The Comparative Panel File: Harmonized household panel surveys from seven countries

Turek, K.L.; Kalmijn, M.; Leopold, T.

published in
European Sociological Review
2021

document version
Publisher's PDF, also known as Version of record

European Sociological Review, 2021, Vol. 37, No. 3, 505–523
doi: 10.1093/esr/jcab006
Advance Access Publication Date: 12 March 2021
Data Brief

The Comparative Panel File: Harmonized Household Panel Surveys from Seven Countries

Konrad Turek [1,*], Matthijs Kalmijn [1] and Thomas Leopold[2]

[1]Netherlands Interdisciplinary Demographic Institute (NIDI-KNAW/University of Groningen), Lange Houtstraat 19, NL-2511 CV The Hague, The Netherlands and [2]University of Cologne, Institute of Sociology and Social Psychology, Albertus-Magnus-Platz, 50923 Köln, Germany

*Corresponding author. Email: k.l.turek@rug.nl

Submitted December 2020accepted February 2021

Abstract

The Comparative Panel File (CPF) harmonizes the world's largest and longest-running household panel surveys from seven countries: Australia (HILDA), Germany (SOEP), United Kingdom (BHPS and UKHLS), South Korea (KLIPS), Russia (RLMS), Switzerland (SHP), and the United States (PSID). The project aims to support the social science community in the analysis of comparative life course data. The CPF builds on the Cross-National Equivalent File but offers a larger range of variables, larger and more recent samples, an easier and more flexible workflow, and an open science platform for development. The CPF is not a data product but an open-source code that integrates individual and household panel data from all seven surveys into a harmonized three-level data structure. The CPF allows analysing individual trajectories, time trends, contextual effects, and country differences. The project is organized as an open science platform. The CPF version 1.0 contains 2.7 million observations from 360,000 respondents, covering the period from 1968 to 2019 and up to 40 panel waves per respondent. In this data brief, we present the background, design, and content of the CPF.

Introduction

In the past few decades, life course research has expanded enormously (Mayer, 2009). This trend is evident in research on multiple life domains, including union and family formation, education and labour market behaviour, and health and wellbeing. The expansion of life course research has been fuelled by societal trends, in particular changing gender roles and rising female labour force participation, globalization and rising uncertainty in the labour market, declines in marriage and fertility, and increasing life expectancy and population ageing. In light of these macro-level trends, attention in cross-national differences has increased. Comparative life course studies have tried to explain differences in terms of economic, cultural, and institutional conditions (Blossfeld and Hakim, 1997; Aassve *et al.*, 2002; Whelan, Layte and Maitre, 2004; Blossfeld *et al.*, 2005; Andress *et al.*, 2006; Liefbroer and Dourleijn, 2006; Boye, 2011; Mohring, 2016; Perelli-Harris and Lyons-Amos, 2016; Leopold, 2018).

To study life courses in comparative perspective, longitudinal data from multiple countries need to be (made) comparable (Slomczynski and Tomescu-Dubrow, 2019). Several large-scale data collections have provided a major boost in this respect (Burkhauser and Lillard,

2005; Dubrow and Tomescu-Dubrow, 2015). The Generations and Gender Survey (GGS) includes retrospective data on life courses in twenty countries. The GGS has also introduced a prospective element and now includes two waves. The Survey of Health, Ageing and Retirement (SHARE) is a prospective survey in 29 countries which has collected seven biannual waves of panel data. In each wave, new countries were added. Both datasets are widely used, but they also have limitations. The GGS is mostly retrospective, which limits the number of variables that can be measured longitudinally (e.g., no income and no wellbeing). The SHARE samples the population aged 50 years and over, which limits the analysis to later stages of the life course.

An alternative approach is to analyse and compare panel surveys from multiple countries. Several countries have collected prospective panel data. The Panel Study of Income Dynamics (PSID) in the United States and the Socio-Economic Panel (SOEP) in Germany served as examples for similar studies in other countries, including the United Kingdom, Australia, Switzerland, Russia, and South Korea. Household panel studies interview their respondents each year, have done so for long periods, cover various life domains, and include all age groups and multiple cohorts. With such a broad scope and long-term perspective, they offer an excellent data source for analyses of changes over the life course.

The comparative analysis of such panel data requires harmonization, i.e., merging equivalent or similar variables into a single dataset with a unified data structure. One prominent endeavour in this respect is the Cross-National Equivalent File (CNEF) from Ohio State University which combines annual household panel studies from eight currently running surveys (Frick *et al.*, 2007).[1] Despite its many merits, the CNEF is limited by its focus on mainly economic topics, its decision to include only measures that are fully equivalent across all surveys, and little flexibility in data handling the data.[2]

In this data brief, we present the Comparative Panel File (CPF)—an open science project aimed at answering the growing need for cross-nationally comparative longitudinal data in the social sciences. The core of the CPF is a Stata code that harmonizes comparative life course data from seven major household panel surveys. CPF includes the same surveys as CNEF[3] but provides an extended harmonization approach focused on two principles. First, it is *inclusive*, covering a broad range of variables and multiple life domains. An inclusive approach means that harmonization is not limited to full equivalence across surveys. In this way, we were able to add several key constructs, including detailed measures of work and family life, work quality, health, wellbeing, and social background. Second, it is *open and dynamic*. By providing an extensively documented open-source code that can be modified and improved by users, the CPF is flexible, easily adjustable, and can be extended in various ways. Contrary to CNEF-data releases, users can modify and add variables, as well as include more recent samples on their own. This open and dynamic design supports technical and substantive solutions for comparability problems. It also contributes to open and replicable science by providing access to data resources and collaborative improvement of research tools. CPF was built by researchers for researchers as a bottom-up initiative that allows to fully utilize the power of open-science community.

Harmonizing national panels has an enormous value for life course research (Burkhauser and Lillard, 2005; Dubrow and Tomescu-Dubrow, 2015; Bryan and Jenkins, 2016; Kühne *et al.*, 2020). Kühne *et al.* (2020) emphasize the importance of household panel data in the context of individual and societal challenges, especially collectively experienced crises such as the global financial crisis and the COVID-19 pandemic. Life course researchers are increasingly interested in continuity and change across individuals' lives, long-term trajectories, and distal outcomes of early life events. Analysts have called for more comparative data that allow studying the interplay between individual and contextual factors, such as social structures, institutional arrangements, public policies, or socio-cultural contexts (Bernardi, Huinink and Settersten, 2019; Piccarreta and Studer, 2019). This growing need is intensified by the development of longitudinal and multilevel methods of analysis (Gelman and Hill, 2007; Bryan and Jenkins, 2016; Brüderl, Kratz and Bauer, 2019). Providing high-quality data required for the use of these models is costly, especially in a cross-national design (Kühne *et al.*, 2020), rendering the harmonization of high-quality secondary data from household panel studies an appealing alternative (Dubrow and Tomescu-Dubrow, 2015).

The CPF aims to provide this alternative, maximizing the comparative research potential of the world's largest and longest-running household panel surveys from seven countries:

- Australia [The Household, Income and Labor Dynamics in Australia Survey (HILDA)],
- Germany [The German Socio-Economic Panel (SOEP)],
- the United Kingdom [The British Household Panel Survey (BHPS) and Understanding Society—The UK Household Longitudinal Study (UKHLS)],

- South Korea [The Korean Labor and Income Panel Study (KLIPS)],
- Russia [The Russian Longitudinal Monitoring Survey (RLMS)],
- Switzerland [The Swiss Household Panel (SHP)], and
- the United States [The Panel Study of Income Dynamics (PSID)].

The CPF provides free and full access to a code that generates a comparative dataset based on these household panel surveys. The code and complete documentation are available at www.cpfdata.com. The project is organized as an open science platform that integrates tools for general communication (online Forum), code development (*GitHub* code repository), and general management of scientific research [Open Science Framework (OSF)]. After securing access to the national panel surveys, users can run our code which combines datasets and waves within a country, constructs harmonized variables, and merges these into one data set for all countries and all waves. The file is organized in a long format containing one record for each person in each wave. The merged file contains data on approximately 360,000 individuals observed at an average of 7.5 waves.

The Idea of CPF

The idea originated in 2019 in the context of the project 'Critical Life Events and the Dynamics of Inequality: Risk, Vulnerability, and Cumulative Disadvantage' (CRITEVENTS). CRITEVENTS is funded by NORFACE through the transnational research programme 'Dynamics of Inequality Across the Life-Course: Structures and Processes (DIAL).'[4] The initial motivation to harmonize data from national panel surveys was the fact that the CNEF release did not include measures for job loss and unemployment. Instead of harmonizing only these variables, we decided to extend the approach pioneered by CNEF to a larger set of key variables of social science research and make the result available to the broader scientific community. Currently, CPF is developed by Konrad Turek and Matthijs Kalmijn at the Netherlands Interdisciplinary Demographic Institute (NIDI-KNAW) and Thomas Leopold at the University of Cologne. The CPF code and entire open science platform were designed and prepared by Konrad Turek and will be continuously developed and improved by the CPF team and the community of users.

