Comparative Panel File (CPF) 2.0

Information and Manual

CPF 2.0 Manual v.1

Konrad Turek

Matthijs Kalmijn

In collaboration with:

Priscilla Zhang & Centerdata

Xiao Xu, NIDI

www.cpfdata.com

Comparative Panel File - Open Science Project

The Netherlands

July 2025

Suggested citation for the Manual

Turek, K., Kalmijn, M. (2025). *Comparative Panel File (CPF) 2.0: Information and Manual.* DOI: 10.31219/osf.io/y3ph6_v1

Citing the main article about the CPF

Turek, K., Kalmijn, M. and Leopold, T. (2021), The Comparative Panel File: Harmonized Household Panel Surveys from Seven Countries, *European Sociological Review*, Vol. 37(3): 505–523, https://doi.org/10.1093/esr/jcab006

Citing the CPF code and project

We kindly ask you to include the following information in publications that use the CPF code:

This paper uses the open code from the Comparative Panel File (CPF) version 2.0 available at www.cpfdata.com created by Konrad Turek, Matthijs Kalmijn, Thomas Leopold, and others. The project was supported by the "NORFACE Joint Research Programme on the Dynamics of Inequality Across the Life-course" and "NWO Open Science Fund" (grant OSF23.2.017). DOI:10.17605/OSF.IO/H3YXQ.

CPF v.2.0 team

Core team:
Konrad Turek, Tilburg University & Netherlands Interdisciplinary Demographic Institute
Matthijs Kalmijn, Netherlands Interdisciplinary Demographic Institute

Support for the Dutch LISS data:
Priscilla Zhang, Centerdata, and Centerdata team

Support for developing the CPF 2.0 platform and webpage:
Xiao Xu, NIDI

Research Assistant for CPF v.1.5:
Isabel Voets, Netherlands Interdisciplinary Demographic Institute

Abstract

The Comparative Panel File (CPF) harmonizes household panel data from seven of the world's most important longitudinal surveys: Australia (HILDA), Germany (SOEP), Great Britain (BHPS/UKHLS), South Korea (KLIPS), Switzerland (SHP), the United States (PSID), and the Netherlands (LISS). These studies offer rich, multi-decade information on individuals and households.

The project aims to support the social science community in analyzing comparative life course data. CPF provides open-source Stata code that links and harmonizes these data into a unified three-level panel structure, enabling comparative life course and social science research. The project is designed for flexibility—users can customize the code to suit their own variables, countries, and samples. It supports both straightforward use and advanced extensions. CPF is not a data product, but a community-driven harmonization tool. Users must download original data from national sources and apply the CPF code to generate the harmonized dataset.

In this manual, we present the design and content of the CPF, explain the logic of the project, workflow, and technical details. We also describe the CPF's open-science platform. The first version of CPF was prepared by Konrad Turek, Thomas Leopold and Matthijs Kalmijn, and published in December 2020.

Comparative Panel File web addresses:

- Website: cpfdata.com
- Forum: cpfdata.com/forum
- GitHub: github.com/cpfdata
- OSF: osf.io/h3yxq
- Contact: contact@cpfdata.com or k.l.turek@tilburguniversity.com

Contents

List of symbols and abbreviations

Symbols

Folder directory:	`D:/CPF/11_CPF_in_syntax/`
Syntax do-file:	`1_Folder_setup`
Data file:	**`CPF_v2.0.dta`**
Variable:	***`education`***
Syntax code:	`global your_dir "D:/CPF"  // <--inster your directory`

Abbreviations

CPF	Comparative Panel File
BHPS	British Household Panel Survey
HILDA	Household, Income and Labor Dynamics in Australia Survey
KLIPS	Korean Labor and Income Panel Study
PSID	Panel Study of Income Dynamics (the US)
SHP	Swiss Household Panel
SOEP	German Socio-Economic Panel
UKHLS	Understanding Society – The UK Household Longitudinal Study
LISS	Longitudinal Internet studies for the Social Sciences (Nehterlands)
AUS	Australia
GER	Germany
KOR	South Korea
SWT	Switzerland
UK	United Kingdom
US	United States
NL	The Netherlands

1. Information about the new version of CPF v.2.0

CPF version 2.0 was published in July 2025. The development of this version was supported by a grant from the NWO (The Dutch Research Council) Open Science Fund (grant ID OSF23.2.017).

CPF is an ongoing open-source project that aims to support the community of social researchers. Please note that users must take responsibility for the final harmonization and analytical decisions, while CPF provides flexible tools and a coding framework only. If you find an error, want to suggest an improvement, or propose an extension, please contact us at contact@cpfdata.com / k.l.turek@tilburguniversity.com or suggest the changes on GitHub https://github.com/cpfdata.

New in CPF v2.0:

1. The CPF version 2.0 encompasses up to 43 waves (between 1968 and 2024), combines data from seven countries, and comprises approximately 3 million observations from more than 400,000 respondents.
2. The **LISS panel** from the Netherlands was added, offering high-quality annual data across diverse life domains. This was done in collaboration with Priscilla Zhang from Centrdata (the authors of the LISS survey).
3. The **Russian part** of the project (RLMS) has been **excluded** from CPF v2.0 in response to Russia's 2022 invasion of Ukraine.
4. The code was adjusted to include **the latest available waves** (data collected up to 2023 and 2024)
5. The CPF code was improved with a more transparent syntax structure and some **modifications** to the harmonization code (based on the team evaluation and users' suggestions). The most visible changes include:
    - New labels for *Self-Rated Health* (*srh5*): (1) Excellent, (2) Very good, (3) Good, (4) Fair, (5) Poor (Instead of: 1 "Very good" 2 "Good" 3 "Satisfactory" 4 "Bad" 5 "Very bad")
    - More precise definition of yearly income and household income, depending on whether it refers to the previous year (based on respondents' reported value) or the current year (estimated based on monthly income). For example, instead of incjobs_y*, two new variables are created: incjobs_py* (reported annual income for the previous year) and incjobs_cy* (estimated annual income for the current year based on monthly income).
    - For *Gender* (variable *female*), a category 2 "Other/No answer" was added (available only in some countries)

- An updated beta version of 'psidtools' has been added to work with the new PSID code: At the moment of preparing the CPF 2.0 (6.2025), Psidtools required a minor update to work with the 2023 wave. The original Psidtools package was compatible only with PSID until wave 2021. The small adjustments made by K.Turek are based on the original code of the Psidtools creator, Prof. Dr. Ulrich Kohler (https://gitup.uni-potsdam.de/ukohler), who deserves all the credit for this amazing tool. The original source code is available at: https://gitup.uni-potsdam.de/ukohler/psidtools/-/blob/main/psid.ado. The CPF v.2.0 version is included in the CPF 2.0 syntax (in the PSID folder).
- Added equivalent household income
- Many minor corrections to the harmonization code (e.g., HH income in UK, kidsn* in KOR and US, relig in US, fedu/medu in GER, parstat6 excluded from US and KOR)
- Updated and improved instructions, comments, and additional information in the syntax files

Table 1. Data included in CPF 2.0

Country	Panel Survey	Start Year	Latest Wave	Data version in CPF 2.0
1. Australia	HILDA	2001	2020*	200c
2. Korea	KLIPS	1998	2023	26w
3. US	PSID	1968	2023	2023er
4. Russia	RLMS	*Excluded*		
5. Switzerland	SHP	1999	2023	Wave 25
6. Germany	SOEP	1984	2023	V40
7. UK	BHPS/UKHLS	1991	2022	10.5255/UKDA-SN-6614-20 (UKHLS waves 1-14; BHPS waves 1-18)
8. Netherlands	LISS	2007	2024	LISS releases available 07/2025

** At the moment of publication, we received no access to the latest waves of HILDA.*

CPF v.2.0 team

CPF 2.0 was designed and developed by:

Konrad Turek, Tilburg University & Netherlands Interdisciplinary Demographic Institute

Support for the Dutch LISS data:

Priscilla Zhang, Centerdata

Consultations & general supprt:

Matthijs Kalmijn, Netherlands Interdisciplinary Demographic Institute

Xiao Xu, Netherlands Interdisciplinary Demographic Institute

Russian RLMS data exclusion from CPF 2.0

Starting with CPF 2.0, the Russian Longitudinal Monitoring Survey (RLMS) data has been excluded from the CPF harmonized cross-country dataset.

This decision is a direct response to the Russian Federation's full-scale invasion of Ukraine on February 24, 2022—a brutal act of aggression that constitutes a grave violation of international law and human rights. In line with international academic principles and ethical responsibility, we have chosen not to include data from the Russian Federation in CPF's collaborative scientific infrastructure. This applies irrespective of where the data are used or by whom. The Comparative Panel File is an open science project grounded in the values of transparency, peaceful cooperation, and solidarity among nations. We stand firmly with Ukraine and with the global scientific community that condemns authoritarian violence, the targeting of civilians, and the suppression of democratic freedoms.

While RLMS data remain technically supported within the CPF codebase for legacy purposes, they are no longer part of the default configuration and will not be updated beyond CPF version 1.5 (based on data up to 2021). Users who wish to include RLMS data for valid academic purposes may do so by manually modifying the survey selection parameters in the configuration files.

The CPF 2.0 release continues to provide comprehensive harmonized panel data from Australia (HILDA), Korea (KLIPS), United States (PSID), Switzerland (SHP), Germany (SOEP), United Kingdom (UKHLS), and the Netherlands (LISS).

2. The idea of CPF

About CPF

Comparative Panel File (CPF) is an ongoing, open science project to harmonise the world's largest and longest-running household panel surveys from seven countries (Turek et al., 2021). The project aims to support the social science community in analyzing comparative life course data. By harmonising individual repeated data covering long periods and several general population surveys, researchers can analyse both time trends and country differences. Currently, CPF includes seven studies:

- **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),
- **Switzerland** (The Swiss Household Panel, SHP), and
- **the United States** (The Panel Study of Income Dynamics, PSID)
- **Netherlands** (Longitudinal Internet studies for the Social Sciences, LISS) ← New in v2.0

Previous versions of CPF include Russia (The Russian Longitudinal Monitoring Survey, RLMS), but from version 2.0 it is removed (see note above).

Rather than a static dataset, CPF offers flexible code (in Stata) that combines original survey data into a unified, three-level panel structure (observations nested within individuals, within countries). Thus, CPF is not a data product. Researchers must download the raw data from national data providers and apply the CPF code to create their harmonized dataset. The open-source nature of the code enables the development and expansion of its areas of application. The main features of CPF include:

- **Open source:** All code is fully transparent, editable, and extendable.
- **Broad and extendable scope:** Covers a broader range of variables (which can be further extended).
- **Flexible and modular:** Users can easily adapt it to include different countries, waves, or variables.
- **Community-driven:** Designed for collaborative development through open-coding frameworks GitHub.

The origins and development

CPF was initiated in 2019 in the context of the NORFACE-funded project "Critical Life Events and the Dynamics of Inequality" (CRITEVENTS)[1]. It was developed by Konrad Turek, Matthijs Kalmijn, and Thomas Leopold. The CPF code and entire open-science platform were designed and prepared by Konrad Turek, who continuously maintains it and coordinates the project (together with an open team of researchers and contributors, including ongoing collaboration with Matthijs Kalmijn and Thomas Leopold). In

[1] The CRITEVENTS project was 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.

particular, adding LISS in CPF 2.0 was done in collaboration with Priscilla Zhang from Centrdata (which conducts the LISS survey). Updates and developments for CPF 1.5 were prepared in large part by Isabel Voets (NIDI-KNAW) under the supervision of Konrad Turek and Matthijs Kalmijn. We would also like to thank Daniel van Wijk (NIDI-KNAW) for his precious comments and suggestions for improving and extending the code.

CPF originated as an attempt to move the data harmonization process to open science, crowdsource cooperation, and provide novel functionalities (Turek, 2025). The idea of harmonising such data is not new, and CPF builds on the foundation laid by CNEF (Cross-National Equivalent File).

Cross-National Equivalent File (CNEF)

The most well-known, long-running, and successful harmonisation project is the Cross-National Equivalent File (CNEF) (Burkhauser et al., 2001; Frick et al., 2007). CNEF 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.[1] Initially, in 1991, the dataset harmonised only a limited set of variables for two countries, the US and Germany.[1] Over the years, the project expanded by adding countries, such as the UK and Canada (discontinued) in 1999, Australia and Switzerland in 2007, and Russia 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 (Allanson, 2011; Büchel & Frick, 2004; Chen, 2009), sociology (DiPrete & McManus, 1996; Ehlert, 2013; McCall & Percheski, 2010; Musick et al., 2020), or demography (Cooke et al., 2009), and less often in research on other topics, such as life satisfaction (Cho & Lee, 2013), and self-employment (McManus, 2003).

However, CPF introduces some key innovations, including more flexibility, transparency, and breadth, making it a more accessible tool for comparative social science. CPF is open and flexible, thereby facilitating a genuine bottom-up approach. The open-source model allows users to inspect, modify, and extend the harmonisation process. CPF fully supports modifications to harmonized variables or the addition of new variables from the source database, depending on the researchers' needs. The code also facilitates work with single surveys, as it instructs how to go from a large set of raw files to an integrated and ready-for-analysis panel dataset (for some surveys, e.g., 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. Compared to CNEF, CPF also significantly broadens the scope by including variables on family, labour market status, wellbeing, social origin, and other factors.

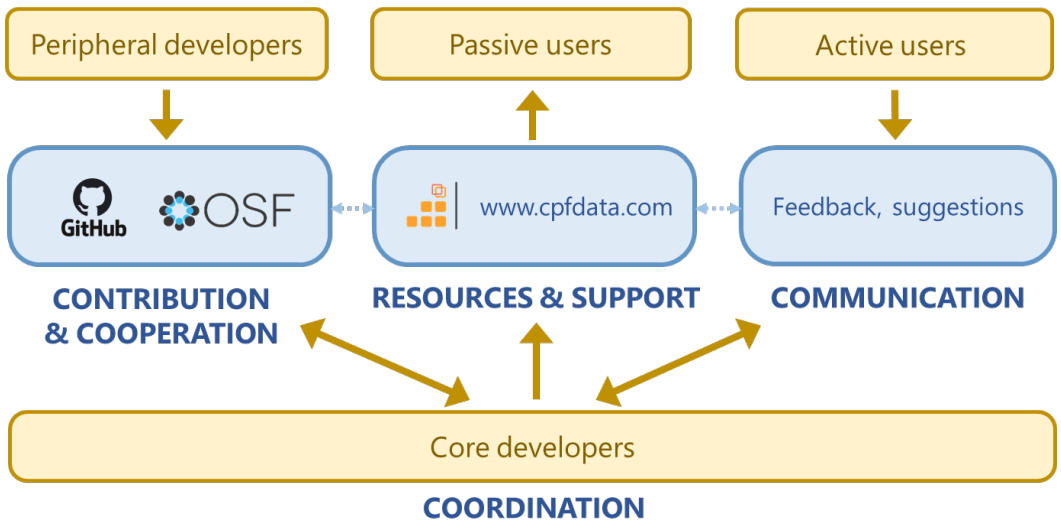
CPF is also more agile. It is not tied to centralized or government-managed releases, allowing it to incorporate new updates quickly. The process of accessing and harmonising data is streamlined: researchers need only secure access to national data sources and run the CPF code—no need for special applications or external file merging.

