. Forrest, R. M. Lerner, & E. M. Faustman (Eds.), *Handbook of Life Course Health Development* (pp. 561–577). Springer International Publishing. https://doi.org/10.1007/978-3-319-47143-3_23

Frick, J. R., Jenkins, S. P., Lillard, D. R., Lipps, O., & Wooden, M. (2007). The Cross-National Equivalent File (CNEF) and its Member Country Household Panel Studies. Schmollers Jahrbuch, 127(4), 627–654

Giesselmann, M., Bohmann, S., Goebel, J., Krause, P., Liebau, E., Richter, D., Schacht, D., Schröder, C., Schupp, J., & Liebig, S. (2019). The individual in context(s): research potentials of the Socio-Economic Panel Study (SOEP) in sociology. *European Sociological Review*, 35(5), 738-755. https://doi.org/10.1093/esr/jcz029

Granda P., Wolf, C., & Hadorn, R. (2010). Harmonizing Survey Data. In J.A. Harkness, M. Braun, B. Edwards, T.P. Johnson, L. Lyberg, P.Ph. Mohler, B.-E. Pennell, & T.W. Smith (eds.), *Methods in Multinational, Multicultural and Multiregional Contexts* (pp. 315-332). Hoboken, NJ: John Wiley & Sons.

Halpin, B. (1998). Unified BHPS work-life histories: combining multiple sources into a user-friendly format. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 60(1), 34-79. https://doi.org/10.1177%2F075910639806000102

İsaoğlu, A. (2010). *Occupational affiliation data and measurement errors in the German Socio-Economic Panel*. SOEP Papers on Multidisciplinary Panel Data Research 318. Berlin: Deutsches Institut für Wirtschaftsforschung.

Jäckle, A., & Eckman, S. (2019). Is that still the same? Has that changed? On the accuracy of measuring change with dependent interviewing. *Journal of Survey Statistics and Methodology*, 8(4), 706–725. https://doi.org/10.1093/jssam/smz021

Köhler, M., & Thomsen, U. (2009). Data Integration and Consolidation of Administrative Data From Various Sources. The Case of Germans' Employment Histories. *Historical Social Research/Historische Sozialforschung*, 34(3), 215-229.

Maré, D. C. (2006). Constructing Consistent Work-life Histories: A guide for users of the British Household Panel Survey. https://www.iser.essex.ac.uk/files/iser_working_papers/2006-39.pdf.

Mathiowetz, N. A., & Duncan, G. J. (1988). Out of work, out of mind: Response errors in retrospective reports of unemployment. *Journal of Business & Economic Statistics*, 6(2), 221-229.

Perales, F. (2014). How wrong were we? Dependent interviewing, self reports and measurement error in occupational mobility in panel surveys. *Longitudinal and Life Course Studies*, 5(3), 299–316. http://doi.org/10.14301/llcs.v5i3.295

Pina-Sánchez, J., Koskinen, J., & Plewis, I. (2014). Measurement Error in Retrospective Work Histories. *Survey Research Methods*, 8(1), 43-55. https://doi.org/10.18148/srm/2014.v8i1.5144

Pina-Sánchez, J., Koskinen, J., & Plewis, I. (2019). Adjusting for Measurement Error in Retrospectively Reported Work Histories: An Analysis Using Swedish Register Data. *Journal of Official Statistics*, 35(1), 203-229. https://doi.org/10.2478/jos-2019-0010

Platt, L., Knies, G., Luthra, R., Nandi, A., & Benzeval, M. (2020). Understanding Society at 10 years. *European Sociological Review*, 36(6), 976-988. https://doi.org/10.1093/esr/jcaa031

Pyy-Martikainen, M., & Rendtel, U. (2009). Measurement errors in retrospective reports of event histories. A validation study with Finnish register data. *Survey Research Methods*, 3(3), 139-155. https://doi.org/10.18148/srm/2009.v3i3.2372

Saris, W. E., & Revilla, M. (2016). Correction for measurement errors in survey research: necessary and possible. *Social Indicators Research*, 127(3), 1005-1020.