The CNEF is a long-running and well-established project which harmonizes international longitudinal surveys of households (Burkhauser *et al.*, 2001; Frick *et al.*, 2007). It has been developed since 1990 under the lead of researchers from Cornell University. Over the years, the project was managed primarily by Dean R. Lillard and administered by Cornell University and Ohio State University.[5] Initially, in 1991, the dataset harmonized only a limited set of variables for two countries, the United States and Germany.[6] Over the years, the project expanded by adding countries, such as the United Kingdom and Canada[7] in 1999, Australia and Switzerland in 2007, and Russia, South Korea, and Japan in later years. The set of topics and variables has been gradually extended, but the main focus remains on income and earnings. CNEF has been used primarily in income-related research in economics (Büchel and Frick, 2004; Chen, 2009; Allanson, 2011), sociology (DiPrete and McManus, 1996; McCall and Percheski, 2010; Ehlert, 2013; Musick, Bea and Gonalons-Pons, 2020), or demography (Cooke *et al.*, 2009), and less often in research on other topics, such as life satisfaction (Cho and Lee, 2013) and self-employment (McManus, 2003).

The CPF project makes several steps forward. First and foremost, CPF has a broad focus, including information about education, family and marital relationship, labour market status, subjective wellbeing and work satisfaction, social origin, and socio-economic status, in addition to the classic economic variables included in CNEF. For several of these variables, CPF also offers more detail; for example, it allows distinguishing between unemployed, retired, self-employed, and entrepreneurs. Second, CPF is open and flexible, thereby facilitating a genuine bottom-up approach. CPF fully supports modifications in harmonized variables or adding new variables from the source database, depending on researchers' needs. Our code is available in full and for all selected countries. It also facilitates work with single surveys, as it provides instructions about how to go from a large set of raw files to an integrated panel data set ready for analysis (for some surveys, e.g., the PSID, this is a complex process). In contrast, CNEF is a data product that offers a set of separate data files but only parts of the code are available. Third, CPF does not depend on direct government funding, which greatly facilitates the speed and direction of its further development. New waves can be added as soon as these are released by the national data centres. CNEF files are released differently by country, and most do not cover recent waves. Fourth, procedures to obtain and use the CPF are streamlined and greatly simplified. The only administrative step needed is obtaining permission to use

the national data from each of the seven national data centres. This may be some work, but when permission is obtained, the openly available code can be run, and the CPF is readily available on the user's computer. CNEF requires an additional application for accessing some surveys, separate CNEF files are provided partly online and partly on a CD sent by mail, and they still have to be integrated.

In sum, we build on the approach pioneered by CNEF and other cross-national data harmonization projects (Dubrow and Tomescu-Dubrow, 2015), but overcome the main limitations for users who require a broader set of variables and more flexibility and control over the data management process. Users can either follow the default workflow and run the code unchanged or modify and improve it for their use (e.g., select countries, add new waves, add new, or modify existing variables). Additionally, CPF is an open science project organized around an online platform (www.cpfdata.com) comprising a website, online Forum, GitHub repository, and OSF. CPF's platform provides tools that support open collaboration, management, documenting and sharing all projects materials. It also facilitates recording user's improvements and suggestions which can be incorporated and shared to allow continuous development and regular updates to the official versions of the code. The CPF platform and the first version of the CPF code were published in December 2020.

Design and Content

CPF data sources

CPF harmonizes household panel studies—general population repeated surveys with household as the primary sampling unit. They regularly (mostly yearly) interview all or selected adult members of sampled households over long periods and collect information about the entire household and its members (Rose, 1995). Version 1.0 of the CPF combines seven most established and longest-running household panel studies globally. All studies are representative of the population of households. As ongoing panel studies, they continuously renew their samples by including new household members (e.g., grown-up children, newly married partners), following new independent household established by respondents (e.g., children leaving parents' homes), by refreshments (e.g., including a new set of households), or by extensions (e.g., including a new type of households, such as new migrant families). Many panels included systematic oversamples of subgroups; these are included in the CPF but identifiable with country-specific variables.

With over 50 years of data collection, the oldest national panel survey is the US's PSID. PSID was initiated to evaluate poverty and economic wellbeing dynamics in the United States, and the first wave was conducted in 1968. Over the five decades, the PSID sample has grown through its genealogic design that allows gathering data from up to seven generations of the same family. It has collected survey information on more than 80,000 individuals (McGonagle *et al.*, 2012; Johnson *et al.*, 2018). The second oldest study is the German SOEP, which began in 1984. Initially, SOEP included only Western Germany, and since the 1990 reunification, it also covers the eastern part of the country, being the only database worldwide covering such a political unification (Giesselmann *et al.*, 2019; Goebel *et al.*, 2019; Siegers, Belcheva and Silbermann, 2020).

PSID and SOEP served as examples for the development of household panel studies in other countries. This included BHPS which began in 1991 and after collecting 18 waves, since 2009, it was integrated into UKHLS. Both of them are included in the CPF as the UK sample. With a target sample size of 40,000 households in wave 1, UKHLS became the largest nationally representative household panel study worldwide (Buck and McFall, 2012; Platt *et al.*, 2020). The Russian RLMS was initiated in 1992, and the first full wave was conducted in 1994. RLMS is the longest-running panel survey of households in Eastern Europe and Asia (Gerry and Papadopoulos, 2015; Kozyreva and Sabirianova Peter, 2015).

CPF also includes younger, but already well-established studies. The KLIPS began in 1998 and has served as an essential source of information to develop and evaluate labour market policies in South Korea (KLI, 2020). The Swiss SHP started as extensive and vital research in 1999 (Tillmann *et al.*, 2016; Voorpostel *et al.*, 2020). Finally, the Australian HILDA that commenced in 2001 is the youngest survey included in the CPF at the moment; it collected already 18 waves (Watson and Wooden, 2020).

All of the studies are extensively used in research. For example, the PSID data were used in almost 5.5 thousand peer-reviewed publications in total until late 2018, very often in the top economic and sociological journals (Johnson *et al.*, 2018). SOEP-based publications amount currently to between 300 and 400 annually (Goebel *et al.*, 2019). Between 2006 and 2018, there were 50 articles based on SOEP published in the European Sociological Review, mostly taking a life course perspective (Giesselmann *et al.*, 2019).

Table 1. Number of waves, observations, and respondents

Country	Survey	First wave	No of waves	Observations		Unique respondents	
				n	%	*n*	%
[1] Australia	HILDA	2001	18	257,418	9.6	30,576	8.5
[2] Korea	KLIPS	1998	21	257,495	9.6	23,535	6.5
[3] United States	PSID	1968	40	457,638	17.0	42,219	11.7
[4] Russia	RLMS	1994	23	274,914	10.2	44,559	12.4
[5] Switzerland	SHP	1999	20	146,765	5.4	21,900	6.1
[6] Germany	SOEP	1984	35	675,693	25.1	94,525	26.3
[7] United Kingdom	BHPS/UKHLS[a]	1991	27	626,787	23.2	102,605	28.5
Total				2,696,710	100	359,919	100

[a]BHPS: 1991–2008, 18 waves and UKHLS: from 2009, 9 waves.

CPF version 1.0 was built on data versions released in 2020, i.e., HILDA version 180, KLIPS version 21, PSID version 2017, RLMS version 2018, SHP version 20, SOEP version 35, and UKHLS version 8. New waves will be continuously integrated into the CPF code; users can also do this independently (see *Workflow C* in *Using the CPF*).

Samples

The CPF is a comparative panel dataset with a three-level hierarchical data structure: repeated individual observations from multiple waves (level 1) are clustered within individuals (level 2), and individuals are clustered within countries (level 3). The CPF version 1.0 covers up to 40 waves (between 1968 and 2018), combines seven countries, and includes around 2.7 million observations from almost 360,000 respondents (Table 1, see also Supplementary Table A1). In the default settings, CPF includes observations from individuals aged 18 and older and meet the following criteria for the interview status:

a. South Korea, Russia, Switzerland, and the United Kingdom: keep all observations (including proxy respondents).
b. Australia and Germany: keep direct respondents only.
c. The United States: only reference persons (heads) and partners (spouses).

Additionally, observations with missing values for gender and age are deleted. Users can easily modify these selection criteria (see: Workflow D—Adjustments to sampling criteria).

