gesis Leibniz Institute for the Social Sciences

GESIS Panel Data Manual

GESIS Panel Data Manual

related to ZA5664 and ZA5665

Jan-Philipp Kolb, Christian Bruch, Matthias Sand, Ingo Konradt, Bernd Weiß and Kai Weyandt

May 2022

Content

1. About this document

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

2. 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.

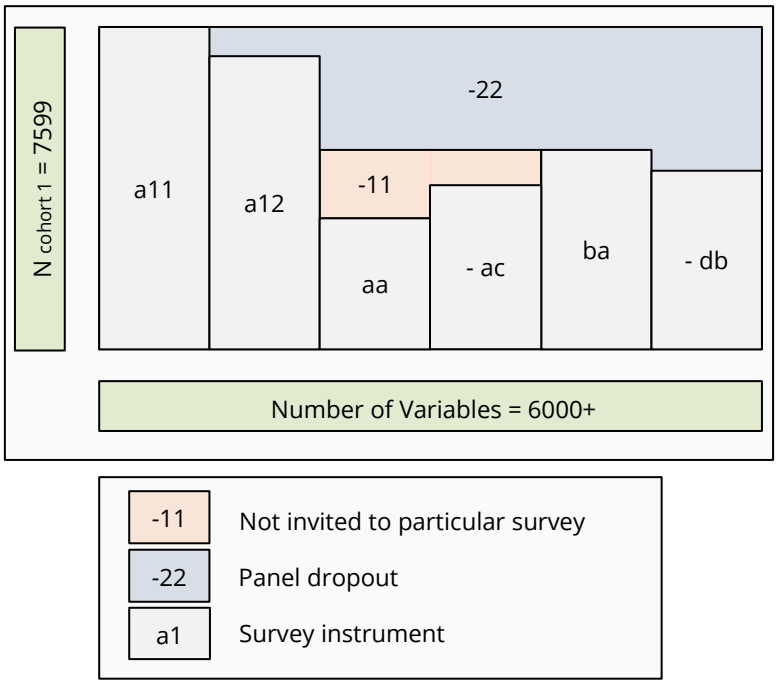

Figure 1: Data dissemination up until 2018

With the integration of the second cohort in 2016 and the third cohort in 2018, 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 master datasets are again part of the dissemenation package.