Main features of CPF

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. 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 a single dataset for all countries and all waves. Users can either follow the default workflow and run the code unchanged or modify and improve it for their specific use (e.g., selecting countries, adding new waves, or adding new or modifying existing variables). The file is organized in a long format, containing one record for each person in each wave. The merged file of CPF version 2.0 contains more than 3 million observations from over 400,000 individuals, observed for an average of 7.1 waves (ranging up to 43 waves).

The CPF is organized as an open science platform that integrates several tools to support open collaboration, management, documentation, and sharing of all materials. The main elements of the platform are the website (central platform) with forum (general communication), GitHub code repository (code development), and Open Science Framework (general management of scientific research). Users' improvements and suggestions will be recorded, incorporated, and shared using open online tools to allow continuous development and regular updates to the official versions of the code. This design – which can be called a coordinated open-source model (Turek, 2025) – balances community-based development with centralized coordination through a core team that supervises development and ensures quality control (Figure 1). Such a hybrid approach supports different types of users: external (peripheral) developers who can directly contribute code through GitHub, active users who provide substantial contributions, such as error detection and detailed suggestions, and passive users who benefit from the harmonized datasets. This combines the flexibility and innovation potential of crowd-based collaboration with the expertise control necessary for scientific applications, creating an ecosystem where users can contribute through microtask workflows. At the same time, the core team maintains overall coordination and quality assurance.

Figure 1. *CPF's coordinated open-source model*



More information about the CPF and the idea of open-source in social sciences can be found in publications:

- **Turek K. (2025)**. Accelerating Social Science Knowledge Production with the Coordinated Open-Source Model, *Quality & Quantity.*
- **Turek K., M. Kalmijn, T. Leopold (2021)** The Comparative Panel File (CPF): Harmonized Household Panel Surveys from Seven Countries. *European Sociological Review* , 37(3): 505–523.

3. CPF data sources – general information

For information about downloading and installing the data, see part 6 → "Installing the source data in CPF folders"

CPF harmonizes national household panel studies—long-running, representative surveys that track individuals and households over time. These surveys are conducted regularly (mostly on a yearly basis) and interview all or a select group of adult household members, collecting rich longitudinal data. CPF combines the seven most established and longest-running household panel studies globally (Table 2). CPF included Russian RLMS, which was replaced by the Dutch LISS panel in version 2.0.

Table 2. *Panel studies included in CPF*

Country	Survey name	Start Year	Webpage
Australia	Household, Income and Labour Dynamics in Australia (HILDA)	2001	melbourneinstitute.unimelb.edu.au/hilda
Korea	Korean Labor and Income Panel Study (KLIPS)	1998	kli.re.kr/klips_eng
US	Panel Study of Income Dynamics (PSID)	1968	psidonline.isr.umich.edu
Switzerland	Swiss Household Panel (SHP)	1999	forscenter.ch
Germany	German Socio-Economic Panel (SOEP)	1984	diw.de/en/soep
UK	British Household Panel Survey (BHPS) / UK Household Longitudinal Study (UKHLS)	1991 / 2009	understandingsociety.ac.uk
Netherlands	Longitudinal Internet Studies for the Social Sciences (LISS)	2007	lissdata.nl
Russia*	Russian Longitudinal Monitoring Survey (RLMS)	1994	www.hse.ru/en/rlms

*Note: * Starting with CPF v2.0, the Russian Longitudinal Monitoring Survey (RLMS) is no longer included*

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 can be identified using country-specific variables. Each panel differs slightly in sample design, oversampling, or questionnaire structure, but CPF aligns core information across them, making cross-country comparisons possible. For details on each survey—including sample design, data access, and citation—see the dedicated subsections below.

Australia: HILDA

The Australian data come from the Household, Income and Labor Dynamics (HILDA). It began in 2001 as a nationally representative longitudinal survey of Australian households. HILDA is developed to study family and labour market dynamics, economic and subjective wellbeing over the life-course (Watson & Wooden, 2020). All working-age members of the household (over 15 years old) are re-interviewed annually. In 2001, ca. 7,682 households and 13,969 individuals participated in the survey. In 2011, 2,153 households (5,477 individuals) were additionally selected, expanding the original sample size. Since then, HILDA has followed over 17,000 Australians each year.

On a less frequent basis, further information on wealth, health care utilisation, eating habits, cognitive functioning, and retirement is collected. HILDA does not conduct interviews with foreign residents in Australia, people living in outlying areas, members of non-Australian defence forces, and foreign diplomatic officers. The HILDA is directed by the Melbourne Institute of Applied Economic and Social Research at the University of Melbourne and funded by the Australian Government Department of Social Services (DSS).

More information:

- Official website: melbourneinstitute.unimelb.edu.au/hilda
- Data are available via the National Centre for Longitudinal Data Dataverse (Australian Government Department of Social Services): https://dataverse.ada.edu.au/dataverse/ncld

For citing, please follow instructions of HILDA, e.g. all publications must carry the following:

> "This paper uses unit record data from Household, Income and Labour Dynamics in Australia Survey [HILDA] conducted by the Australian Government Department of Social Services (DSS). The findings and views reported in this paper, however, are those of the author[s] and should not be attributed to the Australian Government, DSS, or any of DSS' contractors or partners. DOI: ####"

South Korea: KLIPS

The South Korean data come from the Korean Labor and Income Panel Study (KLIPS). It began in 1998 as a national longitudinal survey of households and individuals living in urban areas in South Korea. Interviews with all household members aged 15 and older are conducted annually. KLIPS monitors individuals' economic and labour activities, income and expenditures, education, and job training. An original sample was developed using a stratified clustering method to select districts. Out of 7 metropolitan cities and urban areas in 8 provinces, households were derived based on an equal probability technique. KLIPS set out to re-interview ca. 5,000 households and 13,000 individuals every year.

Over the waves, the sample expands by adding individuals who form family ties with original panel respondents. Such sample growth enables tracking demographic dynamics (e.g., marriages, births, divorces) of initial sample members. In 2009, an additional consolidated sample of 1,415 households was added to address the issues of household attrition and representativeness. Since 2009, KLIPS has contained two panels within one dataset. Following families over decades allows the KLIPS data to contribute to vital improvements in Korean employment policies. The project is carried out by the Korea Labor Institute and Center for Labor Statistics Research. The main funder of the study is the Ministry of Employment and Labor.

For more information, see:

- Official website: www.kli.re.kr/klips_eng
- Data are available via the official website for registered users: www.kli.re.kr/klips_eng

United States: PSID

The US data come from the Panel Study of Income Dynamics (PSID). It covers the period from 1968 and remains the oldest national panel survey worldwide. The study was initially created to evaluate poverty and economic well-being dynamics in the US. Currently, PSID aims to study the dynamics of income and poverty by conducting regular interviews with only one person per family.

From 1968 to 1997, the data were collected every year. Since 1998, interviews have been biennial. In 1968, ca. 5,000 families and 18,000 individuals participated in the survey. Over the 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 in total (Johnson et al., 2018; McGonagle et al., 2012). Respondents had entered the study in three ways. Demographic inflows (births, adoptions, and marriages) brought new members to the families. Formation of new independent households as a result of children splitting off from their parents' homes provided new unit measures for PSID. Additionally, post-1968 immigrant families extended the original sample in 1997/1999. Now, ca. 10000 families participate in the PSID.

To enrich data, PSID collects supplement studies. Children development (CDS) for 18 years old and younger, transition into adulthood (TAS) of those over 18, disability and use of time (DUST) for 60 and older are monitored. PSID is managed by faculty at the University of Michigan. The project's major funders are the National Science Foundation, National Institute on Aging and National Institute of Child Health and Human Development.

More information:

- Official website: simba.isr.umich.edu/
- Data are available via the official website for registered users: simba.isr.umich.edu/Zips/ZipMain.aspx

For citing, please follow the instructions of PSID, e.g.:

> Panel Study of Income Dynamics, public use dataset [restricted use data, if appropriate]. Produced and distributed by the Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI (year data were downloaded)

Switzerland: SHP

Swiss data come from the Swiss Household Panel (SHP) and cover the period from 1999. SHP was run as a longitudinal survey of private households based on a random representative sampling. The project aims to report the dynamics of living conditions change, income, quality of life, and population representation in Switzerland. All household members aged 14 years and older are asked to complete an individual questionnaire annually by telephone. In 1999, ca. 5,074 households and 12,931 individuals participated in the survey.

Over time, a high percentage of non-responses has accumulated in SHP. Thus, in 2004, 2,538 new households were added to the original sample to overcome the issue. In 2013, a further 4,093 households were added to the sample size, although different measures were implemented to return those who refused to participate.

The SHP has been developed and funded by the Swiss National Science Foundation. The project was carried out by the Swiss Federal Statistical Office and the University of Neuchâtel. Since 2008, SHP has been integrated into the Swiss Centre of Expertise in the Social Sciences (FORS) and hosted by the University of Lausanne.

For more information, see:

- Official website: https://forscenter.ch/projects/swiss-household-panel/
- Data are available via FORSbase for registered users: https://forscenter.ch/projects/swiss-household-panel/data/

For citing, please follow the instructions of SHP, e.g. all publications must carry the following:

> "This study has been realised using data collected by the Swiss Household Panel (SHP), which is based at the Swiss Centre of Expertise in the Social Sciences FORS. The project is supported by the Swiss National Science Foundation".

Germany: SOEP

The German data come from the German Socio-Economic Panel (SOEP). SOEP began in 1984 as a representative longitudinal study of private households in Germany for social, behavioural, and economic research. It is designed to measure disparities in resources across individuals over the life course by re-interviewing the same household members aged 17 and older on an annual basis. Initially, SOEP included only Western Germany, and since 1990, after reunification, it also covers eastern parts of the country, being the only database worldwide covering such a political unification (Giesselmann et al., 2019; Goebel et al., 2019; Siegers et al., 2020).

The first wave of SOEP consisted of two samples and ca. 6,000 households from the western states of Germany: (1) with a German household head and (2) with a migrant Greek, Italian, Spanish, Turkish, or Yugoslavian household head. In 1990, the panel data were expanded to include a representative sample from East Germany. The data was further advanced by adding immigrant (1994/95, 2013/2015, and 2020) and refugee (2016 and 2017) samples. Now, ca. 15,000 households and 30,000 individuals participate in the SOEP.

SOEP has been developed by the Research Center 'Sonderforschungsbereich' at the Universities of Mannheim and Frankfurt/Main together with the German Institute for Economic Research (DIW Berlin). Since 1990, SOEP has been fully delegated to DIW under the umbrella of the Leibniz Association with financing from the state governments and the Federal Ministry of Education and Research (BMBF).

More information:

- Official website: www.diw.de/en/soep
- SOEPcompanion: companion.soep.de
- Additional resources (including variables-search system): paneldata.org
- Data are available via the Research Data Center SOEP after granting access: www.diw.de/en/diw_02.c.242211.en/criteria_fdz_soep.html

For citing, please follow instructions of SOEP, e.g.:

> Socio-Economic Panel (SOEP), version 40, data for years 1984-2023 (SOEP-Core v40, EU Edition). 2025. DOI:10.5684/soep.core.v40eu

United Kingdom: BHPS and UKHLS

The UK's CPF sample consists of two studies:

- British Household Panel Survey (BHPS): 1991–2008, 18 waves
- UK Household Longitudinal Study (UKHLS / Understanding Society): 2009–present

BHPS began in 1991 as a multi-purpose panel survey for social and economic research (Buck & McFall, 2012; Platt et al., 2020). It was conducted every year, with the same individuals being re-interviewed each successive year. The first wave of BHPS consisted of ca. 5,500 households and 10,300 individuals drawn from 250 areas of Great Britain. In the following years, additional samples were added, including 1,500 households in Scotland and 1,500 households in Wales (1999), and 2,000 households in Northern Ireland (2001).

Since 2009, BHPS has been integrated into UKHLS. With a target sample size of 40,000 households in the first wave, UKHLS became the largest nationally representative household panel study in the world. Young people aged 10-15 complete a youth questionnaire. Respondents aged 16 and over complete the adult survey and continue to be interviewed when they leave their original households. Data continue to be collected every year, yet the fieldwork is extended to around 2,5 years for each wave (e.g., the 1st wave of UKHLS covers years 2009-2011, the 2nd covers years 2010-2012). Both BHPS and UKHLS have been developed and carried out by the Institute for Social and Economic Research at the University of Essex.

More information:

- BHPS: www.iser.essex.ac.uk/bhps,
- UKLHS: www.understandingsociety.ac.uk
- Data are available via the UK Data Service after granting access: www.ukdataservice.ac.uk.

For citing, please follow instructions of UKHLS, e.g.:

> University of Essex, Institute for Social and Economic Research. (2025). Understanding Society: Waves 1-14, 2009-2021 and Harmonised BHPS: Waves 1-18, 1991-2009. [data collection]. 17th Edition. UK Data Service. SN: 6614, DOI: http://doi.org/10.5255/UKDA-SN-6614-20

Netherlands: LISS

The LISS panel (Longitudinal Internet Studies for the Social Sciences) is a nationally representative household panel from the Netherlands. It was established in 2007 and is managed by Centerdata, an independent non-profit institute affiliated with Tilburg University.

LISS stands out for its methodological rigor: it is based on a true probability sample, drawn from the Dutch population register by Statistics Netherlands (CBS). To ensure full population coverage—including households without Internet access—participants without digital resources are provided with a computer and broadband connection. Self-selection is not possible; panel participation is by invitation only.

The panel is based on a true probability sample of households, drawn from the population register by Statistics Netherlands. It consists of 5,000 households, comprising approximately 7,500 individuals of 16 years and older. One designated respondent per household provides household-level data (e.g. income, composition), while individuals complete monthly online surveys and receive monetary incentives for participation.

CPF includes the **LISS Core Study,** which provides repeated measures on 10 separate domains:

- Background variables (general information about the household)
- Health
- Politics & Values
- Religion & Ethnicity
- Social Integration & Leisure
- Family & Household
- Work & Schooling
- Personality
- Economic Situation: Income
- Economic Situation: Assets
- Economic Situation: Housing

Each LISS Core questionnaire starts at a different month of the year. Consequently, the data collection period for each LISS wave is spread over many months. Additionally, the data collection periods for specific modules have been changing over the years. Some modules (e.g. Health, Politics) were collected over two separate calendar years (and the period has been changing), while some modules were also skipped in particular years.

In addition to the LISS Core Study, LISS includes a set of **Background Variables** that are updated monthly. These cover general household and respondent characteristics such as household composition, income, and living arrangements. One household member reports this information on behalf of the household. These background variables ensure continuous tracking of key structural indicators and are available for integration with other questionnaire modules.

All collected data are made available for your own research through the LISS data archive. Today, LISS is widely used by scholars across disciplines—social sciences, economics, public health, and more. Over 8,000 researchers have registered for the LISS Data Archive, and the panel has been used in hundreds of peer-reviewed studies and policy evaluations. In addition to the Core Study, LISS allows collecting additional data on the same sample using on-request (paid) modules.