An average respondent participated in 7.5 waves (between 6.2 in Russia and 10.9 in South Korea). Out of all 359,919 respondents, 91,625 participated in a minimum of 10 waves and 28,209 in a minimum of 20 waves (Figure 1). Country-specific lines indicate the exact number of waves for which the dataset provides information on sample members. For example, almost 22,000 respondents in the UK participated in only one wave and nearly 16,000 participated in exactly nine waves. In Australia, 18 per cent of the sample (ca. 5,500 respondents) participated in all 18 waves of HILDA.

The oldest survey in CPF, the PSID, covers a period from 1968 and has collected 40 waves until now (Figure 2, see also Supplementary Table A2). The youngest panel study in CPF is HILDA with 18 waves since 2001. CPF includes four countries since 1994, five countries since 1999, and all seven countries since 2001. A substantial increase in the sample size observed for the United Kingdom in 2009 is related to the transition from BHPS to UKHLS. For most of the surveys, data have been collected yearly (after 1997, PSID has switched to 2-year intervals in data collection).

Variables

The goal of CPF is to harmonize variables across surveys. We based our approach on the CNEF but aimed at extending the range of variables included. For example, instead of a simple indicator for being employed or not employed, CPF provides a full range of labour market statuses, including being unemployed, retired, in education, or inactive. Instead of years of schooling, CPF focuses on education level according to the ISCED classification, a measure that has been designed for cross-national comparison. We provide more detailed information on marital status. CPF provides a set of additional variables, such as training participation, satisfaction with different domains, social origin, labour market experience, self-employment and entrepreneurship, work-education skill fit, and perception of job security. An overview of all variables is presented in Table 2.

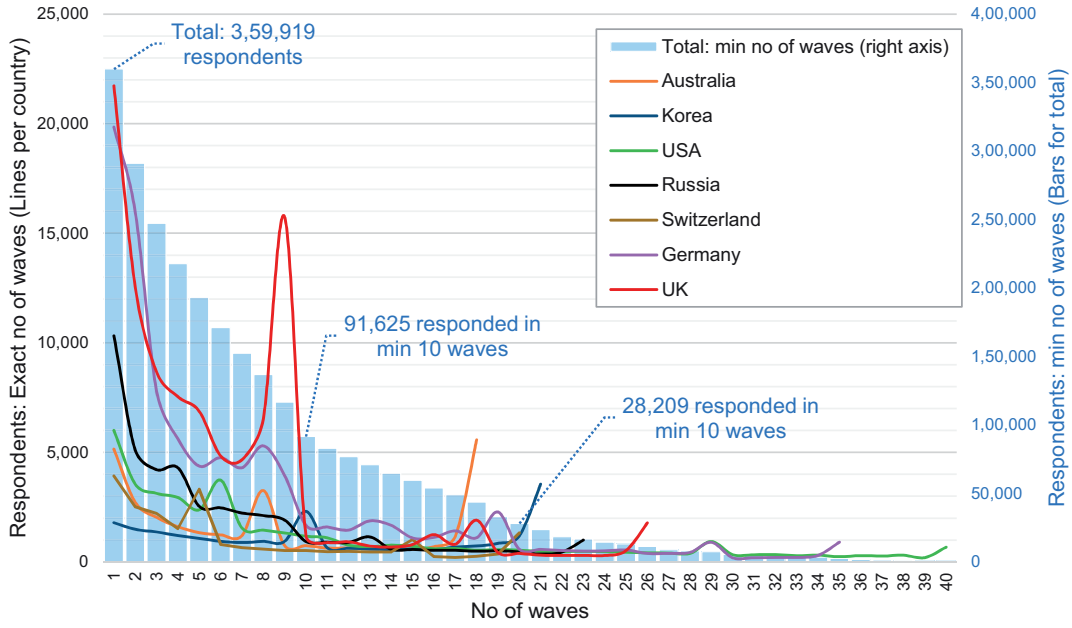

Figure 1. Number of waves in which individuals participated: exact number by survey (left axis) and minimum number for the total sample (right axis)

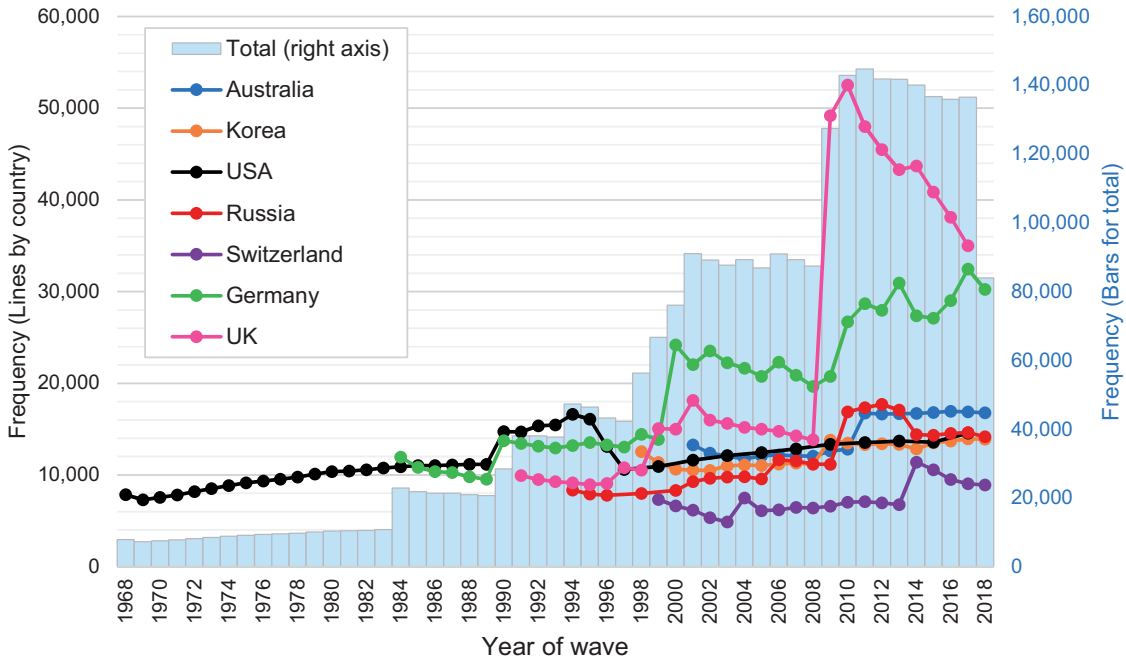

Figure 2. Timeline of the data and number of observations by wave

Table 2. Variables available in the CPF version 1.0

Group of variables	Description	Main variables
Technical	Respondent identifiers, information about wave and interview and other technical information	• Country • Personal and household identification numbers • Wave's number and year • Interview status • Year and month of interview • Sample identifiers
Demographic	Basic demographic characteristics	• Gender • Age • Year of birth
Education	Education level is harmonized using the ISCED classification in four different versions with three, four, and five levels. For example, three levels are (0–2) low, (3–4) medium, and (5–8) high. Variables also include years of education, participation in training, self-assessment of qualifications	• Education: 3/4/5 levels • Participation in training in the past 12 months • Work-education skill fit • Qualifications for job
Marital and relationship status	CPF distinguishes between formal marital status and partnership living-status, which also accounts for living with the partner. Additionally, it includes less precise primary partnership status equivalent to the one used in CNEF. Also, it provides indicators for specific statuses (e.g., divorced) and being never married	• Formal marital status • Partnership living-status • Primary partnership status • Living with the partner • Never married • Widowed • Divorced • Separated
Number of children and household members	There are several children-related variables to account for differences in questionnaires in: • the definition of children, e.g., own-born, adopted, of other family members, any children • the situation of children, e.g., living currently in the household, living elsewhere, children ever had • age of children, e.g., any age, below 18, and below 15 years old	• Number of children in household (aged 0–15, 0–17) • Number of children ever had • Has own children (yes/no) • Number of people in household
Labour market situation and employment	An important goal of the CPF is to provide a comprehensive view of individuals' labour market situation. These include the following areas:	
	Labour market situation: employed, unemployed, retired or disabled, in education, not active, employed but on leave. CPF also identifies maternity leave	• Labour market situation (5/6 categories) • Currently working (self-reported) • Working in the previous year (based on reported working hours) • Being on maternity leave • Never worked
	Level of employment: full- or part-time, number of working hours (several versions, including actual and contracted hours)	• Full- or part-time work (based on working hour/self-reported) • Number of working hours (per year, month, week, day) • Work hours per week: contracted

(continued)

Table 2. (Continued)