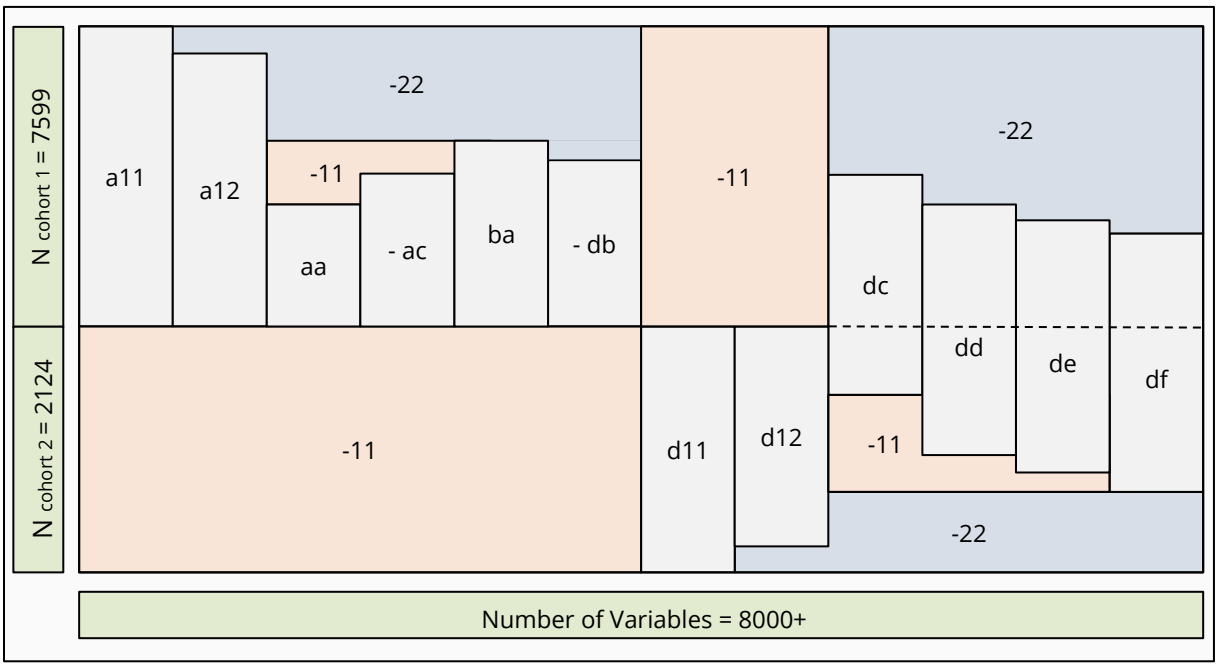

Figure 2: Continuation of the single master dataset exemplary for cohort 2013 and 2016[1]

Since the new dataset concept picturedin Figure 2 would lead to a high amount of non-substantial data for new recruitment cohorts (depicted below the data frame of the initial cohort, a11-a12; 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 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.