The LISS sample was originally drawn in 2007 from 10,000 households, with a recruitment rate of 48% yielding ~8,000 individuals from nearly 5,000 households. Multiple refreshment samples have been added to preserve representativeness and account for attrition. All refreshment samples were drawn from the population register by Statistics Netherlands. LISS maintains a high participation rate. In 2020–2022 overall response ranged from 71% to 75%. In 2019 average monthly individual response was around 80% and household response around ~84%.

More info:

- Official website: www.lissdata.nl
- Data access: LISS Data Archive after registration (https://www.lissdata.nl/use-the-data)

Citing as suggested by LISS-Centerdata:

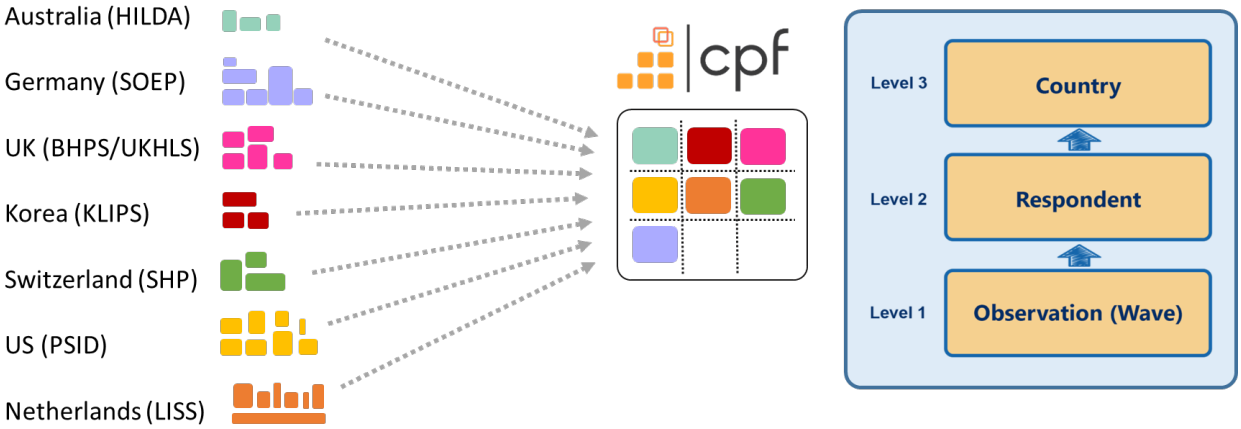
- `In this paper, we make use of data from the LISS panel (Longitudinal Internet studies for the Social Sciences), which is administered and managed by the non-profit research institute Centerdata (Tilburg University, the Netherlands).`
- `The LISS panel data were collected by the non-profit research institute Centerdata (Tilburg University, the Netherlands). Funding for the panel's ongoing operations comes from the Domain Plan SSH and ODISSEI since 2019. The initial set-up of the LISS panel in 2007 was funded through the MESS project by the Netherlands Organization for Scientific Research (NWO).`

4. CPF harmonization code and the outcome dataset: basic information

Data structure and time frame

CPF is a comparative panel dataset with a hierarchical structure of data. The structure consists of three levels (Figure 2): repeated individual observations from multiple waves (Level 1) are clustered within individuals (Level 2), and individuals are clustered within countries (Level 3).

Figure 2. *CPF data structure*



The CPF version 2.0 encompasses up to 43 waves (between 1968 and 2024), combines data from seven countries, and comprises approximately 3 million observations from more than 400,000 respondents (Table 3). The oldest survey is the PSID, which began in 1968 and has collected 43 waves. The second oldest is SOEP, which started in 1984 and has collected 40 waves to date. The youngest panel study in CPF is HILDA, with 23 waves since 2001 (however, only 20 waves are included in CPF 2.0, the rest will be added after securing data access), and the newly added LISS with 17 waves.

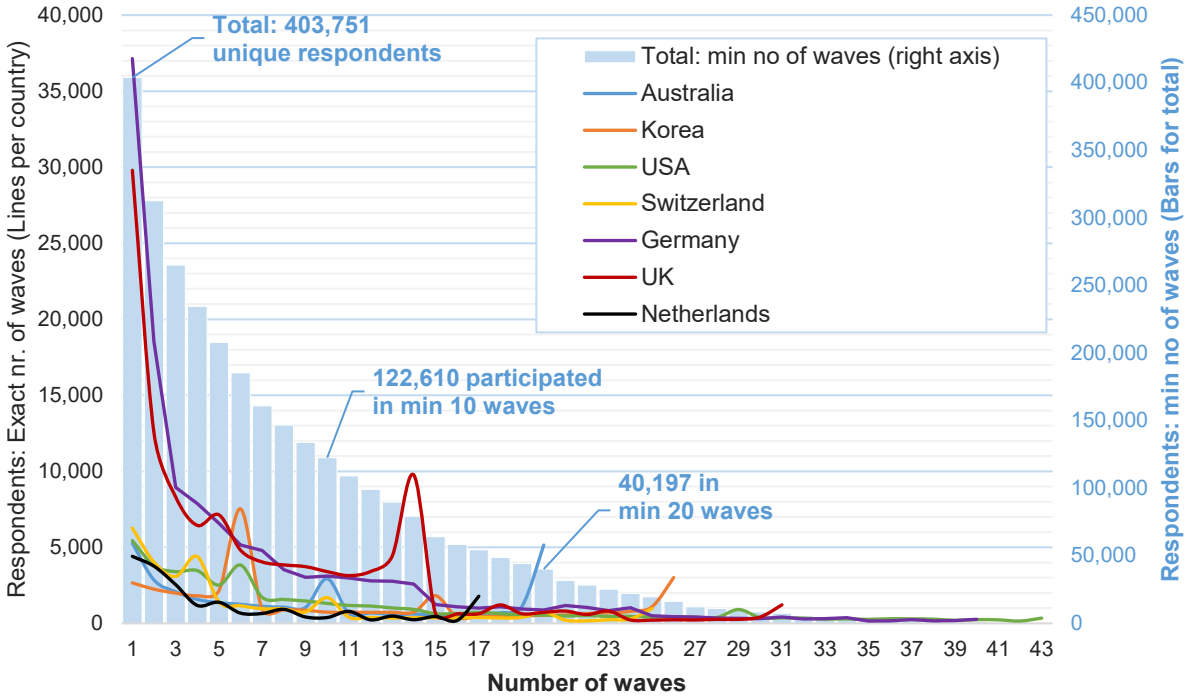
Table 3. *Number of waves, observations, and respondents (CPF v2.0)*

		First	No of	Observations		Unique respondents	
Country	**Survey**	**wave**	**waves**	n	%	n	%
[1] Australia	HILDA	2001	20	290,616	9.42	31,897	7.9
[2] Korea	KLIPS	1998	26	378,766	12.28	35,744	8.85
[3] USA	PSID	1968	43	499,658	16.2	45,077	11.16
[5] Switzerland	SHP	1999	25	203,914	6.61	30,719	7.61
[6] Germany	SOEP	1984	40	812,135	26.33	125,229	31.02
[7] UK	BHPS/UKHLS*	1991	32	782,886	25.38	114,554	28.37
[8] Netherlands	LISS	2008	17	116,355	3.77	20,531	5.09
Total				3,084,330	100	403,751	100

** BHPS: 1991-2008, 18 waves, UKHLS: from 2009, 14 waves*

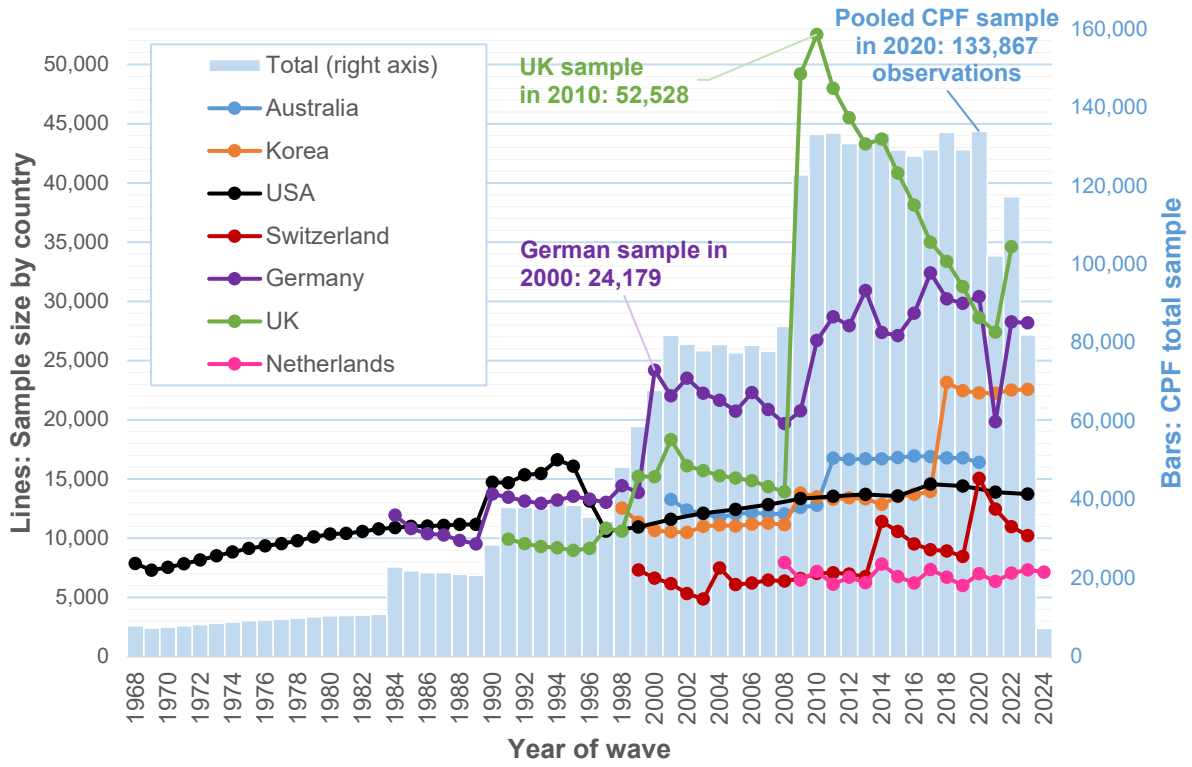
An average respondent participated in 7.6 waves (between 5.7 in the Netherlands and 11.1 in the US). As shown in Figure 3, out of all 403,751 respondents, 122,610 participated in a minimum of 10 waves and 40,197 in a minimum of 20 waves. Country-specific lines indicate the exact number of waves for which the dataset provides information on sample members. For example, there are almost 30,000 respondents in the UK who participated in only one wave, and around 3,400 who participated in exactly ten waves. In Australia, 16.2% of the sample (ca. 5000 respondents) participated in all 20 waves of HILDA.

Figure 3. *Number of waves in which individuals participated: exact number by survey (left axis) and minimum number for the total sample (right axis) (CPF v2.0).*



The oldest survey in CPF is PSID, covering the period from 1968 to the present, with 43 waves (Figure 4). The youngest panel study in CPF is the Netherlands, with 17 waves since 2007/2008, and HILDA with 20 waves since 2001. Since the wave of 2000, the number of participants in SOEP has grown significantly. CPF includes three countries since 1991, four countries since 1998, six countries since 2001, and all seven from 2008 (after excluding Russian data, which were available from 1994). A substantial increase observed for the UK sample in 2009 is related to the transition from the BHPS to the UKHLS. For most surveys, data have been collected yearly (after 1997, the PSID switched to 2-year intervals in data collection).

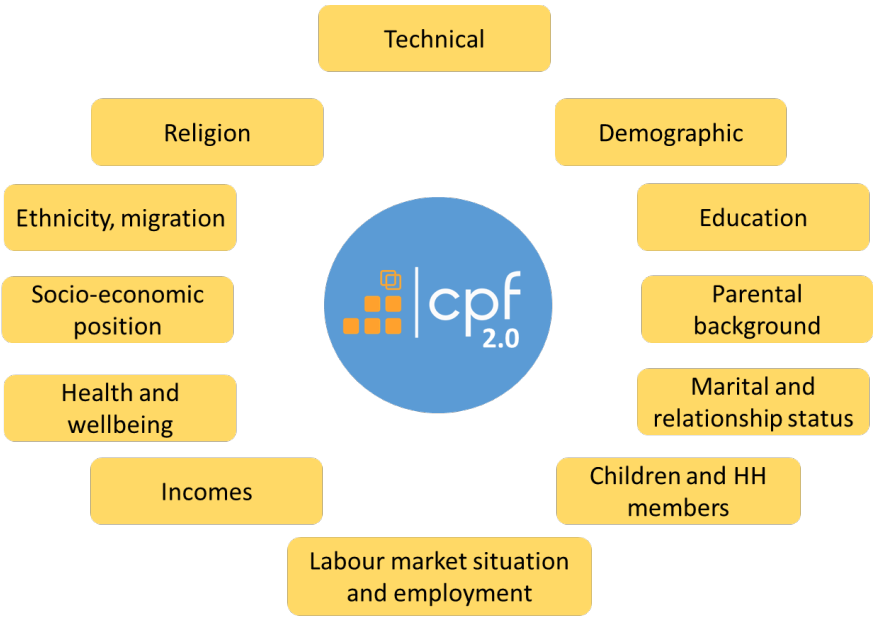
Figure 4. *Timeline of the data and sample sizes by wave (CPF v2.0)*



Variables and harmonization approach

The goal of CPF is to harmonise variables across surveys. CPF initially followed the general logic of the Cross-National Equivalent File (CNEF), but it extended this framework significantly in scope and flexibility. For example, instead of a simple employed/unemployed indicator, CPF provides a full range of labour market statuses (e.g., unemployed, retired, in education, inactive) and uses ISCED-based educational levels rather than just years of schooling. It also offers more detailed information on marital and partnership status, and includes additional variables such as training participation, satisfaction across various life domains, social origin, labour market experience, self-employment and entrepreneurship, work-education skill match, and perceived job security. New variables have been added in version 1.5, including religiosity, ethnicity, and parental background. An overview of all variables is presented in Figure 5 and further detailed in Table 3. Users should consult the CPF Codebook for survey-specific notes, comparability limits, and detailed variable definitions (www.cpfdata.com/download).

Figure 5. *The main groups of variable harmonized in CPF 2.0*



For harmonisation, we explored the items available in the source data for their comparative potential. Some questions had a very similar form across all surveys, 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 note that variations in question wording may result in (likely minor) differences in the frequency distributions. However, correlations with other variables are not necessarily affected (Kaminska & Lynn, 2017; Slomczynski & Tomescu-Dubrow, 2019; Wolf et al., 2017).

Full harmonisation was not always possible, so some variables 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).

We recommend that users read the *Codebook* before using the harmonized CPF variables. If necessary, changes to the harmonization approach can be easily applied at the syntax level (mostly in country-level syntax 02, the procedure is described below in *Workflow D*).

