gesis Leibniz Institute for the Social Sciences

Data Manual

GESIS Panel Data Manual

Jan-Philipp Kolb and Kai Weyandt

December 2018

Content

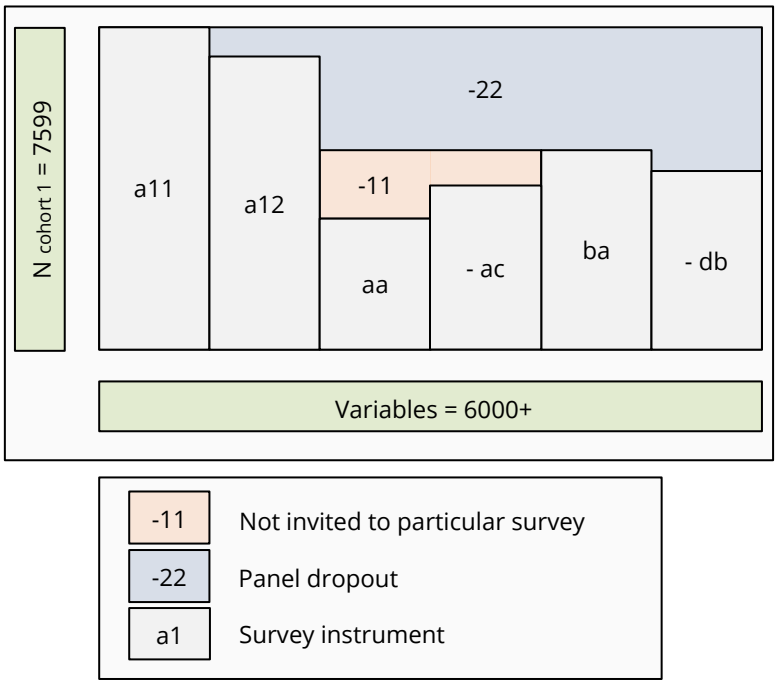
About this document

The GESIS Panel data manual is a supplementary document to the other documentation such as wave reports, study descriptions, and the codebook and refers to each of these documents if necessary.

1. Data structure

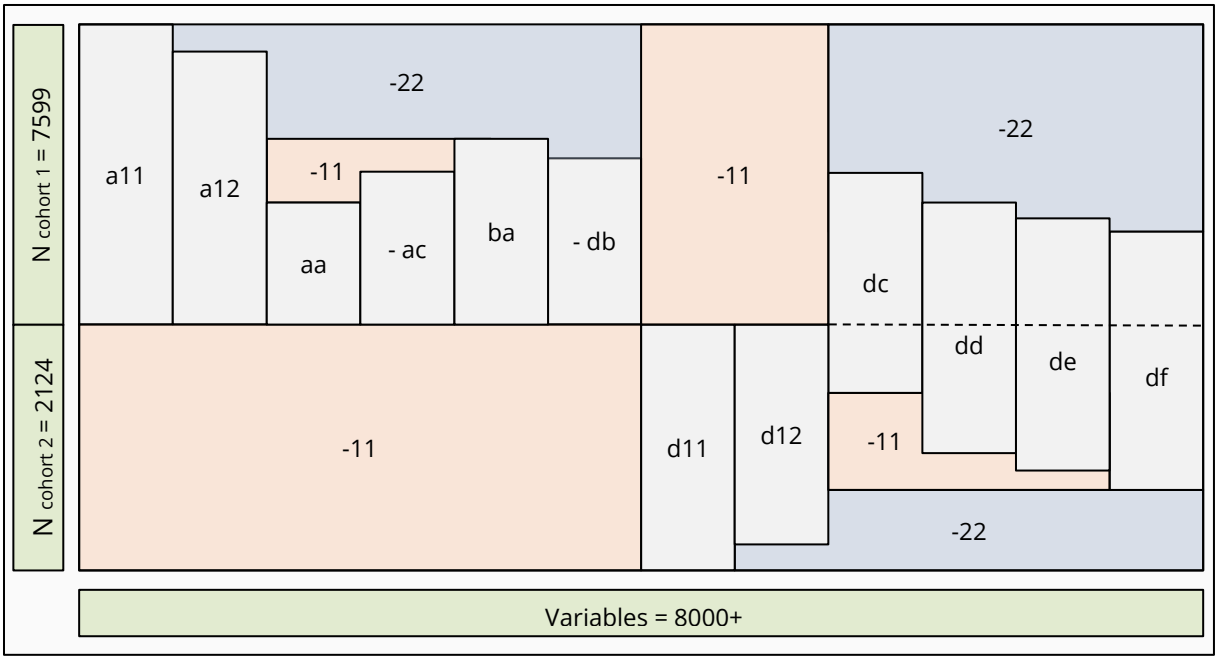
Up until the year 2018, the GESIS Panel disseminated the data including all variables within a single dataset (Figure 1). Due to the rapid growth in the variable count and the integration of the first refreshment cohort („cohort 2“) recruited in 2016, the dissemination package has been modified.