Group of variables	Description	Main variables
	Occupation—classified according to the International Standard Classification of Occupations (ISCO). KLIPS and PSID use different classifications than ISCO. In these cases, crosswalk algorithms were developed. ISCO level 1 and 2 are harmonized for all countries, but if available, CPF provides a more detailed classification in versions ISCO-88 or ISCO-08 at 3- or 4-digit levels	• Occupation: ISCO level 1: 1 digit, 10 categories • Occupation: ISCO level 2: 2 digits, 50+ categories • Additionally, ISCO-08/ISCO-88 with 3 or 4 digits • Supervisory position
	Characteristics of the employee's organization	• Industry: 3 major, 10 sub-major and 17 minor groups • Sector (public) • Size of organization
	More precise and specific identification of actively unemployed, self-employed, entrepreneurs (with employees), and retirees. These indicators are built on information from several variables. For example, individuals are classified as retired when they are not working and meet any of the following criteria: • Self-categorization as retired and age 50+ • Receives old-age pension and age 50+ • Age 65+	• Unemployed: actively looking for work • Self-employed • Entrepreneur (including or not including farmers) • Retired fully • Receiving old-age pension
	Labour market experience measured as years of employment/work	• Total labour market experience (total/full time/part time) • Tenure with current employer
	Perception of job security—whether the respondent is worried about job security (in two versions)	• Secure/insecure • Secure/insecure/hard to say

Incomes	Incomes of individuals and households. Depending on the origin data, information on individual income is included in several variables based on:
	• source of income (total income from jobs and benefits, from all jobs, from the main job)
	• type of income (gross, net)
	• reference period for income (year, month, per hour)
	This approach results in multiple variables but provides clear definitions. For analytical purposes, users can combine particular variables using the nominal values or relative values (e.g., percentiles). CPF provides values as they are included in the source data, without any additional cleaning, imputation, conversion, or inflation-adjustments. Values are in local currency
	Depending on the type of monthly household income in the origin data, information is provided in two versions: before taxes and deduction (gross, pre), after taxes and transfers (net, post). Some datasets provide a negative household
		• Individual income (all types)
		  • Year, net
		  • Month, net
		• Individual labour earnings (all jobs)
		  • Year, gross
		  • Year, net
		  • Month, net
		  • Month, gross
		• Salary from the main job
		  • Year, net
		  • Year, gross
		  • Month, gross
		  • Month, net
		  • Per hour, gross
		• Household income (month)
		  • Gross
		  • Net

(continued)

Table 2. (Continued)

Group of variables	Description	Main variables
	income indicating a loss or debit (e.g., PSID since 1994). Values are in local currencies	

Health and wellbeing	Self-rated health status is based on the standard 5-point scale
	There are three versions of disability-related questions
	Variable for chronic diseases is in a working version: it is not fully harmonized and should be modified by the users according to specific conceptual framework (e.g., defining chronic conditions)
	CPF provides several dimensions of subjective well-being, which can be harmonized for at least several countries. We include two versions of each variable due to differences in original answer scales: with a 5-point scale (1–5 range) and 11-point (0–10 range). If required, the original values were rescaled
		• Self-rated health
		• Receiving disability pension
		• Disability: any type (physical, mental, or nervous condition)
		• Disability: min. category 2 or greater than 30 per cent
		• Chronic diseases (yes/no)
		• Satisfaction with
		  • Life
		  • Work
		  • Financial situation of household
		  • Individual income
		  • Family
		  • Health

Parental background	Parents' education level is coded in 3- and 4-categorical variables similarly to respondent's education level	• Mother's/father's education: 3/4 levels
Socio-economic position	Socio-economic position scales are based on respondents' work status and occupation's ISCO code	• International Socio-Economic Index of occupational status (ISEI) • Treiman's international prestige scale (SIOPS) • German Magnitude Prestige Scale (MPS)

During harmonization, we explored various items available in the source data to verify their comparative potential. Some questions had a very similar form across all surveys (e.g., self-rated health), but many differed in the wording or number of answer categories. In the latter case, we assessed the comparative value and compared distributions of responses. It is important to keep in mind that descriptive statistics of harmonized variables may differ across countries if the original variables had different numbers of answering categories in different countries (Revilla, Saris and Krosnick, 2014). Similarly, differences in the wording of questions may produce differences in the frequency distributions. Correlations with other variables are not necessarily affected by such differences (Kaminska and Lynn, 2017; Wolf *et al.*, 2017; Slomczynski and Tomescu-Dubrow, 2019). We advise users to read the Codebook to be aware of such differences.

Since full harmonization was not always possible, some items are available for a subset of countries. Many of the CPF variables are composed of multiple source variables. For example, retirement is based on information about working status, self-reported retirement status, receiving retirement pension, and age. In many cases, the CPF's code includes data cleaning, such as updating contradictory entries with the most reliable information, filling missing values based on information from other waves or other variables (e.g., for education, age, year of birth, marital status). Users can modify existing or add other variables from the source data by developing the open CPF code (the procedure is described below in *Workflow D*). Detailed information on all variables is provided in the *CPF Codebook*.

Organization of the code

The syntax is designed at two levels: higher and lower (Figure 3). Two *higher-level syntaxes* are short and do not refer directly to variables or data files. Instead, they work as an interface and allow to fill in the necessary information (e.g., file directory) and setup options for harmonization (e.g., which surveys to include). As meta-

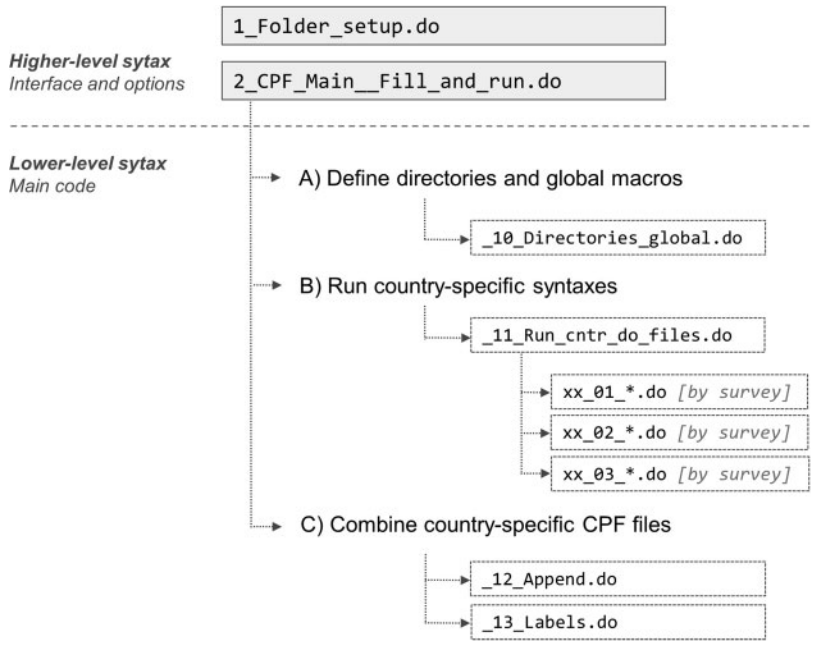

Figure 3. Structure of the CPF syntax

level codes, they call all the necessary codes from a more complex structure of lower-level syntaxes.

For each survey, there are separate *lower-level syntaxes*, and the algorithms are designed differently. However, they all lead through the same three steps: the first constructs initial separate country data in a long format by merging original files, the second harmonizes variables within countries according to the unified template, and the third selects comparative samples. The process results in separate datasets with the same data structure for each country. Then, all country files can be combined into the single CPF harmonized dataset using a higher-level syntax.

Using the CPF

The CPF provides the syntax (programming code in Stata do-file format), the *Manual* explains how to work with the syntax, and the *Codebook* describes all variables. Figure 4 presents a step-by-step guidelines for working with the CPF (for details, see the *Manual*). Users must first apply for access to each of the original datasets independently at national administrator institutions. Access is free of charge, but in most cases, users must describe their research goals and sign a contract. When access is granted, the first syntax can be run to set up a folder structure where original survey files can be extracted. Then, users can easily follow the instructions to build the comparative file in the default way or modify the procedures according to their needs. In the latter case, the code's hierarchical design allows quickly locating all the steps in the algorithms. Country-specific syntaxes are commented and organized in a similar way to facilitate the work.

There are four general ways of working with the CPF syntax, which are called workflows. *Workflow A* describes the default and basic approach which constructs the data without any modifications. Users first have to fill in the necessary information, such as the directory in the first syntax (*1_Folder_setup.do*) and run it to create an appropriate folder structure. Then, they can place the downloaded data in specific folders. The next step uses the second syntax (*2_CPF_Main__Fill_and_run.do*) to call all lower-level syntaxes (*_10*, *_11*, *_12*, and *_13*). Within this structure, syntax *_11* calls all country-specific syntaxes (multiple do-files numbered *_01*, *_02*, *_03*). Many users can be interested in syntaxes *_01* which contain code that integrates all raw files into single and ready for analysis (yet unharmonized) country data sets.