Słomczyński, K.M., & Tomescu-Dubrow, I. (2018). Basic principles of survey data recycling. In T. P. Johnson, B.-E. Pennell, I. A. Stoop, & B. Dorer (Eds.), *Advances in Comparative Survey Methods* (pp. 937–962). Wiley. https://doi.org/10.1002/9781118884997.ch43

Tomescu-Dubrow, I., Slomczynski, K. M., Sawiński, Z., Kiersztyn, A., Janicka, K., Życzyńska-Ciołek, D., Wysmułek, I., Kotnarowski, M. (2021). The Polish Panel Survey, POLPAN. *European Sociological Review*, 37(5), 849–864. https://doi.org/10.1093/esr/jcab017

Turek, K., Kalmijn, M., & Leopold, T. (2021). The comparative panel file: Harmonized household panel surveys from seven countries. *European Sociological Review*, 37(3), 505-523. https://doi.org/10.1093/esr/jcab006

Wright, L. (2020) *Producing Working Life Histories in the BHPS and UKHLS 2017-2020. [Data Collection]*. Colchester, Essex: UK Data Service. 10.5255/UKDA-SN-854327

Appendix 1. CNB-Young Variable list

Variable	Value	Comments
• **RESPONDENT**		List may be expanded
Respid	ID	
YEAR	Calendar year	
YRBIRTH	Year of birth	
MTBIRTH	Month of birth	
DYBIRTH	Day of birth	Optional
GENDER	Gender	

Variable	Value	Comments
RESIDSIZE[a]	Urban / rural	
RESIDREG[a]	Region of residence	
MARITSTAT[a]	Marital / partnership status	
CHILDREN[a]	Number of children	
• **DATA** [b]		
FSTWAVE	First wave R participated - survey year	First wave R respondent completed the adult questionnaire
LSTWAVE	Last wave R participated - survey year	Survey year is a wave identifier, the year of the interview can be different; possibly more than 1 interview this year
CURWAVE	First interview this year - survey year	
Curwave_Yr	Date of first interview this year: year	
Curwave_Mth	Date of first interview this year: month	
Curwave_Day	Date of first interview this year: day	Optional
NXTWAVE	First interview after this year - survey year	
Nxtwave_Yr	Date of first interview after this year: year	
Nxtwave_Mth	Date of first interview after this year: month	
Nxtwave_Day	Date of first interview after this year: day	Optional
PREVWAVE	Last interview before this year - survey year	
Prevwave_Yr	Date of last interview after this year: year	
Prevwave_Mth	Date of last interview after this year: month	
Prevwave_Day	Date of last interview after this year: day	Optional
• **JOBS**		Code the same set of variables for each job recorded in the survey
JOB01_ID	[Job 01] CNB-Young Job ID Number	codes year of wave and number of job
JOB01_Spell_M	[Job 01] Flag for type of job spell (survey specific)	codes different types of jobs for which different information are collected – see Appendix 2A below
JOB01_Jobtype	[Job 01] Type of job (country specific)	e.g., full-time / part-time / marginal / irregular / gig work
JOB01_Wave	[Job 01] Survey wave (year) data collected	wave / survey year (not year of actual interview)
JOB01_Link_M	[Job 01] [Methodological] Link to job in previous wave	methodological information on how are jobs linked between waves
JOB01_Nwaves_M	[Job 01] [Methodological] Total number of waves in which job was recorded	Based on dependent interviewing or job dates and characteristics
JOB01_XwaveID	[Job 01] ID of [this] job across waves	Based on dependent interviewing or job dates and characteristics
JOB01_firstwave_M	[Job 01] [Methodological] First survey wave (year) in which this job recorded	Based on dependent interviewing or job dates and characteristics
JOB01_lastwave_M	[Job 01] [Methodological] Last survey wave (year) in which this job recorded	Based on dependent interviewing or job dates and characteristics
JOB01_incons_M	[Job 01] [Methodological] Flag for inconsistent info on this job across waves	Detailed documentation will be provided
JOB01_Mos	[Job 01] Total months of work this year	
JOB01_M01_ASC	[Job 01][M01 Jan] Activity Status Code	
JOB01_M02_ASC	[Job 01][M02 Feb] Activity Status Code	See Appendix 3