Figure 1: Data dissemination up until 2018



With the integration of the second cohort 2016, the dataset is delivered separated by cohorts and years. Figure 2 illustrates the structure of the continuation of a single master dataset. Each bar represents a survey instrument (e.g., the recruitment interview 2013 a11 or wave ba). On the vertical axis, the observations per cohort are mapped. Since version 27-0-0 cohort specific master datasets are again part of the dissemenation package.

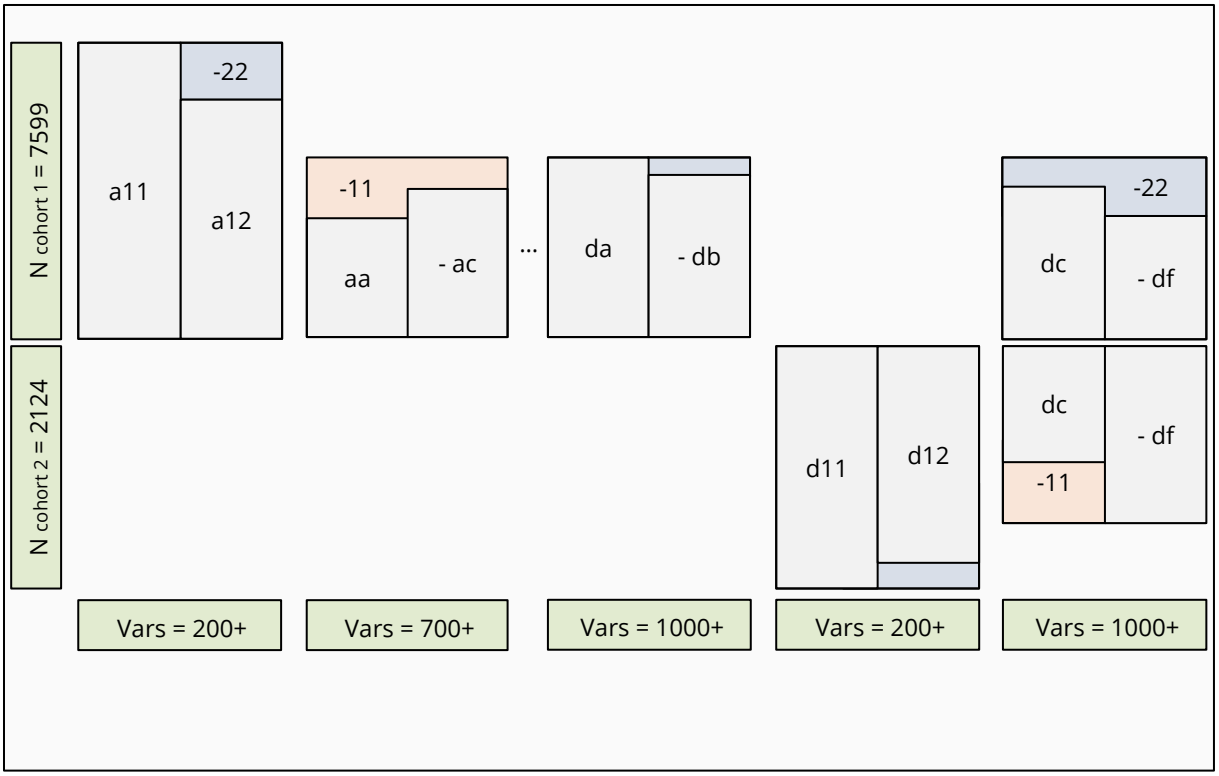
Figure 2: Continuation of the single master dataset[1]