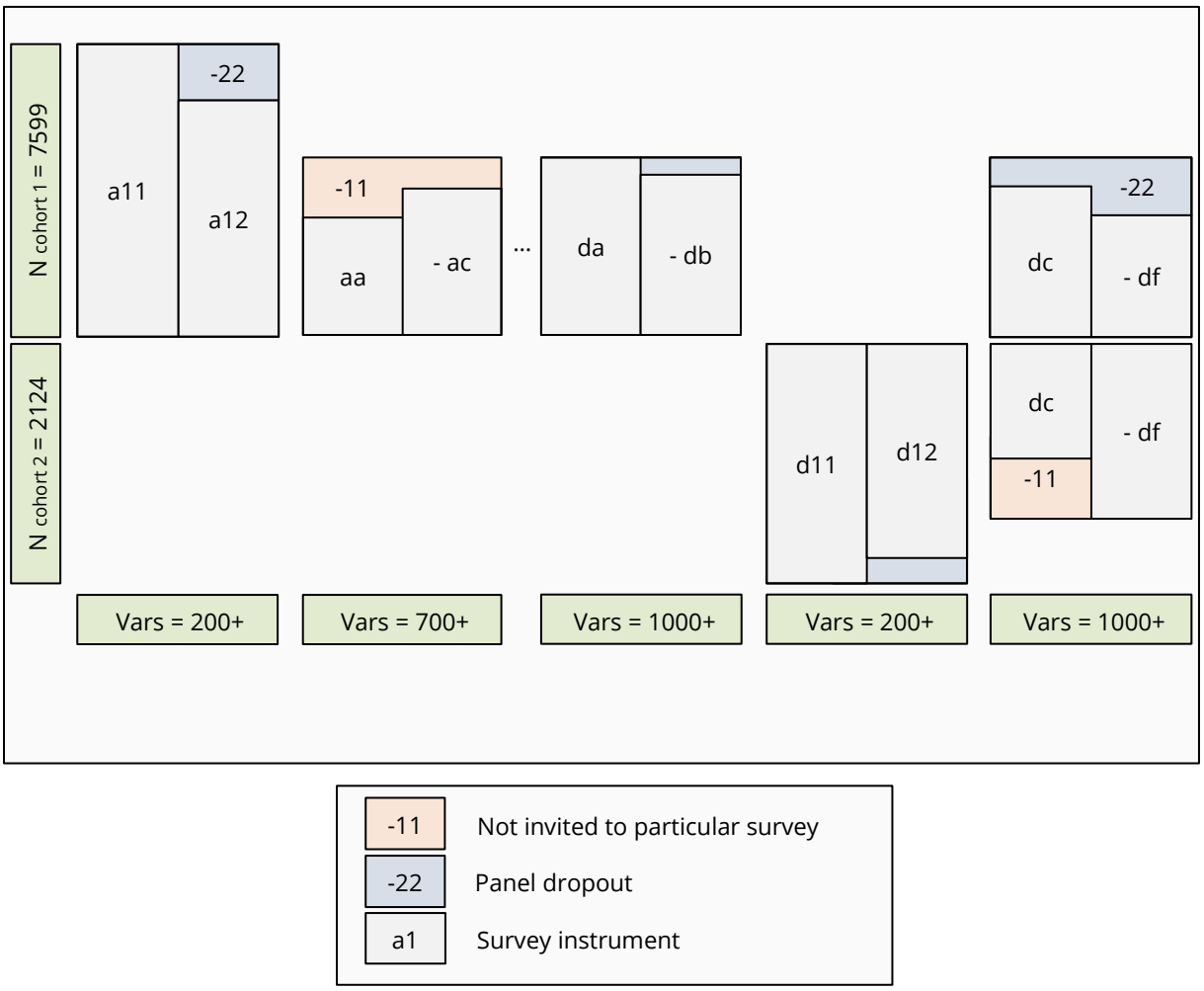

Figure 3: Fragmentation of the dataset

2.1. Types of datasets

Basically, six 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. From version 43-0-0 on the longitudinal (5) demographic dataset is part of the dissemination package and is described in more detail below. Table 1 lists some examples.

Table 1: Overview of existing data sets exemplified for the GESIS Panel Standard Edition ZA5665 of version 43-0-0

	Description	ZA-Number		Cohort		Content[1]		Version number		File type
(0)	Cohort specific Master of initial recruitment (a1)	ZA5665	_	a1			_	v43-0-0	.	dta
(1)	Recr. Int. and profile survey 2016 (d1)	ZA5665	_	d1	_	d11-d12	_	v43-0-0	.	dta
(2)	Wave da-df data of initial recruitment (a1)	ZA5665	_	a1	_	da-df	_	v43-0-0	.	sav
(3)	String var. with more than 25 characters	ZA5665					_	v43-0-0	.	csv
(4)	Wave dc specific dataset of study zs	ZA5664-65	_			dczs	_	v43-0-0	.	csv
(5)	Demographic dataset	ZA5665	_			demography		v43-0-0	.	dta

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 three 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

2.2. 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 (e.g., „ZA5664-65_v22-0-0.csv"). The respective variables in the datasets indicate contain the filename as 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_v43-0-0.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?id=65301

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

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

2.3. Demographic variables and demographic dataset

Basic demographic variables are in the cross-sectional datasets of the respective recruitment (a11, a12 and d11, d12). Apart from that, yearly updates are made within the last or second-last wave of each year. Details on the variables can be found in the study description document and the codebook.

The demographic dataset helps data users to find and use socio-demographic information from the GESIS Panel across all survey waves. The dataset will be updated regularly. For the demographic dataset, certain variables are extracted from the individual waves and compiled to a longitudinal format.

Overview of variable naming scheme and auxiliary variables

All variables have been given mnemonic longitudinal variable names.[2] In addition, there are the following auxiliary variables:

Orig-variables („*_orig“) refer to the original variable names on which the measurement in the cross-section is based.

Flag-variables („*_flag“) indicate dichotomously (0/1) for certain, supposedly time-invariant variables (e.g., year of birth or sex) whether the information has changed (noticeably) over time. This includes variables such as year of birth (yob) or a lower highest level of education (hle) in a follow-up survey, which cannot change, but also variables such as sex, which only very rarely change. Depending on the characteristic of the variable, this may indicate an inconsistency or represent an actual change or subsequent correction. No adjustments or standardizations were made here. The corresponding cases can be checked for plausibility or excluded in further analyses if necessary.

Longitudinal data structure (long format)

For each respondent, information is included in the longitudinal dataset if he or she received an invitation for the corresponding survey wave[3]. Whether the survey was completed or why this was not the case is indicated by the AAPOR disposition code[4] in the dataset.

Updating and dealing with missing values

The socio-demographic longitudinal data reflect the last available information by the respondents.[5] However, this can sometimes lead to outdated information being used in subsequent waves. In this context, the *_orig variables, which indicate the wave in which the characteristic was last collected, must always be considered.