Table 3. *Variables available in the CPF v.2.0*

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 harmonised 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, [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
	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 harmonised 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 organisation.	- Industry: 3 major, 10 sub-major and 17 minor groups - Sector (public) - Size of organisation
	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-categorisation as retired & age 50+ - Receives old-age pension & age 50+	- Unemployed: actively looking for work - Self-employed - Entrepreneur (including or not including farmers) - Retired fully - Receiving old-age pension

Group of variables	Description	Main variables
	- Age 65+	
	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 original 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 original 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 income indicating a loss or debit (e.g. PSID since 1994). Values are in local currencies.	- Individual Income (All types) • year, net • month, net - Individual Labor 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
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 harmonised and should be modified by the users according to specific conceptual framework (e.g. defining chronic conditions). CPF provides several dimensions of subjective wellbeing, which can be harmonised 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 >30% - 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)
Ethnicity	Self-reported ethnicity based on broad categories. For the US, a separate variable is available for hispanicity	- Ethnicity - Hispanicity (US only)
Migration	Country of birth of respondents and (if available) their parents. For foreign-born individuals, country of birth is categorised in global regions. Derived from the country of birth variables is a set of variables relating to the respondents migration status.	- Country of birth - Migration status - Migrant generation (derived)
Religion	A binary variable is available indicating whether the respondent self-identifies as belonging to any religious group. Further information is available on attendance of religious services which is placed on a 4-point scale. If required, the original values were rescaled. Additionally, a separate variable is available for Korea only reflecting religious participation, which may be relevant as a proxy for attendance (which is not available for Korea)	- Religiosity - Attendance religious services - Religious participation

Samples

In the default settings, CPF includes observations from individuals aged 18 and older and meet the following criteria:

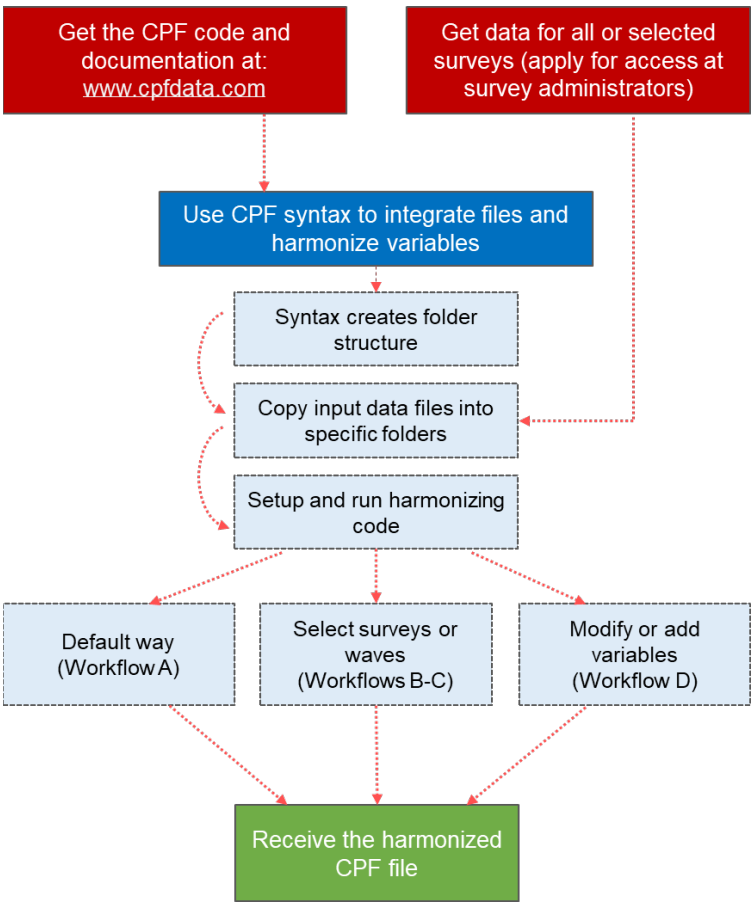
1. **Interview status**:
    a. South Korea, Switzerland, the UK: keep all observations (including proxy responses)
    b. Australia, Germany, Netherlands: keep direct respondents only
    c. The US: only reference persons (heads) and partners (spouses) (see details on PSID for explanation)
1. **Age**: 18 and older
2. **Delete observations** with missing values for gender and age (a minor correction)

Users can easily modify these selection criteria (see: *Workflow D - Adjustments to sampling criteria*).

5. How to work with the CPF code

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. The code allows combining the separate raw survey data into a single harmonised data file. A step-by-step guideline of how to work with the CPF is presented in Figure 6. Users must first apply for access to each of the original datasets (see: *Obtaining the original data*). 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 hierarchical design of the code allows locating all the steps in the algorithms easily. Country-specific syntaxes are commented and organised in a similar way to facilitate the work.

Figure 6. *A step-by-step guide on using the CPF code*



There are four general ways of working with the CPF syntaxes (workflows). *Workflow A* describes the basic approach which constructs the data without any modifications. *Workflows B*, *C* and *D* refer to different modifications of the existing syntaxes, with *Workflow D* being the most flexible and advanced. All approaches are described in the following paragraphs. More details are in the syntax's comments which additionally include references to the *workflows A-D* to highlight places which might require adjustments.

Getting started

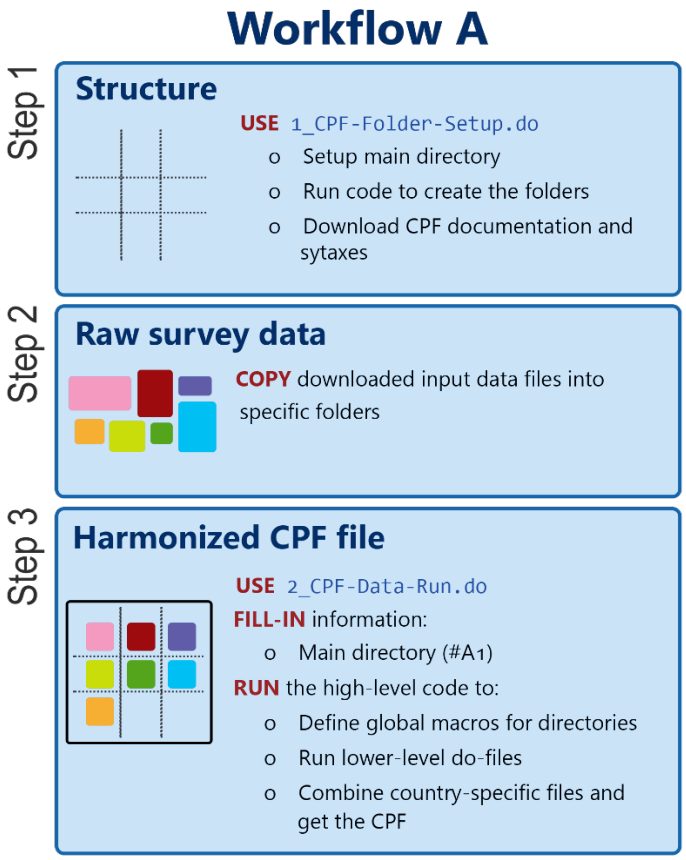
How to start using the CPF:

1. Download the latest CPF-code from the Github, OSF or the CPF webpage (all sources are synchronized and provide the latest version of the code). For example, to code can be downloaded as **CPF-Code-main.zip**. Unpack the entire folder structure with CPF-do-files to `11_CPF_in_syntax`. Or simply rename the un-packed `CPF-Code-main` as `11_CPF_in_syntax`.
2. Run `1_CPF-Folder_Setup.do` according to instruction in the do-file.
3. Copy the original input datafiles (e.g. from SOEP) to specific folders (e.g. `C:\CPF\02_Country_Data_Origin\06_SOEP\data`).
4. Go to `2_CPF-Data-Run.do` and follow instructions regarding the worfklows.

Three steps to get the CPF data (Basic workflow A)

The basic way of working with the CPF syntax leads from the raw data to a CPF harmonised dataset without any modifications (such as modifying variables, adding new variables, adding new waves, or selecting countries – for these see the next part). The approach requires only the use of two higher-level syntaxes (1 and 2). Workflow A consists of three basic steps, as presented in Figure 7:

Figure 7. *The basic way of working with the CPF syntax (workflow A)*



Users must first fill in the necessary information, such as the directory in the first syntax (1_CPF-Folder_Setup.do), and then run it to create an appropriate folder structure (see the section on *Folder Structure*). Then, they can place the downloaded data in specific folders of the `02_Country_Data_Origin` main folder. There is a separate subfolder for each survey (e.g., 01_HILDA) with a 'Data' subfolder, where the files should be copied, e.g., `02_Country_Data_Origin\01_HILDA\Data` (see details in *Obtaining original data*). The next step uses the second syntax (*2_CPF_Main__Fill_and_run.do*)

The third step uses syntax `2_CPF-Data-Run.do` to call all lower-level syntaxes (_10, _11, and _12). Users only have to fill in the address of the main directory in #1. (Also check the information on the number of waves in #3 and file names in #4 – see Workflow C). With this information, the code in parts #5-#7 can be simply run to activate lower-level syntaxes (_10, _11, and _12).

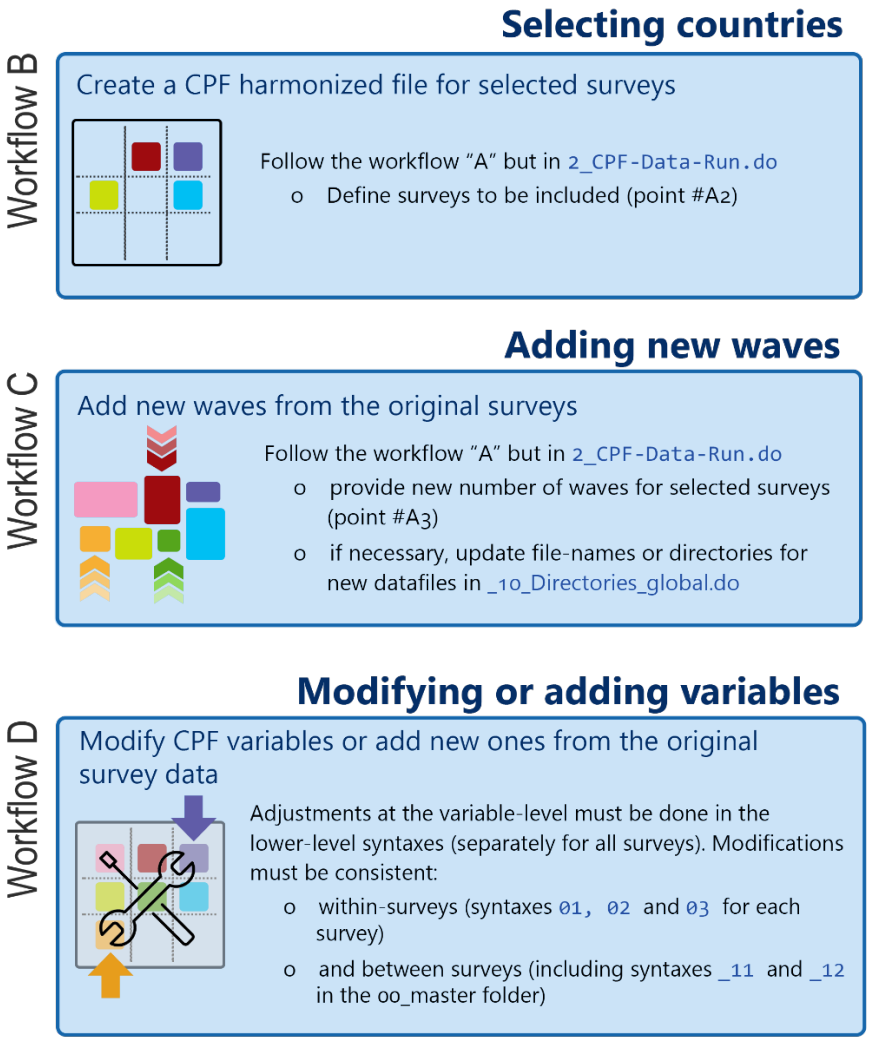
Since CPF v.2.0, previous syntax _11_Run_cntr_do_files.do (which was calling all country-specific syntaxes _01, _02, _03) was removed (and names of other syntaxes were adjusted). Now the reference to country-specific syntaxes is included in 2_CPF-Data-Run.do and required to run each country code separately (this is more efficient and less prone to errors).

Note that many operations are complex, and running the code can easily take an hour or two on an average computer. More details are in the syntax's comments. For any modifications or problems, refer to workflows B, C or D.

Modifying and adding data (Advanced workflows B, C, and D)

The other workflows can be used if users wish to modify or add data (Figure 8).

Figure 8. *Advanced ways of working with the CPF syntax (workflows B, C, and D)*



Selecting countries (Workflow B)

The code easily allows the selection of surveys to be included in the harmonized CPF dataset (e.g, for users who do not have access to all surveys or have no need to harmonise all of them). The only difference in the procedure compared to Workflow A, is to define surveys to be included in `2_CPF-Data-Run.do` (point #A2). For example, to keep all surveys, simply leave all their names in the global macro (use lowercase):

```
global surveys “hilda klips psid shp soep ukhls liss”
```

To select only PSID, LISS and UKHLS, keep only respective names in the code:

```
global surveys “psid ukhls liss”
```

Then, in B1, run only the do-files for the selected countries, e.g., B1C for PSID and B1F for UKHLS.

Adding new waves (Workflow C)

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, however, some users might want to modify the syntaxes on their own. In most of cases, the procedure should be easy and limited to filling in information in `2_CPF-Data-Run.do` on the number of waves in #A3 (for HILDA, KLIPS, SHP, SOEP, and UKHLS), e.g.:

```
global hilda_w "20"  // for 18 waves
global klips_w "26"  // 26 waves
global ukhls_w "14"  // 14 waves (it refers only to the UKHLS, not BHPS waves)
```

For PSID, adding new waves is more complex and must be done manually for each variable (see: *Survey-specific details on PSID*). The latest name of the individual dataset has to be added in:

```
global psid_ind_er "${psid_in}/pack/IND2017ER.txt", clear
```

For PSID, folder names will also need to be updated – this can be done in `1_CPF-Folder-Setup`, under “Additional survey-specific folders”.

The same might be required for new waves of UKHLS – in syntax _10_Directories e.g.:

```
global ukhls_data "UKDA-6614-stata/stata/stata13_se"
```

Note that releases of the new waves can also bring changes to the original data structure, which do not fit the current CPF algorithms, such as changes in names of variables, files, or directories. In such cases, additional adjustments have to be made in higher-level syntax `2` and/or lower-level syntaxes `_01` (see *Workflow D*).

Modifying and adding variables (Workflow D)

Workflow D refers to all other modifications 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 for the master-syntaxes. Depending on the character of modifications, the procedure can be easy or complicated.

Adjustments to variables

1. When adding or modifying variables, users should:
   - Carefully explore questionnaires, codebooks and data
   - Assess the consistency of original variables across waves
   - Assess the consistency of new variables between countries
   - Perform check-up of the new variables within a country (logical rules, distributions, cross-tabulations)
2. Adjustments to variables must be first introduced in the survey-specific syntaxes `01` and `02` stored in the survey folders (e.g., `11_CPF_in_syntax\01_HILDA`). Note that for some surveys, there are multiple syntaxes at levels `01` and `02`. The main steps include:
   - New variables can be added to the main CPF dataset using do-files `01`. The procedure depends on the structure of the original data. For most of the surveys (HILDA, KLIPS, RLMS, UKHLS), all original variables are already available in the **`xx_01.dta`** file created with the `01` syntaxes (note that code for UKHLS drops some variables at the end). Note that for UKHLS, variables that need to be kept in the harmonized file need to be listed under the isvar command for both BHPS and UKHLS. SOEP and SHP require adding variables from multiple source data files using `01` code. The procedure is more complex for PSID, where a whole set of variable names has to be added (so-called item-blocks) using `01_3`. More details can be found in the survey descriptions and instructions in the do-files.
   - Modifications of the new or existing variables (such as recoding, combining multiple variables, renaming etc.) can be done in do-files `02`. This is a place to harmonise the variables across surveys.
   - Always include new or modified variable names in the *keep* commands at the end of the file.
   - Syntaxes `03` do not have to be adjusted in case of variable-level modification.