Since the dataset concept as pictured in Figure 2 would lead to a high amount of non-substantial data for new recruitment cohorts below the data frame of the initial cohort (Recruitment 2013: a11-a12; Waves 2013-2016: aa-db), the division into cohort specific master files as well as year-specific recruitment and wave datasets is realized. The initial rationale for dividing the dataset is illustrated in Figure 3. The location of a variable can be deduced by the first two digits of a variable name, for instance, bczd002a can be found in the Stata dataset „ZA5665_a1_ba-bf_v22-0-0.dta". In 2016, the refreshment cohort in cooperation with the General Social Survey (ALLBUS) was conducted and was integrated into the dataset with the first data publication in 2018. Variables collected after wave db (second wave in 2016) are located in four datasets, the respective master file and the ones for each cohort. As an example dczd005a is included in the datasets „ZA5665_a1_v27-0-0.dta" „ZA5665_a1_da-df_v22-0-0.dta" and „ZA5665_d1_v27-0-0.dta" and „ZA5665_d1_dc-df_v27-0-0.dta". Also, the respective dataset name(s) has been added to the most recent version of the codebook. That is, searching for variable dczd005a now reveals the dataset(s), where this variable can be found.

[1] This structure can be produced with the Stata Do-File "ZA5664-65_merge-and-append-files.do" that is part of the data dissemination package.

Figure 3: Fragmentation of the dataset



1.1. Types of datasets

Basically five types of datasets exist: since version 27-0-0 cohort specific master files (0) are again part of the dissemination package. The next two kinds are (1) Recruitment instruments and (2) Regular Waves which are cohort specific as well, the latter two (3) String Data (4) Client-Side-Paradata (study zs) do not differentiate between cohorts. Here are some examples:

	Description	ZA-Number		Cohort		Content2		Publication		File type
(0)	Cohort specific Master of initial recruitment (a1)	ZA5665	_	a1			_	v27-0-0	.	dta
(1)	Recr. Int. and profile survey 2016 (d1)	ZA5665	_	d1	_	d11-d12	_	v22-0-0	.	dta
(2)	Wave da-df data of initial recruitment (a1)	ZA5665	_	a1	_	da-df	_	v22-0-0	.	sav
(3)	String var. with more than 25 characters	ZA5665					_	v22-0-0	.	csv

[2] Either all data collection events recruitment waves (e.g. a11-a12) or regular panel waves (e.g. wave ba-bf)

	Description	ZA-Number	Cohort	Content2	Publication	File type
(4)	Wave dc specific dataset of study zs	ZA5664-65	_	dczs	_	. csv

With the current structure of the dissemination package, there are two ways of receiving a master data set that includes all variables of all waves and all cohorts:

1. Combining the two cohort-specific master files by appending them
2. Combining the year- and cohort-specific data sets by the means of the provided Stata do-file za5664-65_merge-and-append-files.do

1.2. String variables

The data are provided as Stata (Version 12) and SPSS datasets. Due to restrictions of Stata 12 string variables with more than 25 characters are outsourced to a comma-separated dataset („ZA5664-65_v22-0-0.csv"). The concerning variables in the regular datasets indicate the filename by its value label. This is especially the case for some server-side paradata (zp) as well as the client-side paradata (zs). Latter are outsourced into a separate folder with datasets for each wave (e.g. „ZA5664-65_dczs.csv"), because of the size of the raw data. More Information on the Study zs can be found in the following document: https://dbk.gesis.org/dbksearch/download.asp?db=D&id=53571

Note: Other than the other numeric and short string variables, the outsourced string datasets are not separated for the different recruitment cohorts.

1.3. Demography

Basic demographic variables are located in the datasets of the respective recruitment (a11, a12 and d11, d12). Apart from that, yearly updates are made in fifth (e-waves: e.g., be) and sixth wave (f-waves: e.g., bf) of each year. Details on the variables can be found in the study description document and the codebook.

2. Design Weights

The GESIS panel is a probability-based self-administered panel. The first cohort of the GESIS Panel (GP13) was recruited in 2013. Panel mortality requires regular refreshment. Therefore, the realization of further cohorts is planned. The integration of the second cohort of the GESIS Panel (GP16) requires the calculation design weights because the inclusion probabilities for the respondents in the GESIS Panel differ. For this reason, every observation i in the dataset must receive a design weight, even if the sample design of the first cohort (GP13) was initially a self-weighting design. The second cohort was recruited based on the ALLBUS, where an oversampling for East Germany was implemented.

The first and the second cohort have a different age range. In the first cohort, persons older than 17 and younger than 71 could be part of the panel in 2013. The sample for the second cohort in 2016 does not have an upper age limit.