Cases in which no information was initially available are generally marked as a system missing code. The reason for the item-level non-response can be found in the cross-sectional data, only.

---

[2] Except for the ID variable z000001a for linking to the GESIS Panel Standard or Extended Edition (ZA5665/ZA5664).

[3] The variable aapor does not vary for the interviewer-administered recruitment interviews since only completed interviews are part of the dataset and were generated by default as completed. In the cross-sectional dataset, they do not exist.

[4] The GESIS Panel uses disposition codes based on the standard definitions of the American Association for Public Opinion Research (AAPOR), extended by some own survey-specific codes (see also Stadtmüller et al. 2019, p.15f).

[5] The values are not written back to earlier waves.

Harmonization over time:
In a panel survey, it is always possible that questionnaire items (especially the wording and response categories) may change over time. This must be considered in the longitudinal data processing. In the demographic dataset, the variables affected by such changes are sex, the highest level of education, and the personal and household net income. The GESIS Panel codebook (ZA5664-65_cdb.pdf) and questionnaires (ZA5664-65_fb.pdf) must be used to understand precisely which variables and waves are affected and what the items look like in detail. In the codebook for the demographic dataset (ZA5664-65_cdb_demography.xlsx), the corresponding harmonized variables are marked with an asterisk (*) in the "Measurement differences" row. In the following, the harmonization steps of these variables are explained in more detail:

Sex/gender of respondent (sex)

The measurement of sex or gender differs twofold, at the level of the survey mode (interview-based vs. self-administered) as well as the survey response (biological sex vs. gender). While biological sex is assessed by the interviewer during the recruitment process, in the regular waves the classification is given by the respondent themselves. From wave hf onwards, the questionnaire has also changed in that the answer options "Diverse" and "No entry in civil status register" have been added.[6]

Highest level of education (hle)

The variable for the highest level of education was surveyed differently over time, in particular the number and composition of the available response categories. For the demographic dataset they were combined into a standardized response scale (see also ZA5664-65_cdb_demography.xlsx).

Personal net income (pinc) / Household net income (hhinc)

The personal and household net income in the demographic dataset includes harmonized income data from the recruitment and refreshment surveys as well as from the core study demographic updates in the f-waves. While the recruitment survey and demographic updates differ only in the number of available income categories, which can be easily combined, there are income categories in the refreshment surveys that do not overlap with those from the core study in the f-waves. Therefore, a new derived variable was created that converts the information from the refreshment surveys into a shared classification according to the mean income of the category in question (middle of category). For the cases integrated according to the middle-of-category-approach, the respondents who selected such a category in the refreshment survey are assigned to the new category with a certain bias. If these cases should not be considered in longitudinal analyses, they can be excluded accordingly. For cross-sectional analyses, we recommend using the non-harmonized income data from the original variables[7].

How are the data linked with each other?
Here is an example for cohort 1 and the wave ge from the GESIS Panel Standard Edition in version 41.0.0 (ZA5665):

a. Starting from a cohort-specific master file (Stata).

```
. use ".../ZA5665_a1_ga-gf_v41-0-0.dta", clear
. keep z* ge*
. gen wave = "ge"
```

[6] For data protection reasons, this information is only available in the GESIS Panel Extended Edition (ZA5664).
[7] The income categories in the refreshment sample are based on the measurement from the ALLBUS survey.

```
. merge 1:1 z000001a wave using ".../ZA5665_demography_v42-0-0.dta",
keep(master match)
```

b. Starting from the demographic dataset

```
. use ".../ZA5665_demography_v42-0-0.dta", clear
. keep if wave == "ge"
. merge 1:1 z000001a using ".../ZA5665_a1_ga-gf_v41-0-0.dta", keep(master
match) keepusing(z* ge*)
```

3. 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) and the third cohort of the GESIS Panel (GP18) requires the calculation of design weights because the inclusion probabilities for the respondents in the GESIS Panel differ due to these refreshments. 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 and third cohort were recruited based on the ALLBUS, where an oversampling for East Germany was implemented.