3. Further on, users have to adjust the master (between-surveys) syntaxes `_11` – `_12` stored in the `11_CPF_in_syntax\00_master` folder as follows:
   - New or modified variables must be added to the *keep* code after appending data in `_11`.
   - Then, appropriate labels must be included in `_12`.

Adjustments to sampling criteria

Adjustments to the sampling criteria can be done in syntaxes `03` separately for each survey. They should not require additional modifications in other syntaxes.

Doing analysis with the CPF

To account for the hierarchical data structure, users can refer to the following variables:

- ***country*** – to identify countries (surveys)
- ***pid*** – to uniquely identify respondents (based the original id number from source surveys)

- ***wave***, ***wavey,*** or ***intyear*** – to include the time dimension:
    - ***wave*** – country-specific wave number (counting from 1)
    - ***wavey*** – the main (initial) year of data collection for a given wave
    - ***intyear*** – year of interview

There are different approaches to account for the entire 3-level hierarchical structure. For example, users can include countries as dummies, perform contextual analysis, run separate analyses by country, or use robust (clustered) standard errors. Performing a multilevel model with all three 3-levels is problematic due to the low number of countries (however, the Bayesian approach can be considered in this case). For more information on multilevel and panel analysis, we recommend popular statistical handbooks:

- Gelman A., J. Hill (2007) "Data Analysis Using Regression And Multilevel/Hierarchical Models", New York: Cambridge University Press
- Snijders, T., R. Bosker (1999) "Multilevel Analysis: An introduction to basic and advanced multilevel modeling**",** London: Sage.
- Raudenbush, S.W., A.S. Bryk (2002) "Hierarchical Linear Models: Applications and Data Analysis Methods", Thousand Oaks, CA: Sage Publications
- Joop Hox (2002, 2010) "Multilevel Analysis: Techniques and Applications", Routledge
- Hoffman L., (2015) "Longitudinal Analysis: Modeling Within-Person Fluctuation and Change", Routledge
- Singer J., J. Willett (2003) "Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence", Oxford University Press
- Rabe-Hesketh S., A. Skrondal (2012) "Multilevel and Longitudinal Modeling Using Stata (3rd Edition)", Stata
- McElreath (2020). "Statistical Rethinking: A Bayesian Course (2nd Edition)", CRC Press
- Kruschke (2014) "Doing Bayesian Data Analysis: A Tutorial Introduction with R", Academic Press

For example, in Stata, a simple regression model that accounts only for the country clustering by including a country dummy can be written as:

```
reg satlife5 i.edu3 i.country
```

A panel model, which accounts additionally for repeated observations (2-level model) can be written with the `mixed` command as:

```
mixed satlife5 i.edu3 i.country || pid:,
```

After defining a panel structure with `xtset`, a similar model can be written with the xt-command:

```
* Define panel structure
xtset pid wave // counting waves from 1
* OR:
xtset pid wavey // counting waves by the calendar year
* Panel model
xtreg satlife5 i.edu3 i.country
```

General recommendations:

- Before any substantial analysis, consider the missing values (MV) in the used variables. In most cases, MV should be removed from the analysis, e.g. by recoding them into system-missing values (`.`), of applying `mvdecode` command.
- Given that the CPF data (in version 1.0) contains more than 2.5 million observations, complex statistical analysis can run slowly. In such a case, run the initial analysis on subsamples. Alternatively, consider other statistical software, such as Mplus or R (especially for Bayesian analysis).
- Be aware of different time frames and differences in gaps between years of data collection (e.g. in PSID) when performing longitudinal analysis. Depending on the research question and method, consider using ***wave***, ***wavey*** or ***intyear***.

Troubleshooting

Computer requirements

- The CPF code is available in Stata, and it has been prepared in Stata 18. Running the entire code can easily take an hour or two on an average computer. A faster processor and more working memory will speed up the process.
- It is recommended to have at least 80 GB of storage hard drive space if all countries are included (the original data files require a minimum of 50 GB, and the CPF working and output files need an additional 25 GB).

Large size of files and computer power limitations

- Some surveys, particularly UKHLS and SOEP, have some very large files and operations may be challenging with limited disk space or computer power.
- In such a case, users can add an option to keep only the necessary variables at the end of `01_Prepare_data.do` for particular countries. As default, the CPF code does not drop variables (with some exceptions, e.g., UKHLS) to provide an easier way to add new variables to the harmonized CPF file.
- Users can also run the code step-by-step, in smaller pieces, or delete unnecessary files (e.g. files **01** and **02**, and leave the final **03** file only).

Searching for errors

- Although the structure of the CPF code is complex, there are ways to locate errors if the expected outcome does not appear.
- It's recommended to run the code step-by-step, in smaller pieces, also to learn the procedure better. When the code is run from the higher-level syntax, Stata does not print errors by default. Running the code from the lower-level syntax (following the order of the files, i.e., 01, 02, and 03) allows for better control.

- For each country, the code produces three files (apart from other temporary or working files) numbered **01**, 02, and 03, which correspond to the respective syntax names. If the code does not produce file **02.dta**, it suggests that it's best to search for errors in syntax 02.do.
- Many errors may be related to problems with defining global macros in syntaxes 1, 2, and 10. Especially when adding new waves or working on older releases of the source data.

Older versions of the source files

- CPF version 2.0 was built on data versions available in spring 2025. They include KLIPS ver. 26, PSID ver. 2023, SHP ver. 25, SOEP v. 40, UKHLS ver 14, LISS 2024, and HILDA ver. 2000. Note the HILDA was not update with new waves because access to the data was not granted at the time of publication of CPF 2.0.
- CPF version 1.5 was built on data versions available in 2023. They were released in 2020 (PSID ver. 2019), 2021 (HILDA ver. 2000), 2022 (SOEP ve. 37, RLMS ver. 2021, UKHLS ver 12, SHP ver. 22) and 2023 (KLIPS ver. 24).
- New waves will be continuously integrated into the CPF code; users can also do this independently (see Workflow C in Using the CPF syntax).
- Backward compatibility with older releases may not be available for some variables or surveys due to changes in variable names and file structure in the original data sources
- To work with older versions of the source data, first change information in 2_CPF-Data-Run.do in point #A1. (e.g., change *global soep_w "35"* to "34" or older). However, names of some variables could have changed between waves. In such a case, the syntax (mostly syntax 01 for a particular survey) would have to be modified. In such a case, GitHub provides access to older releases of the CPF code (https://github.com/cpfdata/CPF-Code/releases) and documentation (https://github.com/cpfdata/CPF-Documentation/releases).

Questions and answers

- The best place to questions and answers is by email (contact@cpfdata.com) .

6. CPF syntax: the general design

Folder structure

** Note for CPF 2.0: some do-files have changed names (e.g., from 1_Folder_setup → 1_CPF-Folder-Setup). There were also do-files added to the country-specific folder (_xx_00_) to replace the deleted _11 file.*

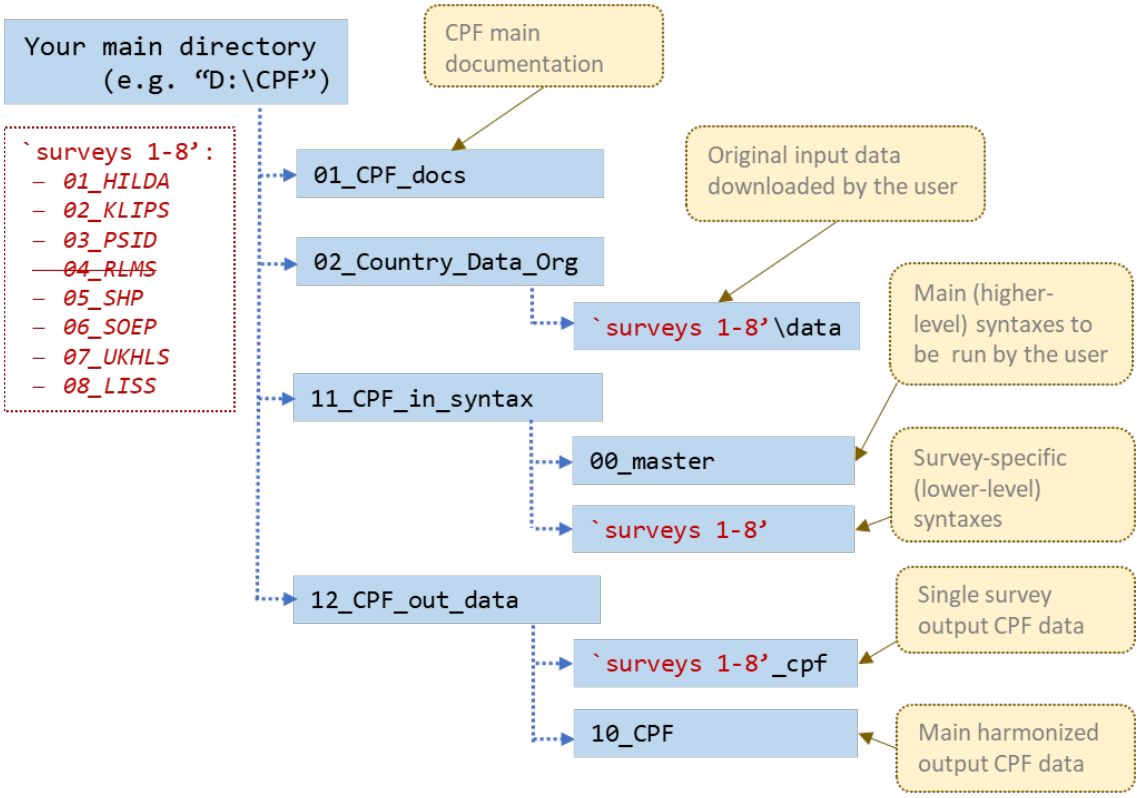
To run the CPF code, users must first create a folder structure as shown in Figure 9. It is done automatically by running `1_CPF-Folder-Setup.do`:

1. First, insert the directory to your CPF folder in #1 ("Your local directory"), e.g.:

```
global your_dir " D:\CPF"  // <--insert your directory
```

2. Running the rest of the code (#2-#3) will create appropriate folders.
3. Then you can copy the original input datafiles to specific folders

Figure 9. *Structure of the folders required for the CPF algorithms*



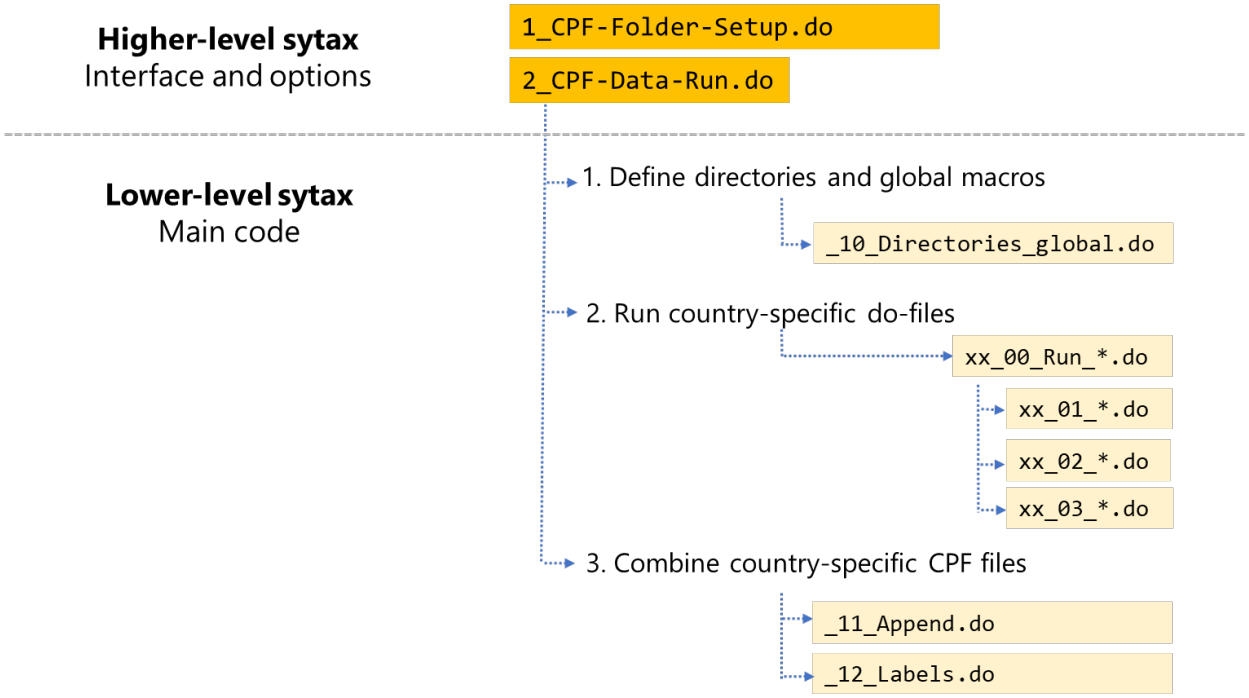
Syntax: higher and lower-level code

CPF syntaxes (codes) are designed at two levels: higher and lower (Figure 10). Two *higher-level syntaxes* are: `1_CPF-Folder-Setup.do` to set up the directory and folder structure, and `2_CPF-Data-run.do` to harmonise the dataset. These are short meta-codes and do not refer directly to variables or data files. Instead, they work as an interface and allow for filling in the necessary information (e.g., file directory) and setup options for harmonisation (e.g., which surveys to include). Higher-level syntaxes call all the required code of 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 follow the same three steps: the first constructs initial separate country data in a long format by merging the original files, the second harmonizes variables within countries according to a common 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 a single CPF harmonised dataset using a higher-level syntax.

Users can easily follow the instructions to build the comparative file in the default way (Workflow A) or modify the procedures according to their needs. In the latter case, the hierarchical design of the code allows for the quick location of all the steps in the algorithms. Country-specific syntaxes are commented and organised in a similar way to facilitate the work. Many users can be interested in syntaxes *_01, which contain code that integrates all raw files into a single and ready-for-analysis (yet un-harmonised) country data set*.

Figure 10. *Structure of the higher and lower level syntaxes*



Design of the lower-level code

Harmonisation of the data is done in four steps using a lower-level syntaxes specific for each survey (stored in `11_CPF_in_syntax\`survey'\`). The first step is to construct the base separate-country data in a long format, the second – to harmonise variables within a country-files, the third – to select the sample, and the fourth – to combine all country files into a single CPF harmonised dataset. All of these steps can be run from the higher-level syntax `2_CPF….do` (point B1) which activates country-specific syntax _xx_00 (e.g., _au_00_Run_HILDA.do), which runs all country-level syntaxes described below.