2.1. The calculation of the inclusion probabilities

The calculation of inclusion probabilities requires the clarification of the different options a respondent can enter the GESIS Panel. A person can access the GESIS Panel via the first cohort from 2013 and the second cohort in 2016. All panelists in the first cohort had the same inclusion probability π_i^{GP13} whereas for the second cohort, individuals from East Germany had a higher chance than those from West Germany to join the GESIS Panel. If we exclude the possibility of a person coming into the GESIS Panel twice, we have to substract the likelihood that a person is drawn into the first and second cohort at the same time.

This difference in the age limit means that people born before 1943 or after 1995 are underrepresented in the GESIS Panel. We can use the following formula to calculate the inclusion probability for each panel member i:

$$\pi_i = (\pi_{i,k}^{\mathrm{GP16}} + \pi_i^{\mathrm{GP13}}) - (\pi_{i,k}^{\mathrm{GP16}} \cdot \pi_i^{GP13})$$

For example, $\pi_{i,k}^{\mathrm{GP16}}$ is the inclusion probability for the second cohort. Due to oversampling, these inclusion probabilities must be calculated separately for East and West ($k \in \{\mathrm{East}, \mathrm{West}\}$). A person in the groups described above (that is, relatively young and older persons) can only have entered the GESIS Panel through the second cohort.

The general inclusion probability for a GESIS Panel respondent is obtained by dividing the size of the gross sample, for example, n^{GP13}, by the proportion of the respective target population N^{GP13}.

$$\pi_i^{\mathrm{GP13}} = \frac{n^{\mathrm{GP13}}}{N^{\mathrm{GP13}}}$$

For the ALLBUS population, we have to consider that the inclusion probability differs between East and West. We can compute the inclusion probability for the refreshment from the following formula:

$$\pi_{i,k}^{\mathrm{GP16}} = \frac{n_k^{\mathrm{GP16}}}{N_k^{\mathrm{GP16}}}$$

The k indicates whether a person is resident in East or West. Persons residing in Germany aged 18-70 years in 2013 were part of the target population of the original GESIS Panel base sample. That was 56,369,000 people. The target population for the refreshment sample in 2016 was determined by the microcensus 2015. This step became necessary because the data for 2016 were not available at the time of the implementation of the second cohort. In this case, persons are part of the target population that are German-speaking and older than 18.

$$N^A = N^A_{West} + N^A_{East} = 55{,}586{,}000 + 1{,}204{,}000 = 67{,}626{,}000$$

The gross sample of the GESIS Panel included $n^{GP} = 21{,}870$ persons. Using this information we can compute the four inclusion probabilities in following table.

Table 1 Inclusion probabilities for the GESIS Panel

Probability	Value
$\pi^{\text{GP16}}_{\text{West}}$	0.0001318
π_{West}	0.0005197
$\pi^{\text{GP16}}_{\text{East}}$	0.0002796
π_{East}	0.0006674

For example, $\pi^{\text{GP16}}_{\text{West}}$ is the inclusion probability of people living in the West who, due to their age, can only have reached the GESIS Panel via the ALLBUS refresher sample. A person that lives in the West who could get into the GESIS Panel via both samples has the inclusion probability π_{West}.

2.2. The calculation of design weights

The corresponding design weights can be calculated by inverting the inclusion probability, for example:

$$d = \frac{1}{\pi^{\text{GP16}}_{\text{West}}}$$

In social science surveys, it is common practice to normalize this weighting to the sample size. The sum of the weights should be equal to the sample size:

$$\sum d = n^{GP13} + n^{\text{GP16}}$$

The general transformation factor for the design weights is:

$$d_i^* = (n^{\text{GP16}} + n^{\text{GP13}}) \cdot \frac{d_i}{\sum_{n=1}^{n} d_i}$$

This normalized weight in the delivered dataset of the GESIS Panel is provided in the form of variable z000011a. The following values result:

Table 1 Inclusion probabilities and design weights for the GESIS Panel

Group	Probability	Weight	Normalized Weight	Number of cases
{Only GP16, West}	0.0001318	7587.497	3.6153640	155
{Both cohorts, West}	0.0005197	1924.099	0.9168132	7593
{Only GP16, East}	0.0002796	3576.946	1.7043780	87
{Both cohorts, East}	0.0006674	1498.265	0.7139077	1888

If the second cohort is not part of the analysis, it is not necessary to use the design weights for an analysis. If data from the second or following cohorts is used, the design weights must be taken into account.