The different cohorts have a different age range. In the first cohort, only persons older than 17 and younger than 71 could be part of the panel in 2013 (born between 1942 and 1995). The samples for the subsequent refreshment cohorts do not have an upper age limit.

3.1. The calculation of the inclusion probabilities

The calculation of inclusion probabilities requires the determination of the different options a respondent may enter the GESIS Panel. Most persons can access the GESIS Panel via the first cohort from 2013, the second cohort in 2016 or through the third cohort in 2018. However, depending on the age of a panelist, persons of a certain age may only stem from the second and/or third cohort. All panelists in the first cohort had the same inclusion probability π_i^{GP13} whereas for the second and third 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 more than once, we need to correct for the likelihood that a person has been sampled by two or more cohorts at the same time.

The inclusion probability for a GESIS Panel respondent to be selected in 2013 is obtained by dividing the size of the gross sample, for example, n^{GP13}, by the proportion of the respective target population N^{GP13}.

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

N^{GP13} takes the value 55.948.331 and n^{GP13} the value 21.870.

For the ALLBUS population, we must consider that the inclusion probability differs between East and West Germany due to the disproportional allocation. We can compute the inclusion probability to be drawn for the refreshment in 2016 from the formula

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

and the inclusion probability for the refreshment in 2018 from the formula

$$\pi_{i,k}^{\mathrm{GP18}} = \frac{n_k^{\mathrm{GP18}}}{N_k^{\mathrm{GP18}}}$$

The k indicates whether a person is resident in East or West Germany ($k \in \{\mathrm{East}, \mathrm{West}\}$).

The values in 2016 are N_{East}^{GP16} = 11.856.132, N_{West}^{GP16} = 55.867.847, n_{East}^{GP16} = 3.366 and n_{West}^{GP16} = 7.326. The values in 2018 are N_{East}^{GP18} = 11.896.481, N_{West}^{GP18} = 56.953.526, n_{East}^{GP18} = 3.621 and n_{West}^{GP18} = 7.881.

When including all three samples (cohorts) into one dataset the inclusion probabilities must be adjusted in the sense of a multiple frame approach. Furthermore, it must be considered that, due to age limits, not all persons can be part of every sample. In case of the GESIS Panel, an upper limit on age has been set that is not present in the refreshment cohorts based on the ALLBUS. This difference in the age limit means that people born before December 1942 or after December 1995 are underrepresented in the GESIS Panel. Due to the lower age limit in the GESIS Panel and the ALLBUS, people that are younger than 18 at the survey time cannot participate. Thus, three different age groups must be considered, including persons that can be selected for the different cohorts in a varying extent and thus, inclusion probabilities must be computed for these age classes in a different way.

The first group include panel members that were born between December 1942 and the end of November 1995. Such persons can be selected theoretically for each cohort. Thus, the inclusion probability for a panel member i of this age class in the presence of three cohorts is computed by:

$$\pi_{i,k} = (\pi_{i,k}^{\mathrm{GP18}} + \pi_{i,k}^{\mathrm{GP16}} + \pi_{i}^{\mathrm{GP13}}) - (\pi_{i,k}^{\mathrm{GP16}} \cdot \pi_{i}^{GP13}) - (\pi_{i,k}^{\mathrm{GP18}} \cdot \pi_{i}^{GP13}) - (\pi_{i,k}^{\mathrm{GP18}} \cdot \pi_{i,k}^{GP16}) + (\pi_{i,k}^{\mathrm{GP18}} \cdot \pi_{i,k}^{\mathrm{GP16}} \cdot \pi_{i}^{GP13})$$

Due to oversampling, these inclusion probabilities must be calculated separately for East and West ($k \in \{\mathrm{East}, \mathrm{West}\}$).