The other workflows serve to modify or add data. *Workflow B* allows selecting surveys to be included in the harmonized CPF dataset by changing the list of surveys in syntax 2. For example, to harmonize PSID,

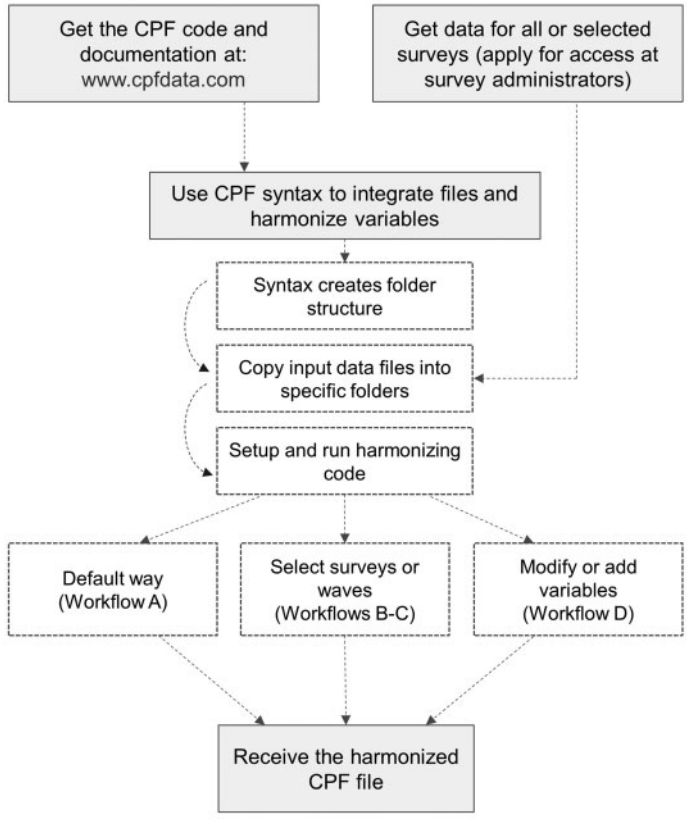

Figure 4. A step-by-step guide through using the CPF code

RLMS and UKHLS, users only have to keep respective names in the code (e.g., global surveys 'psid rlms ukhls') and the rest of the procedure is limited to the default running of the entire code in syntax 2. *Workflow C* serves to add new waves when they become available for the surveys. The CPF code will be regularly adjusted to incorporate new waves, users can also modify the syntaxes on their own. In most cases, the procedure should be limited to filling-in information in syntax 2 on the number of waves (e.g., global soep_w '35' for 35 waves of SOEP) or updating filenames. However, if the approach to naming variables, folders or data files in the original data changes in the future, additional adjustments have to be made in higher-level syntax 2 and/or lower-level syntaxes _01. *Workflow D* refers to all other modification of the existing structure of the CPF data. Users can modify variables, add new ones, or modify the criteria for sample selection. Any adjustments of this type must be made in the lower-level syntaxes, separately and consistently for all surveys and the master-syntaxes. The *Manual* and the code's comments include appropriate instructions. The outcome is always a hierarchical long data file with a set of harmonized variables. Note that running the entire CPF code can easily take an hour or two on an average computer. It is recommended to have at least 80 GB storage hard drive space if all countries are included (the original data files require minimum 50 GB and the CPF working and output files need additional 25 GB).

Basic Characteristics of the Data

Respondents' birth year stretches from 1870 to 2001 (Figure 5). The oldest cohorts are included in the longest-running PSID and SOEP, but most of the studies contain high numbers of respondents from cohorts between the late 1920s and early 1990s. Table 3 shows the distribution of age groups and birth cohort. We can

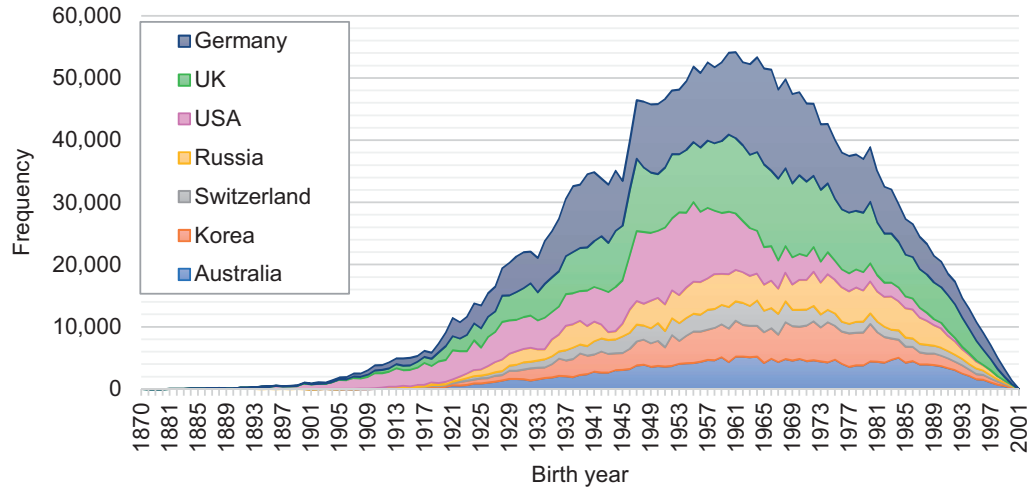

Figure 5. Distribution of birth cohorts (year of birth) by survey

trace each cohort as it ages, not for the entire age range, but when cohorts are pooled, the full age range can be observed.

Table 4 presents the basic socio-demographic characteristics of the sample. Cross-country harmonization of variables is particularly challenging for constructs that are based on different measurement methodology or that refer to different institutional arrangements. This is often the case for occupation and education classifications. Occupations in CPF are harmonized according to the International Standard Classification of Occupations (ISCO). For KLIPS and PSID which use different classifications, we developed crosswalk algorithms into ISCO at 1- and 2-digit levels. Some structural differences, however, remain. For example, the number of South Korea employees classified as managers is very low (1.5 per cent) compared to other countries (which corresponds to the official Korean statistics).[8] In case of education, we focused on harmonization according to the International Standard Classification of Education (ISCED), which is usually more relevant for comparative analysis than years of education. Due to differences in the approach and precision of measurement, it is not possible to harmonize education at all ISCED levels. In CPF, we created variables with three, four, and five categories, but we recognize that alternative approaches are possible. It is therefore essential that the open-source code facilitates such modification.

Open Science Platform for CPF

CPF is an open science project, which means that it provides access to all resources, including the programming code. Furthermore, the code can be improved and developed by anyone who wishes to contribute to the project. To allow the open access and community-based development, we have built an open science platform that connects several tools: website, online Forum, GitHub, and OSF (Figure 6).

The central element is the project's website (www.cpfdata.com) that contains all important information and news. It also provides access to the latest documentation and the major version of the code (which are stored at the GitHub repository). The website also includes an online Forum (www.cpfdata.com/forum) that serves general communication, discussions, questions, and suggestions related to the project.

GitHub (www.github.com/cpfdata) is a code hosting platform, especially useful for collaborations in open-source code development. It allows users to access the main and alternative versions of the code, share their modifications, track changes, and continuously integrate them into consecutive versions. The main function of this platform for the CPF is to host the code and provide tools for its further development. Users can offer extensions, improvements, or alternative versions of the code and all changes are recorded, providing version control functionality. GitHub also stores projects' up-to-date documentation.

OSF (www.osf.io/h3yxq) is one of the most popular open science platformsfacilitating open collaboration in research. OSF integrates many tools and services which support managing, organizing, documenting, and sharing all aspects of a project. OSF allows pre-registering studies, storing code, and data; it is linked to preprint services and many scientific platforms. These features