Step 1. Preparation of the long-format base file for each country

- **Input:** original surveys' datafiles (stored in `02_Country_Data_Origin\`survey'\Data`)
- **Syntaxes**: `11_CPF_in_syntax\`survey'\xx_01_*.do`
- **Result**: data file(s) in a long format: `12_CPF_out_data\`survey'_cpf\`**`xx_01_*.dta`**

First, for each country, we must construct a base file in long format, which contains all (or a selected subset) of the source variables provided by the data supplier. A long-format of panel data means that the repeated observations are clustered by individuals so that each row of data refers to respondent's information from a specific wave (contrary to wide data, where wave-specific information is provided in separate variables, e.g. health_wave1, health _wave2).

The procedure varies by country, and its complexity depends on the data structure. Most of the datasets require combining (appending and merging) separate files for specific waves and/or types of surveys (e.g., individual, household, or topic-specific questionnaires). Codes for this stage often include a group renaming of variables to a format that is required in a long-format file (e.g., removing wave-reference in variable names). For the PSID, the procedure is much more complicated, since it needs first to retrieve and combine sets of variables that refer to the same question or concept. The RLMS, on the other hand, is already provided in a long format for specific types of questionnaires. For some countries, a pre-selection of variables and initial cleaning is done at this stage. In addition to the raw data files, in some cases, CPF also uses selected CNEF variables (separate CNEF data files are provided for HILDA, SHP, and SOEP).

Step 1 uses lower-level syntaxes `xx_01_*.do`. The result of this step is one or more data files **`xx_01_*.dta`** with a large number of non-harmonised variables in a long data format.

Step 2. Harmonisation of variables within a country

- **Input**: `12_CPF_out_data\`survey'_cpf\`**`xx_01_*.dta`**
- **Syntaxes**: `11_CPF_in_syntax\`survey'\xx_02_*.do`
- **Result**: a single data file for each country with a harmonized variable structure (`12_CPF_out_data\`survey'_cpf\`**`xx_02_CPF.dta`**)

In the second step, we use variables from the **`xx_01_*.dta`** base-file(s) to construct new harmonized variables. The harmonised variables are the same for all countries in terms of names, format and response categories. However, some of them are available only for selected surveys.

The harmonisation process involves:

- Recoding and combining original variables into new variables
- When necessary, creating additional versions of variables (not fully harmonised)
- Selecting additional variables to keep (e.g. weights, sample characteristics, other country-specific variables)
- Basic data cleaning (which is not a full preparation for analysis)

Step 2 uses lower-level syntaxes xx_02_*.do. A result of this step is a single datafile **xx_02_CPF.dta** with a set of harmonised variables in a long data format.

Step 3. Sample selection

- **Input**: 12_CPF_out_data\`survey'_cpf**xx_02_CPF.dta**
- **Syntaxes**: 11_CPF_in_syntax\`survey'\xx_03_ Sample_selection.do
- **Result**: a single data file for each country with a harmonized variable structure (12_CPF_out_data\`survey'_cpf**xx_03_CPF.dta**)

During the third step, we select the final sample to be included in the main CPF dataset. The selection is based on the interview status (keeping different types of interviewed respondents or proxy-interviews), age criteria (age 18+) and missing values (keeping individuals with information on age and gender). For some surveys, e.g. PSID or SOEP, selection of the sample is not straightforward, and users might want to adjust the criteria based on their research needs. Step 2 uses lower-level syntaxes xx_03_ Sample_selection.do. A result of this step is a single datafile **xx_03_CPF.dta** with harmonised sample criteria.

Step 4. Combining country data into a one harmonised CPF dataset

- **Input**: harmonized separate survey datafiles 12_CPF_out_data\`survey'_cpf**xx_03_CPF.dta**
- **Syntaxes**: 11_CPF_in_syntax\00_master_11_Append.do and _12_Labels.do
- **Result**: a single CPF data file for all countries (12_CPF_out_data\10_CPF**CPFvX.X.dta**)

Finally, all separate country-files with a harmonised structure of the data are merged into a single harmonised CPF dataset. It is done by running _11 syntax. Additionally, labels for variables and categories are added in syntax _12. The result is the final CPF dataset, e.g. **CPFv2.0.dta.**

7. Survey-specific details: handling the data and code

Installing the source data in CPF folders

Users must first apply for access to each of the original datasets independently at the 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, data can be extracted to specific CPF subfolders in the `02_Cntry_Data_Orgin/`survey'/Data` folders, as explained below. With new waves, users have to modify global macros in #3 of syntax 2 (see Workflow C), e.g.:

```
global klips_w        "21"          // number of waves
```

However, if the approach to naming variables, folders or datafiles in the original data changes in the future, additional adjustments have to be made in higher-level syntax 2 and/or lower-level syntaxes xx_01. Backward compatibility with older releases may not be available for some variables or surveys due to changes in variables names and file structure (but the syntax can be modified). New waves will be continuously integrated into the CPF code; users can also do this independently (see *Workflow C* in *Using the CPF syntax*).

Note: the specific file and folder names for the datasets may change between waves. The names given below serve as an illustration of the data structure and may not correspond to the names given in the latest data release.

01_HILDA – Australia

Apply for the data via the National Centre for Longitudinal Data Dataverse (Australian Government Department of Social Services): https://dataverse.ada.edu.au/dataverse/ncld. Unpack downloaded files, such as `STATA 190c (1-Combined Data Files)` and `STATA 190c (2-Other Data Files)`, to subfolders indicated as "Combined" and "Other" in the "Data" folder. The final structure should look as follows:

02_Cntry_Data_Orgin\01_HILDA\Data
- ↳ STATA 190c (Combined)
  - ↳ **Combined_r190c.dta**
  - ↳ **…**
- ↳ STATA 190c (Other)
  - ↳ **Household_r190c.dta**
  - ↳ **…**

Note that names of downloaded folders change between waves. The CPF-names for Hilda are given in the syntax in 10. The number 190 in folders' names refers to the current version (number of waves) of HILDA which is filled in #3 of syntax 2 as, e.g. 19 (see Workflow C):

```
global hilda_w  "19"          // version of HILDA, number of waves
```

02_KLIPS – South Korea

Data are available via the official website for registered users: www.kli.re.kr/klips_eng. Unpack all downloaded files directly in the Data folder, e.g.:

02_Cntry_Data_Orgin\02_KLIPS \Data

↳ **eklips01h.dta**

With new waves, users have to modify global macros in #3 of syntax 2 (see Workflow C), e.g.:

```
global klips_w        "21"          // number of waves
```

03_PSID – US

The logic behind PSID differs from other datasets and is much more complex (see *Survey-specific details* for PSID). To organise the data, we use **psidtools** ado (Kohler, 2015)[2], which can be downloaded using:

```
ssc install psidtools
```

Data are available via the official website for registered users: https://simba.isr.umich.edu/Zips/ZipMain.aspx:

1. Download all Family Files (one per wave, e.g. fam2019er.zip) and place them in into Family and Ind Files (zip). Do not unpack.
2. Download **Cross-year Individual: 1968-XXXX** zipped file and place it in Family and Ind Files (zip). Do not unpack.
3. Leave all files in the Family and Ind Files (zip) folder unpacked but additionally unpack the **Cross-year Individual: 1968-XXXX** zipped file (e.g. **ind2019er.zip**) to Data/Cross-year Individual 1968-XXXX/pack. It should contain a txt file with vales named, e.g. **IND2019ER.txt** (which is defined in 10 as `global psid_ind_er "${psid_in}\pack\IND${psid_w}ER.txt"` based on the latest PSID year indicated in 2 as `global psid_w`).

CPF syntax manages further reorganisation of the files. E.g. after running the lower-level syntaxes 01 for PSID, the code will unpack and combine required files into PSIDtools_files folder, and syntax 03 will create a number of item-specific files in the temporary directories.

After downloading, the PSID data-folder should look as following:

02_Cntry_Data_Orgin\03_PSID\Data

↳ Cross-year Individual 1968-2019

↳ pack

↳ **IND2019ER.txt**

↳ **...** *(place for automatically created file* ***psid_crossy_ind.dta****)*

↳ Family and Ind Files (zip)

↳ Cross-year Individual 1968-2019.zip

[2] PSIDTOOLS is Stata' module to facilitate access to PSID, developed by Ulrich Kohler from the University of Potsdam. See: https://ideas.repec.org/c/boc/bocode/s457951.html.

- ↳ **fam1968.zip**
- ↳ **ind2019er.zip**
- ↳ **…**

↳ PSIDtools_files

- ↳ **…** *(place for automatically unpacked files, e.g **fam1968.dta**)*

05_SHP – Switzerland

Data are available via FORSbase for registered users: https://forscenter.ch/projects/swiss-household-panel/data. Unpack all folders from **Data_STATA.zip** into the main Data folder. It should then contain several folders with different types of datasets. The main source of the individual- and household-level data are files in SHP-Data-W1-W21-STATA folder (e.g. **shp99_p_user.dta**). Additionally, CPF refers to other folders, including SHP-Data-CNEF-STATA and SHP-Data-WA-STATA. After unpacking, the structure should look as follows:

02_Cntry_Data_Orgin\05_SHP\Data

- ↳ SHP-Data-Biography-STATA
  - ↳ **SHP0_bh_user.dta**
  - ↳ **…**
- ↳ SHP-Data-CNEF-STATA
  - ↳ **shpequiv_1999.dta**
  - ↳ **…**
- ↳ SHP-Data-Imputed-Income-Wealth-STATA
  - ↳ **imputed_income_hh_long_shp.dta**
  - ↳ **…**
- ↳ SHP-Data-Interviewers-STATA
  - ↳ **shp00_v_user.dta**
  - ↳ **…**
- ↳ SHP-Data-SHP-3-W1-STATA
  - ↳ **shpiii_cs_user.dta**
  - ↳ **…**
- ↳ SHP-Data-W1-W21-STATA
  - ↳ W1_1999
    - ↳ **shp99_p_user.dta**
    - ↳ **shp99_h_user.dta**
  - ↳ **…**

↳ `SHP-Data-WA-STATA`

↳ **shp_ca.dta**

↳ **...**

The number of waves in the folder name `SHP-Data-W1-W21-STATA` is accounted for automatically after filling it in in the syntax `2_CPF_Main__Fill_and_run.do` in #3. Be aware, however, of any future changes in folder names.

06_SOEP – Germany

Data are available via the Research Data Center SOEP after granting access:
www.diw.de/en/diw_02.c.242211.en/criteria_fdz_soep.html

Data should be unpacked into `Data` keeping additionally the wave-specific subfolder (e.g. `soep.v35`), which contains then all the SOEP files.

02_Cntry_Data_Orgin\05_SHP\Data

↳ `soep.v35`

↳ **abroad.dta**

↳ **pl.dta**

↳ **...**

The wave-specific subfolder name is accounted for automatically after filling in the number of waves in the syntax `2_CPF_Main__Fill_and_run.do` in #3.

07_UKHLS - UK

Data are available via the UK Data Service after granting access: www.ukdataservice.ac.uk.

Data should be unpacked into `Data` with keeping additionally the specific subfolders' path (e.g. `UKDA-6614-stata\stata\stata11_se`), which contains then all the wave-specific folders. These folders (e.g. `bhps_w1`, `ukhls_w1`) contain the data files for each wave.

The specific path may change from wave to wave and has to be properly included in syntax `2_CPF_Main__Fill_and_run.do` in #4, e.g.:

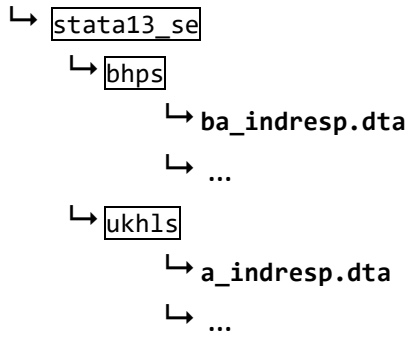
```
* UKHLS

global ukhls_data "UKDA-6614-stata\stata\stata13_se"
```

The structure should look like following:

02_Cntry_Data_Orgin\07_UKHLS\Data

↳ `UKDA-6614-stata`

↳ `stata`



08_LISS – Netherlands

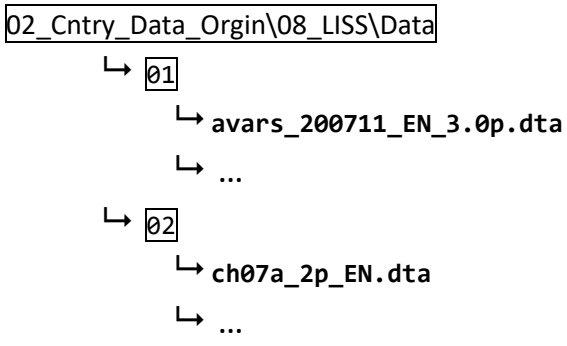
Data are available at https://www.dataarchive.lissdata.nl/.

The procedure of downloading the files may take some time. Data must be downloaded from the Background Variables and LISS Core Study. This has to be done file-by-file, wave-by-wave, thus can take some time (especially for Background Variables). Files to include in CPF are Stata .dta files, which can be either directly downloaded or unpacked from zip files.

One downloaded, dta. files have to be copied into specific subfolders of the `Data` folder: from `01` to `11`:

1 Backgroung variables
2 Health
3 Religion and Ethnicity
4 Social Integration and Leisure
5 Family and Household
6 Work and Schooling
7 Personality
8 Politics and Values
9 Economic Situation: Assets
10 Economic Situation: Income
11 Economic Situation: Housing

The structure should look like the following:



```
...
└→ 05
    └→ cf08a_2p_EN.dta
    └→ ...
...
└→ 11
    └→ cd08a_EN_1.2p.dta
    └→ ...
```

Handling the country-level code, adding new variables

This part presents additional details and instructions for the lower-level (survey-specific) codes.

01_HILDA – Australia

1. au_01_Prepare_data– to prepare data
   a. Step 1: Variable Renaming: Processes combined wave files (**Combined_{w}{year}c.dta**)
   b. Step 2: Append Waves: Appends all processed wave files into single longitudinal dataset
   c. Step 3: Final Dataset Creation: Saves final combined dataset: **au_01_combined_2001_20{wave}.dta**
2. au_02_1_Harmonize (p1- cnef)– code for the CNEF data file
3. au_02_2_Harmonize (p2- combined)– code for the original variables from all waves
4. au_02_3_Combine_p1p2 – combines p1 and p2 into a **xx_02_CPF.dta**
5. au_03_Sample_selection – sample selection

02_KLIPS – South Korea

1. ko_01_Prepare_data – to prepare data (a detailed description is included in the do-file)
   a. Install ado **renvars** (for renaming)

```
net install http://www.stata-journal.com/software/sj5-4/dm88_1
```

   b. Step 1: Prepare Person Files (p-files) Processing
      - • Processes individual person files (**eklips{w}p.dta**) for each wave
      - • Standardizes variable names by removing wave identifiers from variable names
      - • Appends all person waves into single file: **ko_all_p.dta**
   c. Step 2: Prepare Household Files (h-files) Processing

- • Processes household files (**eklips{w}h.dta**) for each wave
- • Applies similar variable name standardization as person files
- • Appends all household waves into single file: **ko_all_h.dta**

d. Step 3: Combine p & h files
- Merges person and household files within each wave using household ID
- Appends all combined person-household waves
- Creates final integrated dataset with both individual and household information

e. Step 4: Final Dataset Creation
- Saves final combined dataset: **ko_01.dta**

2. ko_02_Harmonize– to prepare harmonised variables
3. ko_03_Sample_selection – sample selection

03_PSID – US

The philosophy behind PSID data differs from other surveys. PSID is the oldest ongoing research in the CPF set and the most challenging to incorporate. Unfortunately, unlike in other surveys, adding new waves of PSID to CPF cannot be achieved at the file-level (e.g. by updating a file name). The reason is that names of variables which refer to the same construct (e.g. age, employment status) change from wave to wave (e.g. the variable for employment status is named ER30509 in 1986, ER33111 in 1994, ER34317 in 2015, and ER34516 in 2017). Therefore, all variables must be retrieved separately from all waves by searching in PSID's online system (https://simba.isr.umich.edu). This is a challenge for users who would like to add new items or new waves to the CPF data. To add to the complexity, items are stored in separate variables for Reference Persons (called Heads in older waves) and Partners (called Spouses in older waves). Also, similar questions were sometimes framed differently for Reference Persons and Partners and included in different waves. Therefore, the syntax for PSID is more complex and requires additional clarification.