Variable	Value	Comments
JOB01_M03_ASC	[Job 01][M03 Mar] Activity Status Code	
JOB01_M04_ASC	[Job 01][M04 Apr] Activity Status Code	
JOB01_M05_ASC	[Job 01][M05 May] Activity Status Code	
JOB01_M06_ASC	[Job 01][M06 Jun] Activity Status Code	
JOB01_M07_ASC	[Job 01][M07 Jul] Activity Status Code	See Appendix 3
JOB01_M08_ASC	[Job 01][M08 Aug] Activity Status Code	
JOB01_M09_ASC	[Job 01][M09 Sep] Activity Status Code	
JOB01_M10_ASC	[Job 01][M10 Oct] Activity Status Code	
JOB01_M11_ASC	[Job 01][M11 Nov] Activity Status Code	
JOB01_M12_ASC	[Job 01][M12 Dec] Activity Status Code	
JOB01_start_M	[Job 01][Methodological] Reported date started: previous wave	Flag whether data reported in next wave or retrieved from previous
JOB01_start_Yr	[Job 01] Reported date started: year	
JOB01_start_Mth	[Job 01] Reported date started: month	
JOB01_start_Day	[Job 01] Reported date started: day	Optional
JOB01_end_Yr	[Job 01] Reported date ended: year	
JOB01_end_Mth	[Job 01] Reported date ended: month	
JOB01_end_Day	[Job 01] Reported date ended: day	Optional
JOB01_OCCC	[Job 01] Country Occupational Code	SKZ for Poland, census for the US, etc.
JOB01_ISCO08	[Job 01] ISCO-08 Code	
JOB01_IndCC	[Job 01] Country industry code	
JOB01_Industr	[Job 01] International industry code	Standard Industrial Classification; 2 digits
JOB01_Supvis	[Job 01] Supervises subordinates	
JOB01_HrsWeek	[Job 01] Hours per week	
JOB01_Income	[Job 01] Monthly earnings	at the end of the spell (current earnings for censored spells)
JOB01_Income_M	[Job 01][Methodological] Series of flags for income information	imputed income / wave in which income reported / unit / overtime present
JOB01_Contract	[Job 01] Employment contract	For all surveys except NLSY79 Young Adults
JOB01_Contingent	[Job 01] Is job temporary	For NLSY79 Young Adults
JOB01_Benef_xx	[Job 01] Access to benefits	Series of variables available for the U.S. only
JOB01_Endrsn	[Job 01] Reason for job loss	
JOB01_Newemp	[Job 01] Same employer before	for SOEP, UKHLS and POLPAN
• **ACTIVITIES**		Code the same set of variables for each activity recorded in the survey
ACT01_ID	[Activity 01] CNB-Young ACT ID Number	codes year of wave and number of spell (see codebooks for link to source data)
ACT01_Type	[Activity 01] Type of activity spell	See appendix 2B below
ACT01_Wave	[Activity 01] CNB-Young Wave (year) data collected	wave / survey year (not year of actual interview)