Table 3. Frequency by age groups and birth cohort

Age group	Birth cohort							
	1920s and earlier	30s	40s	50s	60s	70s	80s	90s and later
Australia								
18/29						7,511	30,531	20,235
30/39					9,802	23,432	12,898	
40/49				9,781	26,255	11,800		
50/59			7,233	22,881	12,034			
60/69		4,685	17,340	9,738				
70/79	3,597	10,427	6,716					
80/max	7,227	3,295						
Korea								
18/29					293	15,300	21,128	10,119
30/39				318	15,988	27,590	10,440	
40/49			204	14,560	24,494	12,833		
50/59		176	9,703	21,296	11,032			
60/69	87	7,479	15,622	9,231				
70/79	3,317	10,989	6,871					
80/max	4,963	3,462						
United States								
18/29		269	15,509	36,014	27,610	14,140	12,825	3,211
30/39	366	10,497	22,926	40,908	21,534	14,896	6,934	
40/49	11,567	14,378	22,789	21,942	13,102	4,601		
50/59	22,230	13,738	10,912	12,772	4,477			
60/69	27,206	6,540	6,483	3,878				
70/79	19,253	2,924	1,413					
80/max	9,157	637						
Russia								
18/29					1,864	14,424	31,611	14,526
30/39				2,304	11,775	26,064	13,839	
40/49			1,831	13,137	21,304	12,164		
50/59		2,065	8,065	22,676	10,682			
60/69	1,537	9,306	13,367	10,046				
70/79	5,800	12,393	5,225					
80/max	5,924	2,985						
Switzerland								
18/29						4,398	10,389	8,210
30/39					8,222	9,445	4,453	
40/49				7,638	16,440	6,411		
50/59			5,928	13,888	8,811			
60/69		3,885	11,182	7,283				
70/79	2,176	6,525	5,796					
80/max	3,138	2,547						
Germany								
18/29				3,367	27,133	31,543	39,541	29,239
30/39			3,344	21,400	38,970	43,479	23,469	
40/49		3,563	19,398	33,950	54,142	27,963		
50/59	2,376	17,570	27,350	40,066	24,461			
60/69	11,823	24,385	33,642	16,600				
70/79	18,807	23,840	13,242					
80/max	15,545	5,485						

(continued)

Table 3. (Continued)

Age group	Birth cohort							
	1920s and earlier	30s	40s	50s	60s	70s	80s	90s and later
United Kingdom								
18/29					7,187	25,143	50,972	34,755
30/39				6,689	27,367	54,372	22,801	
40/49			6,595	22,559	61,806	27,506		
50/59		4,157	20,823	50,456	27,749			
60/69	3,928	14,453	46,103	22,066				
70/79	14,884	28,190	17,335					
80/max	20,544	8,347						

Table 4. Basic characteristics of the sample (column percentages)

	Australia	Korea	United States	Russia	Switzerland	Germany	United Kingdom	Total
Gender								
Male	47.2	47.9	44.7	42.1	44.6	47.6	45.9	46.0
Female	52.8	52.1	55.3	57.9	55.4	52.4	54.1	54.0
N	257,418	257,495	457,637	274,916	146,765	675,693	626,166	2,696,090
Education: 3 levels								
(0–2) Low	29.0	33.3	25.7	19.3	9.8	19.6	29.4	24.5
(3–4) Medium	38.1	48.5	56.1	53.3	57.3	57.0	37.3	49.3
(5–8) High	32.9	18.2	18.2	27.4	33.0	23.5	33.3	26.1
N	257,277	257,452	452,265	274,331	146,763	663,515	603,507	2,655,110
Formal marital status								
Married/registered	50.4	66.3	70.3	54.4	58.5	61.4	57.5	60.6
Never married	35.2	21.2	12.1	24.4	25.6	23.2	27.9	23.6
Widowed	5.2	8.7	5.7	12.5	5.5	6.0	6.8	7.0
Divorced	6.3	3.0	7.9	8.3	9.0	7.0	6.0	6.7
Separated	2.9	0.7	4.0	0.4	1.5	2.4	1.8	2.2
N	257,374	257,441	457,607	274,454	146,761	669,286	626,412	2,689,335
Employment status								
Employed	64.4	57.6	67.7	57.9	68.8	58.8	57.7	60.9
Unemployed	3.5	2.9	5.0	7.4	2.0	7.4	4.6	5.2
Retired, disabled	21.8	11.2	13.9	25.2	16.6	20.9	26.6	20.5
Not active/home	8.9	23.3	12.1	6.4	8.6	9.0	7.0	10.1
In education	1.5	5.0	1.3	3.1	4.0	3.9	4.2	3.4
N	257,418	257,493	425,355	274,893	146,765	675,685	626,126	2,663,735
Occupation (ISCO level 1) of employed								
[1] Managers	12.1	1.5	11.1	6.7	9.2	5.8	14.1	9.2
[2] Professionals	20.8	11.1	14.6	17.6	21.3	16.7	14.3	16.1
[3] Technicians	16.1	8.9	15.0	17.3	25.4	21.8	14.4	16.9
[4] Clerical support	12.5	14.9	10.7	5.3	11.9	11.1	13.8	11.6
[5] Services and sale	12.9	21.0	13.5	16.9	12.4	11.3	17.8	14.8
[6] Skilled agricult.	2.9	7.3	0.5	0.4	3.4	1.3	1.1	1.9
[7] Craft and related	9.3	11.8	10.6	13.8	9.2	15.9	8.5	11.6
[8] Plant and machine	6.0	12.3	9.2	14.6	2.5	8.1	6.8	8.5
[9] Elementary occup.	7.4	10.9	13.8	7.4	4.6	7.6	9.1	9.2
N	165,720	146,852	301,351	157,866	99,086	377,193	354,208	1,602,276

Note: Missing values were removed. 'Armed forces' not shown in occupations due to low frequency.

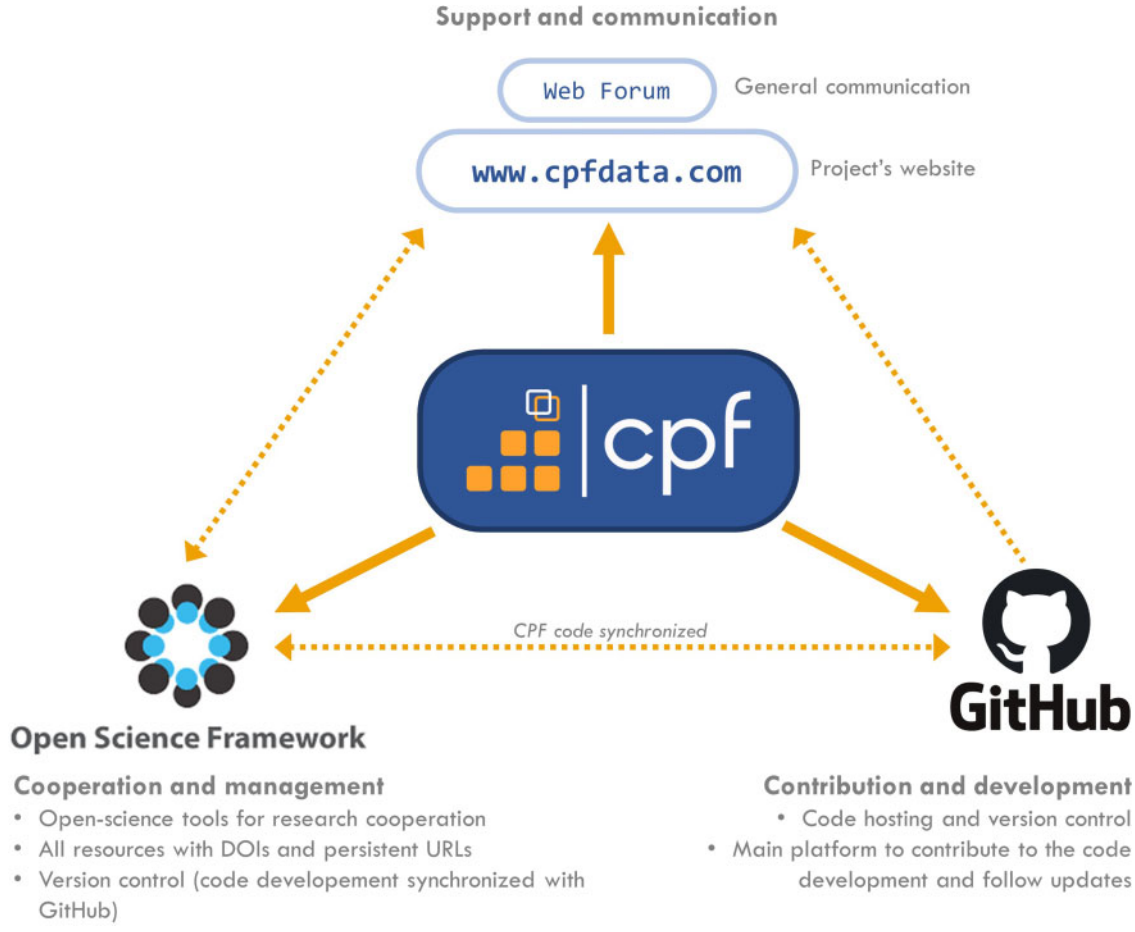

Figure 6. The structure and tools of the CPF's Open Science Framework

facilitate collaborative workflow on projects andallow to document the progress. Similarly to GitHub, OSF uses a version control system to record all changes to the project. OSF allows additionally to register the project at each stage and creates an archival version of the project with a unique hyperlink. All materials can receive permanent links and DOIs. Importantly, OSF includes a GitHub add-on which directly links files stored at GitHub repository into the OSF project. In this way, changes to the code that are introduced either through GitHub or OSF are synchronized and the code is always up-to-date.