1. us_01_1_Create_psid_crossy_ind

This lower-level code is run from 2_CPF_Main… syntax to create **psid_crossy_ind.dta**. Note, however, that code us_01_1 uses the input code provided by PSID in the zipped file (e.g. IND2017ER) which might have to be updated with new waves. Also, the code us_01_1 refers to the individual data file at the end of the **infix** part, which has to be updated in the 2_CPF_Main.

2. us_01_2_Create_waves_psidtools

The **psidtools** implemented in this code will automatically unpack the files, prepare and copy them into PSIDtools_files (one family file per wave and one individual file). These are the base input files for the CPF dataset. After this, the zipped files in Family and Ind Files (zip) folder can be deleted.

Note for CPF 2.0 about psidtools:

At the moment of preparing the CPF 2.0 (6.2025), Psidtools required a minor update to work with the 2023 wave. The original Psidtools package was updated only until the PSID 2021 wave. Thus,

it is necessary to install the temporary version of Psidtools. It is based on the original code of the Psidtools creator, Prof. Dr. Ulrich Kohler (https://gitup.uni-potsdam.de/ukohler). The source code is available at: https://gitup.uni-potsdam.de/ukohler/psidtools/-/blob/main/psid.ado. The code was updated by K.Turek (minor adjustments at the end of the code, e.g., 'stop at 2021' changed to 'stop at 2023') and is included in the CPF 2.0 syntax (in the PSID folder).

3. us_01_3_Get_vars

When adding new variables or new waves to PSID, users must modify the us_01_3_Get_vars. It contains a **combvars** program, which is a wrapper for the **psidtools** command. **Combvars** is used to

- combine variables across waves - strings of variables related to the same item are different for each wave (e.g. age: ER30004, ER30023, etc.)
- reshape them into a long format (using psidtools)
- save them as separate files
- add names to global macro for further use (merging)

The **combvars** program uses files created by **psidtools** in PSIDtools_files folder. However, names of the variables have to inserted by hand (names can be found and copied from the PSID's online search tools at https://simba.isr.umich.edu/Zips/ZipMain.aspx) into specific item-lists. For example, the code to add and combine original variables which refer to the age of respondent is:

```
combvars age, list("[68]ER30004 [69]ER30023 [70]ER30046 [71]ER30070 [72]ER30094
[73]ER30120   [74]ER30141   [75]ER30163   [76]ER30191   [77]ER30220   [78]ER30249
[79]ER30286   [80]ER30316   [81]ER30346   [82]ER30376   [83]ER30402   [84]ER30432
[85]ER30466   [86]ER30501   [87]ER30538   [88]ER30573   [89]ER30609   [90]ER30645
[91]ER30692   [92]ER30736   [93]ER30809   [94]ER33104   [95]ER33204   [96]ER33304
[97]ER33404   [99]ER33504   [01]ER33604   [03]ER33704   [05]ER33804   [07]ER33904
[09]ER34004   [11]ER34104   [13]ER34204   [15]ER34305   [17]ER34504")
```

Unlike in other surveys, the names of variables in PSID change from wave to wave. Thus, names must first be combined across waves. Each of the items on the list refers to a specific wave (e.g., [68] refers to the wave from 1968). The last item on the list refers to the last wave (here: [17]ER34504). If new waves are to be added, users have to add an item to the list with the name of a variable in the latest wave (e.g. [19]ER…). It has to be done for all variables separately.

Thus, the syntax us_01_3 contains the following steps:

- Step1: Define **combvars** program to be used in this syntax and run `global vars""`
- Step 2: Run **combvars** to combine vars across waves. This will create separate long files for each item
- Step 3: Combine single-item long files from step 2 into **us_01.dta**
- Step 4: Add variables which constant across all waves to **us_01.dta** (get them from long **psid_crossy_ind.dta)**

Command “Add new time-constant vars - only if necessary” can be used to add new time-constant variables once the **`us_01.dta`** is already created.

Command “Add new files - only if necessary” can be used to add a new block of items using **combvars** (after creating **`us_01.dta`**).

EXAMPLE: adding new waves

1. Old code:

```
combvars age, list("[68]ER30004 [69]ER30023 [70]ER30046 [71]ER30070 [72]ER30094
        [73]ER30120 [74]ER30141 [75]ER30163 [76]ER30191 [77]ER30220 [78]ER30249
        [79]ER30286 [80]ER30316 [81]ER30346 [82]ER30376 [83]ER30402 [84]ER30432
        [85]ER30466 [86]ER30501 [87]ER30538 [88]ER30573 [89]ER30609 [90]ER30645
        [91]ER30692 [92]ER30736 [93]ER30809 [94]ER33104 [95]ER33204 [96]ER33304
        [97]ER33404 [99]ER33504 [01]ER33604 [03]ER33704 [05]ER33804 [07]ER33904
        [09]ER34004 [11]ER34104 [13]ER34204 [15]ER34305 [17]ER34504 [19]ER34704
        [21]ER34904 [23]ER35104")
```

2. Find age https://simba.isr.umich.edu/default.aspx
   - copy the last wave with age variable - `ER35104`
   - search for `ER35104`
   - select 'i' (information) icon under 'Name'
   - copyt the last wave's code for age: `[23]ER35104`

3. Add new wave to the code:

```
combvars age, list("[68]ER30004 ...
        [21]ER34904 [23]ER35104")
```

4. us_02_Harmonize
   - Selecting observations – the default option is “Keep 2: heads & partners”. The current version of CPF is not adjusted to include other family members. However, users may choose different sets of observations if necessary.
   - For ISCO variables, the code refers to external do-files (due to their length they are stored separately). Note that ISCO recoding can take a lot of time.

5. us_03_Sample_selection

In the default option, it keeps spouses and partners only (Keep 2), repeating the code in 02.

05_SHP – Switzerland

1. `ch_01_1_Prepare_data_Equiv_to_long`– to prepare supplementary CNEF variables
2. `ch_01_2_Prepare_data_Waves_to_long`– to prepare individual data

   This is also a place for adding new variables

- there are 3-4 places you have to put the name of a new variable from the wave-specific files you want to add
- these places are indicated as:

```
*>>>
*>>> NEW VARS [x*x; y*] 1/3:
*>>>
```

- you must adjust the formatting of the name in each case
- ***x*x*** - variables with year inside of the name, e.g. ***p17e50*** (3 places to add)
- ***y**** - variables with the year at the end of the name, e.g. ***educat17*** (4 places to add)
- please, verify if the results are correct, there are a few rules which help to check it

3. `ch_02_1_Harmonize_ (p1- equiv)`– to prepare supplementary variables from CNEF
4. `ch_02_2_Harmonize_ (p2-waves)`– to prepare the main variables
5. `ch_02_3_Combine_p1p2`– combine the main file with CNEF variables. Additionally, some missing values are cross-filed.
6. `ru_03_Sample_selection` – sample selection

06_SOEP – Germany

1. `ge_01_Prep_data`– to prepare data

   The ge_01_Prepare_data.do file processes SOEP panel data by selecting relevant variables from multiple specialized files and combining them into a single harmonized dataset for CPF analysis. Becouse of the size and number of source SOEP data files, we first select the variables that are included in the harmonization.

   Step 1: Prepare Specific Files - Due to SOEP's large size and numerous source files, variables are first selected from each specialized file:
   - ppathl.dta (Master File): e.g., Person identifiers, demographics, sample information
   - hpathl.dta (Household Path): e.g., Household identifiers and basic household characteristics
   - pgen.dta (Generated Variables): e.g., Labor income, employment status, ISCO codes
   - health.dta (Health Module):
   - pequiv.dta (CNEF Equivalents): e.g., Standardized income variables
   - pl.dta (Person Questionnaire): e.g., Detailed work variables, satisfaction
   - pkal.dta (Calendar Data): e.g., Previous year employment history
   - biobirth.dta (Birth Biography): e.g., Number of children
   - biol.dta (Life Course): Parental information, mother tongue
   - bioparen.dta (Parental Background): e.g, Parents' education, occupation

   Step 2: Merging Separate Files
   - Merges household data using household ID and year

Step 3: Final Dataset Creation

- Creates final integrated dataset: ge_01.dta

Adding new variables:

If users want to add new variables from the source data files, they should add them to the specific files.

Code `ge_01_Prep_data` is where new raw variables from the original SOEP data files should be added. Users must identify the specific original data file and variables and add the name(s) in an appropriate place in the syntax. For example, if variable ***newvar*** comes from SOEP's **`health.dta`**, the original name of a variable must be added to the `KEEP` command under the headline `*# health.dta #`., e.g.:

```
*############################
*#                          #
*#    health.dta            #
*#                          #
*############################
*
use "${soep_in}\health.dta", clear
*
keep                ///
...                 /// already added variables
newvar              // the newvar added
*
```

In case when the original data file is not listed in the syntax, it can be added in a similar way as the other files. First, open the new datafile, keep variables, and save the file under new name:

```
*############################
*#                          #
*#    newdatafile.dta       #
*#                          #
*############################
use "${soep_in}\newdatafile.dta", clear    // open the file

keep  ///
...   // add variables

rename persnr pid    // rename if neccessary
rename hhnr hid      // rename if neccessary
*
sort  pid syear
*
save "${soep_out_work}\gnewdatafile_1.dta", replace
```

Then, add the new file to the final `merge` command, e.g.:

```
*############################
*#                          #
*#     MERGING              #
*#                          #
*############################

***** HH + P
use "${soep_out_work}\gppathl_1.dta", clear
merge m:1 hid syear using "${soep_out_work}\ghpathl_1.dta" , keep(1 3) nogen

****
...
merge 1:1 pid syear using "${soep_out_work}\gnewdatafile_1.dta", keep(1 3) nogen
// adjust the code depending on the structure of the datafile
...
```

2. `ge_02_Harmonize` – to prepare harmonised variables
3. `ge_03_Sample_selection` – sample selection

07_UKHLS - UK

1. `uk_01_Prepare_data`– to prepare data

   The uk_01_Prepare_data.do file processes UKHLS/BHPS by combining two related longitudinal surveys into a single harmonized dataset for CPF analysis.

   Setup and Configuration:

   - Configures paths and wave parameters for both BHPS and UKHLS datasets
   - Defines wave structures: BHPS (waves 1-18, letters a-r) and UKHLS (waves 19+)
   - Merges stable characteristics from xwavedat file

   Step 1: Combine Individual Files

   - Variable Selection: Defines comprehensive list of variables for harmonization from both surveys
   - BHPS Processing: Processes individual files (b{w}_indresp) for waves 1-18, s
   - UKHLS Processing: Processes individual files ({w}_indresp) for waves 19+,
   - Merging time-invariant variables
   - Output: Combined individual dataset (uk_01_bhps_hls.dta)

   Step 2: Combine Household Files

   - BHPS Household Processing: Processes household files (b{w}_hhresp) for waves 1-18
   - UKHLS Household Processing: Processes household files ({w}_hhresp) for waves 19+

- Output: Combined household dataset (uk_01hh_bhps_hls.dta)

Step 3: Merge Individual and Household Data

- Merges individual and household files using household ID and wave
- Creates final integrated person-household dataset

Step 4: Final Dataset Creation

- Saves final combined dataset: uk_01.dta

Notes:

- Operations require much disk space. Therefore, temporary files are deleted
- Also, the combined file is large. For this reason, one of the last procedures in the syntax is `DROP` to delete variables which will not be used in further harmonisation.
- Users might have to adjust the command when adding new variables. Additionally, if the order of the variables changes with new editions, the `DROP` command must be modified (or deleted).

2. `uk_02_Harmonize` – to prepare harmonised variables
   - waves from BHPS and UKHLS have mostly separate sets of variables
3. `uk_03_Sample_selection` – sample selection

08_LISS – Netherlands

1. `nl_01_Prepare_data` – to prepare data
2. `nl_02_Harmonize` – to prepare harmonised variables
3. `nl_03_Sample_selection` – sample selection

The `nl_01_Prepare_data.do` file processes data from the LISS background variables and topic-specific data files, ultimately producing a harmonized dataset for CPF. The process includes the following steps (detailed description is included in the syntax):

Step 1: Background Variables Processing

1a. Prepare Background Variables

- Processes files from folder *01*
- Creates *wave* and *month* variables
- Generates standardized file names for subsequent processing.

1b. Combine Background Variables

- Appends all monthly background variable files into one (large) file: **`liss01_bckgr_all.dta`**

1c. Compute Incomes from All Months

- Computes income aggregates from all months to be used in the main file during harmonization

1d. Select Reference Months

- Selects reference months per year and generates the main background file for CPF integration with reference months
- Uses month 4 as the reference month (optimal for Work & Schooling topics)
- Alternative reference months available for other topics:
    - Month 1: optimal for Health, Politics, Religion
    - Month 6: optimal for Economic situation, Personality
    - Month 7 or 11: optimal for Health (waves 8-9)
    - Month 12: optimal for Politics, Personality (waves 8-9)
- Implements backup selection for cases without month 4 data (selects closest available month)
- Creates the final reference file: **`liss01_bckgr_ref.dta`** (contains only 1 observation per person per wave)

Step 2: Topic Files Processing

2a. Specific File Preparation

- Renaming and basic cleaning of the source files (for topic *08* only

2b. Process All Source Files

- Processes files from folders *02-11*
- Accounts for complexities in data collection periods and wave assignments, especially:
    - Topic 02 (Health): Skips wave 9 (2016); assigns correct year per wave.
    - Topic 08 (Politics): Skips wave 10 (2017); combines with wave 11 due to data changes.
    - Topic 09 (Economic Situation: Assets): Adjusts for biennial data collection.
- Creates some standardized variables (e.g., intdate, intyear, intmonth, wavey
- Outputs wave-topic files (e.g., **`liss02_08.dta`**) into corresponding "**`temp/'topic'/…`**" folders.

2c. Variable Format Correction

- Converts selected variables to string format where necessary.
- Handles missing values and resolves data type inconsistencies.