Variable	Value	Comments
ACT01_Link_M	[Activity 01] [Methodological] Link to job spell	methodological information on whether questions on breaks in specific jobs
ACT01_JobID	[Activity 01] ID of the job interrupted	For surveys with items on job breaks
ACT01_Mos	[Activity 01] Total mos of activity this year	
ACT01_M01_ASC	[Activity 01][M01 Jan] Activity Status Code	
ACT01_M02_ASC	[Activity 01][M02 Feb] Activity Status Code	
ACT01_M03_ASC	[Activity 01][M03 Mar] Activity Status Code	
ACT01_M04_ASC	[Activity 01][M04 Apr] Activity Status Code	
ACT01_M05_ASC	[Activity 01][M05 May] Activity Status Code	
ACT01_M06_ASC	[Activity 01][M06 Jun] Activity Status Code	See Appendix 3
ACT01_M07_ASC	[Activity 01][M07 Jul] Activity Status Code	
ACT01_M08_ASC	[Activity 01][M08 Aug] Activity Status Code	
ACT01_M09_ASC	[Activity 01][M09 Sep] Activity Status Code	
ACT01_M10_ASC	[Activity 01][M10 Oct] Activity Status Code	
ACT01_M11_ASC	[Activity 01][M11 Nov] Activity Status Code	
ACT01_M12_ASC	[Activity 01][M12 Dec] Activity Status Code	
ACT01_Dates_M	[Activity 01][Methodological] Reported dates: previous wave	Flag whether data reported in next wave or retrieved from previous
ACT01_start_Yr	[Activity 01] Reported date started: year	
ACT01_start_Mth	[Activity 01] Reported date started: month	
ACT01_start_Day	[Activity 01] Reported date started: day	Optional
ACT01_end_Yr	[Activity 01] Reported date ended: year	
ACT01_end_Mth	[Activity 01] Reported date ended: month	
ACT01_end_Day	[Activity 01] Reported date ended: day	Optional
• **SUMMARIES**		
Y_WorkMos	[Year] Number of months of work	Can be differentiated by type of work
Y_EducMos	[Year] Number of months in education	
Y_UnempMos	[Year] Number of months of unemployment	
Y_PleaveMos	[Year] Number of months of parental leave	
Y_HlthprobMos	[Year] Number of months health problems	Long-term illness, disability
Y_HomeMos	[Year] Number of months homemaker / care duties	
Y_NJobs	[Year] Number of different jobs	Total number of jobs reported this year
Y_Nterm	[Year] Number of job terminations	For the calculation of the precarity index
CY_WorkYrs	[Cumulative] Total work experience in years	
CY_WorkMos	[Cumulative] Total work experience in months	
CY_EduYrs	[Cumulative] Total education in years	
CY_UnempMos	[Cumulative] Total unemployment in months	
CY_PleaveMos	[Cumulative] Total parental leave in months	

Variable	Value	Comments
CY_HlthprobMos	[Cumulative] Total number of months health problems	
CY_HomeMos	[Cumulative] Total number of months homemaker / care duties	
Y_Workincome	[Income] Total income from work	Reported in survey
CY_Workincome	[Derived income] Total income from work	Based on job spell data
ACT_M01_Main	[Activity][M01 Jan] Main activity	
ACT_M02_Main	[Activity][M02 Feb] Main activity	
ACT_M03_Main	[Activity][M03 Mar] Main activity	
ACT_M04_Main	[Activity][M04 Apr] Main activity	
ACT_M05_Main	[Activity][M05 May] Main activity	Main selected based on the first digit of type of spell (see Appendix 2) in the following order: work (1) - education (2) – unemployment (3) – inactivity / all other states / activities (4-7).
ACT_M06_Main	[Activity][M06 Jun] Main activity	
ACT_M07_Main	[Activity][M07 Jul] Main activity	
ACT_M08_Main	[Activity][M08 Aug] Main activity	
ACT_M09_Main	[Activity][M09 Sep] Main activity	
ACT_M10_Main	[Activity][M10 Oct] Main activity	
ACT_M11_Main	[Activity][M11 Nov] Main activity	
ACT_M12_Main	[Activity][M12 Dec] Main activity	
NTY_Workedu	[Transitions] Work to education	In case of overlaps work is the main activity
NTY_Workunemp	[Transitions] Work to unemployment	
NTY_Workinactv	[Transitions] Work to inactivity	
NTY_Eduwork	[Transitions] Education to work	
NTY_Eduunemp	[Transitions] Education to unemployment	
NTY_Eduinactv	[Transitions] Education to inactivity	
NTY_Unempwork	[Transitions] Unemployment to work	
NTY_Unempedu	[Transitions] Unemployment to education	
NTY_Unempinactv	[Transitions] Unemployment to inactivity	
NTY_Inactvwork	[Transitions] Inactivity to work	
NTY_Inactvedu	[Transitions] Inactivity to education	
NTY_Inactvunemp	[Transitions] Inactivity to unemployment	
NTY_Jobjob	[Transitions] Directly between jobs	