Concluding Remarks

In this data brief, we proposed the CPF as a tool for harmonizing comparative life course data. The CPF was developed to meet the expectations of social researchers who increasingly signal their interest in cross-national panel studies (Burkhauser and Lillard, 2005; Dubrow and Tomescu-Dubrow, 2015; Kühne *et al.*, 2020). A harmonized dataset of seven long-running household panel studies offers diverse and ample research options, particularly in studying changes in the life courses of consecutive cohorts. The CPF was inspired by the CNEF, but it has been built in a different way with the paramount goal of supporting researchers' flexibility and allowing for improvements in the open-source code.

Like other harmonization projects, the CPF is limited by the differences in questionnaires and survey designs of studies included (Kaminska and Lynn, 2017; Slomczynski and Tomescu-Dubrow, 2019). For many applications, a full comparative equivalence is required but not always available (Wolf *et al.*, 2017). Surveys also apply different designs and samples or show differences in response and attrition rates. Addressing these

problems may require developing harmonized survey weights (Kish, 1999; Zieliński, Powałko and Kołczyńska, 2019). Against the backdrop of these challenges, the contribution of the CPF is that it enables modifications of the code so that users have full control over the analysis. They can adjust or improve the algorithms and add new variables.

The development of the CPF will continue. Contrary to other international governmental-founded harmonization projects, the CPF is a bottom-up and ongoing initiative that can facilitate various forms of cooperation within a community of researchers. Plans for the nearest future include improving existing variables and the code, adding new variables, and developing the online environment. Further plans involve integrating additional surveys, such as the Longitudinal Internet studies for the Social Sciences (LISS) from the Netherlands, or the Japan Household Panel Survey (JHPS/KHPS). Therefore, we invite interested users to provide feedback or contribute to the development of the project. With unrestricted access to the harmonization code and other online resources, the CPF will support the open science community of social researchers.

Citing the CPF Database

If you publish with the CPF, please include the following note:

This paper uses code from the Comparative Panel File (CPF) version 1.1 available at www.cpfdata.com. CPF is created by Konrad Turek, Matthijs Kalmijn, and Thomas Leopold. The initial version of CPF has been developed in the CRITEVENTS project (PI: Thomas Leopold) and funded by an ERA-NET Cofund grant within the NORFACE Joint Research Programme on the Dynamics of Inequality Across the Life-course (DIAL). doi: 10.17605/OSF.IO/H3YXQ. For details, see: Turek, K., Kalmijn, M., and Leopold, T. (2021). The Comparative Panel File (CPF): Harmonised Household Panel Surveys from Seven Countries. European Sociological Review.

Supplementary Data

Supplementary data are available at *ESR* online.

Notes

1 Other important harmonization projects are *The European Union—Survey on Income and Living Conditions* (EU-SILC) and the *Luxembourg Income Study* (LIS). There are also discontinued studies, such as the *European Community Household Panel* (Burkhauser and Lillard, 2005; Dubrow and Tomescu-Dubrow, 2015).

2 CNEF is being reorganized since late 2020 and aims at providing more transparency in data processing, a modernized webpage and easier access to the data.

3 CNEF also includes the Japan Household Panel Survey. This survey is not yet harmonized for the CPF but will be added in the future.

4 This article forms part of the CRITEVENTS project. The CRITEVENTS project is financially supported by the NORFACE Joint Research Programme on the Dynamics of Inequality Across the Life-course, which is co-funded by the European Commission through Horizon 2020 under grant agreement No. 724363.

5 The CNEF involved a cooperation with national source data administrators, an international group of researchers from United States, Germany, United Kingdom, Switzerland, Australia, Korea, Russia, and Canada. CNEF was funded by several institutions, including the US National Institute on Aging, the German Institute for Economic Research, and Cornell University.

6 The CNEF was built on the model implemented in the Luxembourg Income Study (LIS), which harmonizes micro-level household surveys data from over 25 countries (Burkhauser *et al.*, 2001; Frick *et al.*, 2007). LIS was limited by the cross-sectional character of the data, and difficulties in accessing the data due to confidentiality issues. CNEF aimed at harmonizing more accessible panel data and including a broader range of research topics than LIS.

7 The Canadian Survey of Labour and Income Dynamics (SLID) is discontinued.

8 According to the Statistics Korea, the share of managers among employed persons is ca. 1.5 per cent (www.kosis.kr/eng, online table 'Employed persons by gender/occupation').

Acknowledgements

The authors thank the colleagues who supported development of the first version of CPF, in particular (in alphabetical order) Eldad Davidov, Dina Maskileyson, Aleja Rodríguez, Katya Sytkina, Gert Thielemans, and Gordey Yastrebov. This study uses the following datasets: The British Household Panel Survey, BHPS, and Understanding Society—The UK Household Longitudinal Study, UKHLS. University of Essex, Institute for Social and Economic Research, NatCen Social Research, Kantar Public (2019). Understanding Society: Waves 1–9, 2009–2018 and Harmonised BHPS: Waves 1–18, 1991–2009. 12th Edition. UK Data Service. SN: 6614, http://doi.org/10.

5255/UKDA-SN-6614-14. Socio-Economic Panel (SOEP), data for years 1984–2018, version 35, SOEP, 2020, doi: 10.5684/soep.v35. https://www.diw.de/en/soep. Panel Study of Income Dynamics, public use dataset. Produced and distributed by the Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI (2020). https://psidonline.isr.umich.edu. The Household, Income and Labour Dynamics in Australia (HILDA) Survey, GENERAL RELEASE 18 (Waves 1–18), Department of Social Services; Melbourne Institute of Applied Economic and Social Research, 2019, doi: 10.26193/IYBXHM, ADA Dataverse, V5. Korean Labor & Income Panel Study (KLIPS) version 21. Copyright Korea Labor Institute, 2020. www.kli.re.kr/klips_eng. Russia Longitudinal Monitoring Survey, RLMS-HSE, version 2018, conducted by National Research University 'Higher School of Economics' and ZAO 'Demoscope' together with Carolina Population Center, University of North Carolina at Chapel Hill and the Institute of Sociology RAS. (RLMS-HSE sites: http://www.cpc.unc.edu/projects/rlms-hse, http://www.hse.ru/org/hse/rlms). Swiss Household Panel (SHP), version 20, SHP is based at the Swiss Centre of Expertise in the Social Sciences FORS. The project is supported by the Swiss National Science Foundation. https://forscenter.ch/projects/swiss-household-panel. The Cross-National Equivalent File project is sponsored by the National Institute on Aging (Grant: 5-R01AG040213-10) and the Eunice Kennedy Shriver National Institute of Child Health and Human Development (Grants: 1-R03HD091871-01 and 1-R03HD100924-01) and was conducted by The Ohio State University. www.cnef.ehe.osu.edu.

References

Aassve, A. *et al.* (2002). Leaving home: a comparative analysis of ECHP data. *Journal of European Social Policy*, **12**, 259–276.

Allanson, P. F. (2011). On the characterisation and economic evaluation of income mobility as a process of distributional change. *The Journal of Economic Inequality*, **10**, 505–528.

Andress, H. J. *et al.* (2006). The economic consequences of partnership dissolution - a comparative analysis of panel studies from Belgium, Germany, Great Britain, Italy, and Sweden. *European Sociological Review*, **22**, 533–560.

Bernardi, L., Huinink, J. and Settersten, R. A. (2019). The life course cube: a tool for studying lives. *Advances in Life Course Research*, **41**, 41.

Blossfeld, H.-P. and Hakim, C. (1997). *Between Equalisation and Marginalisation: Women Working Part-Time in Europe and the United States of America*. Oxford: Oxford University Press.

Blossfeld, H.-P. *et al.* (2005). *Globalisation, Uncertainty, and Youth in Society*. London: Routledge.

Boye, K. (2011). Work and wellbeing in a comparative perspective – the role of family policy. *European Sociological Review*, **27**, 16–30.

Brüderl, J., Kratz, F. and Bauer, G. (2019). Life course research with panel data: an analysis of the reproduction of social inequality. *Advances in Life Course Research*, **41**, 41.

Bryan, M. L. and Jenkins, S. P. (2016). Multilevel modelling of country effects: a cautionary tale. *European Sociological Review*, **32**, 3–22.

Büchel, F. and Frick, J. R. (2004). Immigrants in the UK and in West Germany? Relative income position, income portfolio, and redistribution effects. *Population Economics*, **17**, 553–581.

Buck, N. and McFall, S. (2012). Understanding society: design overview. *Longitudinal and Life Course Studies*, **3**, 5–17.

Burkhauser, R. V. *et al.* (2001). The Cross-National Equivalent File: a product of cross-national research. In Becker I., Ott N., Rolf G. (Eds.), *Social Insurance in a Dynamic Society*. Frankfurt: Campus Fachbuch.