Step 3: Combine wave-files within topics

- Output files: **`temp/topics/liss'topic'`**, e.g., **`liss02`** – topic 2 with all waves

Step 4: Merge All Topics

- Creates a combined dataset for topics *02-11* (without background variables yet): **`liss01_all_topics.dta`**

Step 5: Final Dataset Creation - Merge Background & Topics

- Merges the background file (**`liss01_bckgr_ref.dta`**) with the topics file (**`liss01_all_topics.dta`**), creating final CPF-LISS dataset: **`nl_01.dta`** (keeps only participants with background variables)

Waves design in LISS

** Please see the [EXCEL FILE] for details.*

Each LISS Core questionnaire starts at a different month of the year. Consequently, the data collection period for each LISS wave is spread over many months. Additionally, the data collection periods for specific modules have been changing over the years. Some modules (e.g. Health, Politics) were collected over two separate calendar years (and the period has been changing), while some modules were also skipped in particular years.

Therefore, constructing the CPF-LISS files requires designing waves that will cover all modules collected over a longer period and combine them into a single wave. Additionally, the background variables are measured each month, so respondents have up to 12 measurements per year – we include this information as a single measurement (e.g, by selecting the reference month).

Please, consult the LISS data collection table (in CPF materials) for an overview of the LISS waves' design for CPF integration.

For CPF, the wave number of several LISS Core questionnaires differs from the wave number registered in the LISS Data Archive. For example, the first wave of the LISS Core Study was conducted in November 2007 with the LISS Core questionnaire *Health* and ended in April 2008. In the following years the data collection periods of *Health* changes. From the second wave of *Health*, the beginning of the collection period starts in November and ends in December of the same year.

The only exception was in 2015 when *Health* was fielded in July-August. From 2016, *Health* was fielded in November and December of the same year again.

The first wave of other LISS Core questionnaires mostly start in 2008. In the years 2014-2016, the time period between the waves of most questionnaires shifts, which causes changes in data collection periods in the following years.

Additional notes:

- 07 Personality – there is no wave 9 (2016). The first wave of the LISS Core questionnaire *Personality* was conducted in May 2008 and ended in August 2008. From the second wave of *Personality*, the beginning of the collection period starts in May and ends in June of the same year. In 2014 and 2015, *Personality* was conducted in November and December. *Personality* was not fielded in 2016. From 2017, *Personality* was conducted in May and June of the same year again.

- 08 Politics and Values - there is no wave 10 (2017) due to modification in the LISS data collection schedule. New data were combined with the next wave 11. The first wave of the LISS Core questionnaire *Politics and Values* was conducted in December 2007 and ended in March 2008. From the second wave, the beginning of the collection period starts in December and ends in January in the following year. In 2014, *Politics and Values* was not conducted in December. From 2017, the collection period starts in December and ends in March in the following year.

- 09 Economic Situation: Assets - asked every 2nd wavey only. The first wave of the LISS Core questionnaire *Economic Situation: Assets* was conducted in June 2008 and ended in September 2008.

Economic Situation: Assets is the only LISS Core questionnaire that was fielded every two years. From 2024, *Economic Situation: Assets* will be conducted every year.

Background variables

The *Background variables* contain the most important general characteristics of LISS panel households. The *Background variables* are measured every month using a separate questionnaire. For every LISS panel household, a single contact person completes this questionnaire. The questionnaire needs to be completed when joining the panel, before the household can start completing other questionnaires. Thereafter, the contact person is presented the *Background variables* questionnaire every month to enter any changes that may have occurred.

Some of the questions in the *Background variables* concern the household, and others concern the individual household members. All questions of the questionnaire are completed by the household contact person only. Some questions in the *Background variables* are similar to questions in the LISS Core questionnaires. Discrepancies between these variables may occur.

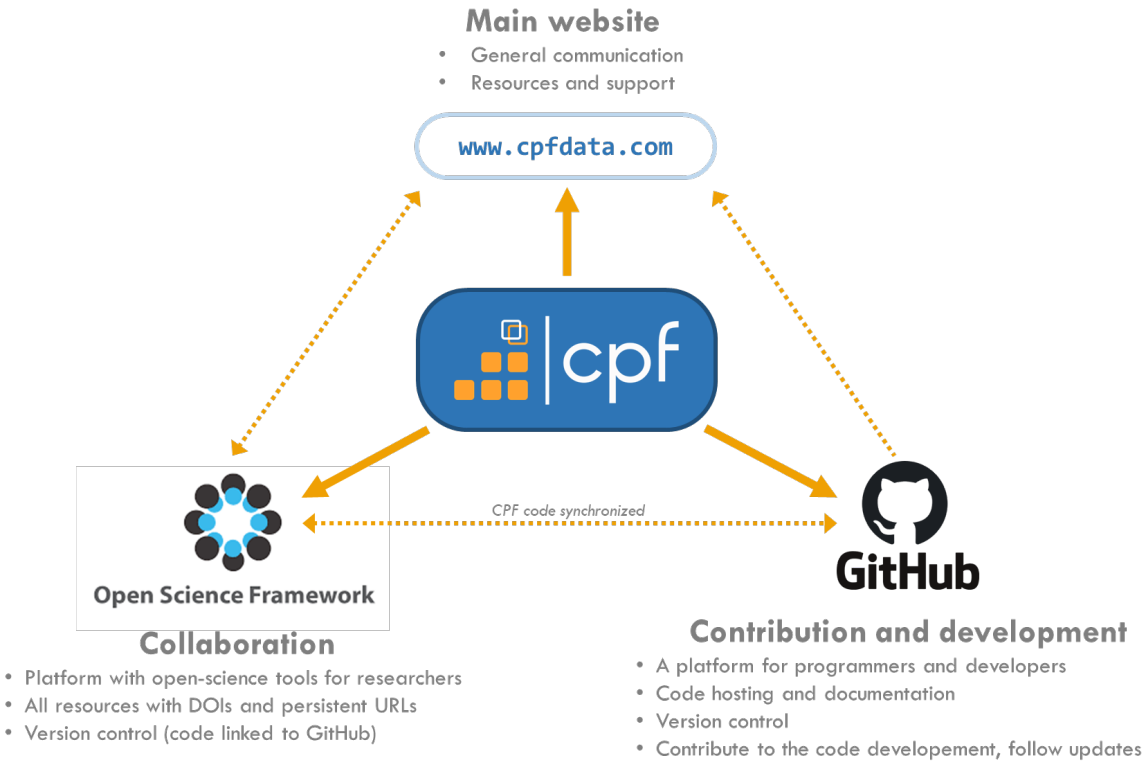
A data file containing details from the *Background variables* is made available for every month. The variables in these data files are presented per person, including variables that contain information on the entire household. These data represent a snapshot of the situation at the end of the field work period (a month) concerned.

8. Open-science platform for CPF

Tools and services

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 facilitate open access and community-based development, we have created an open-science platform that integrates several tools, including a website, GitHub, and OSF (Figure 11).

Figure 11. *CPF's open science platform*



The central element is the project's **website** (www.cpfdata.com) that contains all important information, documentation, and the latest major version of the code. The website also includes an online forum. The **GitHub** (www.github.com) is precisely oriented at the development of the CPF code. GitHub is a code hosting platform for collaboration in code development, especially useful for managing open-source projects. 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. Extensions, improvements or alternative versions of the code can be offered by all researchers and programmers who register free of charge at the GitHub platform. Importantly, all changes are recorded, providing version control functionality.

The Open Science Framework (OSF; www.osf.io) is one of the most popular open-science platforms, facilitating open collaboration in research. OSF integrates many tools and services which support

managing, organising, documenting and sharing all aspects of a project. Among other features, OSF allows for the pre-registration of studies, storing code and data; it is linked to preprint services and numerous scientific platforms. It facilitates collaborative workflow on projects, allows for documenting the work and progress. Similarly to GitHub, OSF uses a version control system, so all changes to the project are recorded. 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 be registered this way, receiving permanent links and DOIs. Importantly, OSF includes a GitHub add-on that directly links files stored in a GitHub repository to the OSF project. This way, changes to the code can be introduced either through GitHub or OSF, and they are synchronised so that the code at the OSF is always up to date.

Links to the resources:

- Website: www.cpfdata.com
- Forum: www.cpfdata.com/forum
- GitHub: www.github.com/cpfdata
- OSF: www.osf.io/h3yxq

Help and support

The up-to-date documentation of CPF can always be found at the project's website: http://www.cpfdata.com/download. Questions regarding the CPF code can be asked by email contact@cpfdata.com. CPF is an independent project developed on a voluntary basis. As such, it does not engage employees responsible for support and help. The CPF team will try to answer all the questions, but extensive support cannot always be provided.

Contribution and cooperation

User's improvements and suggestions will be recorded, incorporated and shared using open online platforms (i.e. web forum and GitHub code repository) to allow continuous development and regular updates to the official versions of the code.

CPF is an open and ongoing project. We invite interested users to provide feedback or contribute to the development of the code (through GitHub or OSF). We are also happy to cooperate in research or support the development of the research network by linking people and institutions. Do not hesitate to contact us!

Acknowledgments

This study uses the following datasets, for which we are grateful to the data providers:

- **The British Household Panel Survey (BHPS)** and **Understanding Society – The UK Household Longitudinal Study (UKHLS)**. Developed by the *University of Essex, Institute for Social and Economic Research (ISER)*, in collaboration with *NatCen Social Research* and *Kantar Public*. Data provided by the *UK Data Service*, Study Number: 6614. https://www.understandingsociety.ac.uk/
- **Socio-Economic Panel (SOEP)**, developed by the *German Institute for Economic Research (DIW Berlin)* under the umbrella of the *Leibniz Association*, with funding from the *Federal Ministry of Education and Research (BMBF)* and German state governments. https://www.diw.de/en/soep
- **Panel Study of Income Dynamics (PSID)**, public use dataset produced and distributed by the *Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI*. https://psidonline.isr.umich.edu
- **The Household, Income and Labour Dynamics in Australia (HILDA) Survey**, managed by the *Melbourne Institute of Applied Economic and Social Research* at the *University of Melbourne*, and funded by the *Australian Government Department of Social Services (DSS)*. Data available via *ADA Dataverse*. https://dataverse.ada.edu.au/dataverse/ncld
- **Korean Labor and Income Panel Study (KLIPS)**. Conducted by the *Korea Labor Institute*, funded by the *Ministry of Employment and Labor*. https://www.kli.re.kr/klips_eng
- **Swiss Household Panel (SHP)**, developed by the *Swiss Centre of Expertise in the Social Sciences (FORS)* and supported by the *Swiss National Science Foundation*. https://forscenter.ch/projects/swiss-household-panel
- **Longitudinal Internet Studies for the Social Sciences (LISS)**. Managed by *Centerdata*, an independent research institute located at *Tilburg University*, Netherlands. Based on a true probability sample drawn by *Statistics Netherlands (CBS)*. Funded initially by the *Netherlands Organization for Scientific Research (NWO)* through the MESS project. Data available via the *LISS Data Archive*. https://www.lissdata.nl
- **The Cross-National Equivalent File (CNEF)**, developed and distributed by *The Ohio State University*, with funding from 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, 1-R03HD100924-01). https://www.cnef.ehe.osu.edu

The CPF team would also like to thank the colleagues who supported the development of CPF over the years, in particular (in alphabetical order) Eldad Davidov, Dina Maskileyson, Aleja Rodríguez, Katya Sytkina, Daniel van Wijk, and Gordey Yastrebov.

References

Allanson, P. F. (2011). On the characterization and economic evaluation of income mobility as a process of distributional change. *The Journal of Economic Inequality*, *10*(4), 505-528. https://doi.org/10.1007/s10888-011-9172-5

Büchel, F., & Frick, J. R. (2004). Immigrants in the UK and in West Germany ?Relative income position, income portfolio, and redistribution effects. *Population Economics*, *17*(3). https://doi.org/10.1007/s00148-004-0183-4

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

Burkhauser, R. V., Butrica, B. A., Daly, M. C., & Lillard, D. R. (2001). The Cross-National Equivalent File: A product of cross-national research. In I. Becker, N. Ott, & G. Rolf (Eds.), *Social Insurance in a Dynamic Society*. Campus Fachbuch.

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*(1), 75-100. https://doi.org/10.1111/j.1475-4991.2008.00307.x

Cho, J., & 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*(2), 179-195. https://doi.org/10.1007/s11482-013-9237-7

Cooke, T. J., Boyle, P., Couch, K., & Feijten, P. (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., & McManus, P. (1996). Institutions, Technical Change, and Diverging Life Chances: Earnings Mobility in the United States and Germany. *American Journal of Sociology*, *102*(1), 34-79.

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. https://doi.org/10.1016/j.rssm.2012.11.001

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 : Zeitschrift für Wirtschafts- und Sozialwissenschaften*, *127*, 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

Goebel, J., Grabka, M. M., Liebig, S., Kroh, M., Richter, D., Schröder, C., & Schupp, J. (2019). The German Socio-Economic Panel (SOEP). *Journal of Economics and Statistics*, *239*(2), 345-360. https://doi.org/10.1515/jbnst-2018-0022

Johnson, D., McGonagle, K., Freedman, V., & Sastry, N. (2018). Fifty Years of the Panel Study of Income Dynamics: Past, Present, and Future. *Ann Am Acad Pol Soc Sci*, *680*(1), 9-28. https://doi.org/10.1177/0002716218809363

Kaminska, O., & Lynn, P. (2017). Survey-Based Cross-Country Comparisons Where Countries Vary in Sample Design: Issues and Solutions. *Journal of Official Statistics*, *33*(1), 123-136. https://doi.org/10.1515/jos-2017-0007

McCall, L., & Percheski, C. (2010). Income Inequality: New Trends and Research Directions. *Annual Review of Sociology*, *36*(1), 329-347. https://doi.org/10.1146/annurev.soc.012809.102541

McGonagle, K. A., Schoeni, R. F., Sastry, N., & Freedman, V. A. (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: Familes, Schools and Persisting Stratification* (pp. 171-200). https://doi.org/10.1016/s1479-3539(03)14008-6

Musick, K., Bea, M. D., & Gonalons-Pons, P. (2020). His and Her Earnings Following Parenthood in the United States, Germany, and the United Kingdom. *American Sociological Review*, *85*(4), 639-674. https://doi.org/10.1177/0003122420934430

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

Siegers, R., Belcheva, V., & 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., & Tomescu-Dubrow, I. (2019). Basic Principles of Survey Data Recycling. In T. P. Johnson, B.-E. Pennell, I. A. L. Stoop, & B. Dorer (Eds.), *Advances in Comparative Survey Methods: Multinational, Multiregional, and Multicultural Contexts (3MC),* . John Wiley & Sons.

Turek, K. (2025). Accelerating social science knowledge production with the coordinated open-source model. *Quality & Quantity*, *59*(S2), 767-795. https://doi.org/10.1007/s11135-024-02020-7

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

Watson, N., & Wooden, M. (2020). The Household, Income and Labour Dynamics in Australia (HILDA) Survey. *Journal of Economics and Statistics*, *Published online*. https://doi.org/10.1515/jbnst-2020-0029

Wolf, C., Joye, D., Smith, T., & Fu, Y.-c. (2017). Harmonizing Survey Questions Between Cultures and Over Time. In C. Wolf, Y.-c. Fu, D. Joye, & T. Smith (Eds.), *The SAGE Handbook of Survey Methodology*. SAGE Publications. https://doi.org/10.4135/9781473957893