[a] Additional variables may be needed to flag cases in which the time-variant respondent characteristics had been taken from a different data source than the next subsequent wave.

[b] Additional methodological files, organized in long format with person-wave units for easy linking with the main CNB-Young datasets, include information on the dates of all survey waves, flags allowing to assess the potential quality of the data for a given respondent by wave (e.g., mode of data collection, proxy survey, etc.), and information on the presence of retrospective life-course survey items that are not included in all the waves (e.g., detailed educational histories in POLPAN are gathered only in 2018, information on all maternity leaves in NLSY79 Young Adults only since 2008, etc.).

Appendix 2. Survey-specific type of spells / activities

A. Job activity spells (codes for variable JOB01_Spell_M in Appendix 1).					
Code	Label	SOEP	BHPS UKHLS	NLSY	POL-PAN
10	**working**				
11	current main job in wave	x	x	x	x
12	past (terminated) job in wave	x	x	x	x
13	current secondary job in wave	x	x	x	x
14	additional job spells with limited data collection [a]	x	x	x	x
B. Other activities (codes for variable ACT01_Type in Appendix 1).					
Code	Label	SOEP	BHPS UKHLS	NLSY	POL-PAN
20	**in education**		x	x	x
21	in school / university / vocational school;	x	x	x	x
22	vocational training / apprenticeship;	x	x		
23	in further training / retraining / occupational training;	x	x	x	x
30	**unemployed**				
31	unemployed: self-defined		x		x
32	registered unemployed	x			
33	unemployed looking for work			x	x
40	**maternity / parental leave**		x	x	
41	maternity / parental leave: paid	x		x	x
42	maternity / parental leave: unpaid / paid statutory minimum			x	x
50	**in retirement / early retirement / disability pay;**	x			
51	retirement		x		x
52	disability pay		x		x
53	not working due to long-term illness, health condition				x
60	**homemaker**	x	x		x
70	**military service / voluntary service**			x	x
71	voluntary social year / federal volunteer service	x			
72	military inactive duty			x	
73	military training			x	

[a] These job spells differ across surveys. In the UKHLS, they code information related to gig work, which is covered in a separate module of the questionnaire independently of the activity status history items. In SOEP, these include spells on which we only possess information from the monthly calendar items. In NLSY, this concerns, among others, military jobs, for which occupational and income data are coded in a different way. Finally, in POLPAN, certain jobs performed abroad are recorded independently from the main career module in the data collection. All such spell types have distinct codes and are described in the codebook / documentation.

Note: while we record all the spells that include longitudinal information, the CNB-Young layout does not include additional information on activities that are gathered only at the time of the survey. An example is offered by search for employment in SOEP, which (contrary to NLSY or some waves in POLPAN) is observed only as a current (ongoing) activity, without information allowing to compute or estimate the number of months spent in this activity. Such additional information on activities at the time of the survey can easily be added to the dataset by researchers willing to do so, via matching by respondent ID and the year of the interview.

Some of the spells are not recorded in all the survey waves (e.g., military service in POLPAN only up to 2008).

Appendix 3. A full list of Activity Status Codes (ASC)

000	no activity this month
101	intermediate month of activity
111	first month of activity reported
112	first month of activity censored (job started before date of last interview)
121	last month of activity reported
122	last month of activity censored (job current in interview in which it was reported)
131	first and last month of activity reported (activity lasting one month)
132	first or last month of activity censored
-001	missing data on whether activity present this month & year
-101	missing data on whether activity present this month but activity present this year