Burkhauser, R. V. and Lillard, D. R. (2005). The contribution and potential of data harmonisation for cross-national comparative research. *Journal of Comparative Policy Analysis: Research and Practice*, **7**, 313–330.

Chen, W.-H. (2009). Cross-national differences in income mobility: evidence from Canada, the United States, Great Britain and Germany. *Review of Income and Wealth*, **55**, 75–100.

Cho, J. and Lee, A. (2013). Life satisfaction of the aged in the retirement process: a comparative study of South Korea with Germany and Switzerland. *Applied Research in Quality of Life*, **9**, 179–195.

Cooke, T. J. *et al.* (2009). A longitudinal analysis of family migration and the gender gap in earnings in the United States and Great Britain. *Demography*, **46**, 1, 147–167.

DiPrete, T. A. and McManus, P. (1996). Institutions, technical change, and diverging life chances: earnings mobility in the United States and Germany. *American Journal of Sociology*, **102**, 34–79.

Dubrow, J. K. and Tomescu-Dubrow, I. (2015). The rise of cross-national survey data harmonisation in the social sciences: emergence of an interdisciplinary methodological field. *Quality & Quantity*, **50**, 4, 1449–1467.

Ehlert, M. (2013). Job loss among rich and poor in the United States and Germany: who loses more income? *Research in Social Stratification and Mobility*, **32**, 85–103.

Frick, J. R. *et al.* (2007). The Cross-National Equivalent File (CNEF) and its member country household panel studies. *Schmollers Jahrbuch: Zeitschrift Für Wirtschafts- Und Sozialwissenschaften*, **127**, 627–654.

Gelman, A. and Hill, J. (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Gerry, C. J. and Papadopoulos, G. (2015). Sample attrition in the RLMS, 2001. *Economics of Transition*, **23**, 2, 425–468.

Giesselmann, M. *et al.* (2019). The individual in context(s): research potentials of the socio-economic panel study (SOEP) in sociology. *European Sociological Review*, **35**, 738–755.

Goebel, J. *et al.* (2019). The German Socio-Economic Panel (SOEP). *Journal of Economics and Statistics*, **239**, 345–360.

Johnson, D. *et al.* (2018). Fifty years of the panel study of income dynamics: past, present, and future. *The Annals of the American Academy of Political and Social Science*, **680**, 9–28.

Kaminska, O. and Lynn, P. (2017). Survey-based cross-country comparisons where countries vary in sample design: issues and solutions. *Journal of Official Statistics*, **33**, 123–136.

Kish, L. (1999). Cumulating/combining population surveys. *Survey Methodology*, **25**, 2, 129–138.

KLI. (2020). *Korean Labor and Income Panel Study (KLIPS) Waves* 1–21. User's Guide. Korea: Korea Labor Institute.

Kozyreva, P. and Sabirianova Peter, K. (2015). Economic change in Russia: twenty years of the Russian Longitudinal Monitoring Survey. *Economics of Transition*, **23**, 293–298.

Kühne, S. *et al.* (2020). The need for household panel surveys in times of crisis: the case of SOEP-CoV. *Survey Research Methods*, **14**, 195–203

Leopold, L. (2018). Education and physical health trajectories in later life: a comparative study. *Demography*, **55**, 901–927.

Liefbroer, A. C. and Dourleijn, E. (2006). Unmarried cohabitation and union stability: testing the role of diffusion using data from 16 European Countries. *Demography*, **43**, 203–221.

Mayer, K. U. (2009). New directions in life course research. *Annual Review of Sociology*, **35**, 413–433.

McCall, L. and Percheski, C. (2010). Income inequality: new trends and research directions. *Annual Review of Sociology*, **36**, 329–347.

McGonagle, K. A. *et al.* (2012). The panel study of income dynamics: overview, recent innovations, and potential for life course research. *Longitudinal and Life Course Studies*, **3**, 2, 268– 284.

McManus, P. A. (2003). Parents, partners, and credentials: self-employment mobility in the United States and Germany. *In Inequality across Societies: Families, Schools and Persisting Stratification*, **171**. 200.

Mohring, K. (2016). Life course regimes in Europe: individual employment histories in comparative and historical perspective. *Journal of European Social Policy*, **26**, 2, 124–139.

Musick, K., Bea, M. D. and Gonalons-Pons, P. (2020). His and her earnings following parenthood in the United States, Germany, and the United Kingdom. *American Sociological Review*, **85**, 639–674.

Perelli-Harris, B. and Lyons-Amos, M. (2016). Partnership patterns in the United States and across Europe: the role of education and country context. *Social Forces*, **95**, 251–281.

Piccarreta, R. and Studer, M. (2019). Holistic analysis of the life course: methodological challenges and new perspectives. *Advances in Life Course Research*,

Platt, L. *et al.* (2020). Understanding society at 10 years. *European Sociological Review*,

Revilla, M. A., Saris, W. E. and Krosnick, J. A. (2014). Choosing the number of categories in agree-disagree scales. *Sociological Methods & Research*, **43**, 73–97.

Rose, D. (1995). Household panel studies: an overview. *Innovation: The European Journal of Social Science Research*, **8**, 7–24.

Siegers, R., Belcheva, V. and Silbermann, T. (2020). *SOEPcore v35—Documentation of Sample Sizes and Panel Attrition in the German Socio-Economic Panel (SOEP) (1984 until 2018)*: DIW/SOEP: SOEP Survey Papers, 826.

Slomczynski, K. M., and Tomescu-Dubrow, I. (2019). Basic principles of survey data recycling. In Johnson T. P., Pennell B.-E., Stoop I. A. L., Dorer B. (Eds.), *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*, Hoboken, NJ: John Wiley & Sons.

Tillmann, R. *et al.* (2016). The Swiss Household Panel Study: observing social change since 1999. *Longitudinal and Life Course Studies*, **7**, 64–78.

Voorpostel, M. *et al.* (2020). *Swiss Household Panel Userguide (1999-2018), Wave 20*. Lausanne: FORS.

Watson, N. and Wooden, M. (2020). The Household, Income and Labour Dynamics in Australia (HILDA) Survey. *Journal of Economics and Statistics, 241, 1, 131–141*.

Whelan, C. T., Layte, R. and Maitre, B. (2004). Understanding the mismatch between income poverty and deprivation: a dynamic comparative analysis. *European Sociological Review*, **20**, 287–302.

Wolf, C. *et al.* (2017). Harmonising survey questions between cultures and over time. In Wolf C., Y.-C Fu., Joye D., Smith T. (Eds.), *The SAGE Handbook of Survey Methodology*. London: SAGE Publications.

Zieliński, M. W., Powałko, P. and Kołczyńska, M. (2019). The past, present, and future of statistical weights in international survey projects: implications for survey data harmonization. In Johnson B. T., Pennell B.-E., Stoop I. A. L., Dorer B. (Eds.), *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC)*, Hoboken, NJ: John Wiley & Sons.

Konrad Turek is a sociologist, social researcher and data analyst working as a postdoctoral researcher at the Netherlands Interdisciplinary Demographic Institute (NIDI-KNAW/University of Groningen) in the Work & Retirement group. His primary research interests include changing and ageing labour markets, life course inequalities, lifelong learning, ageing policies, age management and retirement patterns. He received a Horizon 2020 Marie Sk?odowska-Curie Individual Fellowship to study the role of human capital investments for retirement transitions. He also co-coordinated the Human Capital Study the largest labour market research in Poland.

Matthijs Kalmijn is a full Professor of Sociology at the University of Groningen and senior researcher at the Netherlands Interdisciplinary Demographic Institute (NIDI-KNAW) in The Hague. His main research fields are family, life courses, and intergenerational relationships. He was also co-director of several large-scale surveys in the Netherlands, including the survey Parents and Children in the Netherlands (OKiN), the Netherlands Kinship Panel Study (NKPS), and the Netherlands Longitudinal Lifecourse Study (NELLS). Kalmijn has a PhD from UCLA (1991). He is principal

investigator of the ERC Advanced Grant project Intergenerational Reproduction and Solidarity in an Era of Family Complexity.

Thomas Leopold is a full Professor of Methods of Empirical Social Research at the Institute of Sociology and Social Psychology at the University of Cologne. He completed a PhD at the University of Bamberg and was a Max Weber Postdoctoral Fellow at the European University Institute. His interests are in social demography, family relations, and life course research. He is principal investigator of the ERC Starting Grant project KINMATRIX: Uncovering the Kinship Matrix: A New Study of Solidarity and Transmission in European Families. He is also principal investigator of the NORFACE project CRITEVENTS (20172020), funded by the European Commission via an ERA-NET Cofund grant.