A person born between December 1995 and the end of November 1998 can only be selected by the second or the third cohort. They cannot be part of the first cohort. Thus, the inclusion probability for one of these panel members i is computed by

$$\pi_{i,k} = (\pi_{i,k}^{\mathrm{GP16}} + \pi_{i,k}^{\mathrm{GP18}}) - (\pi_{i,k}^{\mathrm{GP16}} \cdot \pi_{i,k}^{\mathrm{GP18}})$$

A person born after November 1998, who is at least 18 during the survey of the third cohort can only be included via the third cohort. Thus, their inclusion probability remains $\pi_{i,k}^{\mathrm{GP18}}$. In the GESIS Panel dataset, the inclusion probabilities are provided by the variables z000012a and z000012b.

3.2. The calculation of design weights

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

$$d_{i,k} = \frac{1}{\pi_{i,k}}$$

This design weights in the delivered dataset of the GESIS Panel is provided in the form of variables z000010a und z000010b.

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

$$\sum \sum d_{i,k} = {n_{net}}^{\mathrm{GP13}} + {n_{net}}^{\mathrm{GP16}} + {n_{net}}^{\mathrm{GP18}} = n_{net}$$

The general transformation factor for the design weights is:

$$d_{i,k}^{*} = n_{net} \cdot \frac{d_{i,k}}{\sum_{i=1}^{n_{net}} \sum_{k \in \mathrm{East,West}} d_{i,k}}$$

These standardized weights in the delivered dataset of the GESIS Panel will be provided in the form of wave and instrument specific variables. The standardization of weights leads to weights that have a mean of 1 and a total sum equal to the net sample size, which may be beneficiary to determine whether a particular person has a lower than average (weight above 1) or a higher than average (weight below 1) probability of inclusion. Due to the by-wave variation of the net sample size, that standardization needs to be conducted each wave.

If only the first cohort is part of the analysis, it is not necessary to use the design weights for an analysis. If data from the second and/or following cohorts is added, the design weights must be considered.

3.3. How and when to use which design weights

Generally, when one wants to estimate inferences based on the latest published wave of the GESIS Panel, we recommend using the latest (standardized) corresponding design weight. When analyzing older waves of the GESIS Panel, the design weights published for these waves should be used.

As already described, the wave specific weights are standardized on a wave's current sample size[8]. Hence, the sum of those weights is equal to the number of persons that have answered the current wave's questionnaire. The unstandardized weights, hence, z000010a and z000010b, provide an estimation of the size of the (current) target population. Therefore, their sum would lead to

$$\sum d_{i,k} = \hat{N}.$$

If one wants to estimate (basic) inferences of the survey sample, we recommend using the standardized weights. The same applies for the estimation of (relative) distributions of a survey variable. Due to the properties of the standardized weights, a relative distribution could easily be derived via the aggregation of weights by category and the division by sample size.

If one wants to estimate inferences for the target population (absolute values), the unstandardized weights should be used. Relative distributions can be derived via aggregation of these weights by category followed by the division of the sum of unstandardized weights. However, these should be identical to the previously described relative distributions[8].

Another point of importance is the usage of different weights and therefore the different weights that may result. As already described, the weights will be standardized by wave, due to varying sample sizes. In that case we would recommend to either use the unstandardized weights that remain constant up to the next refreshment of cohorts or take notice of the variation of each standardization (e.g., by aggregating the weighted data by wave).

The additional refreshment cohorts may however further impact the analysis of "older" cohorts. As shown in the previous section, adding a new cohort would automatically lead to a change in inclusion probabilities of the already existing panelists. That can be explained by the fact that these panelists also have a (theoretical) probability to be sampled by the same sample, the most recent refreshment cohort stems from. Their probability of being part of the survey therefore increases and in turn, their design weight decreases. Using new weights, when analyzing data from previous waves that did not contain a particular refreshment cohort would therefore lead to biased estimates. The correct usage of weights is exemplarily displayed in Table 2.

[8] Beware for wave hz the sample is limited to the online participants and does not include the offline participants.

Table 2: Usage of design weights of the GESIS panel

ID	Cohort	Weight 1	Weight 2	Weight 3	…
1	Base	a	${d_{i,k}}^{b}$	${d_{i,k}}^{c}$	
2	Base	a	${d_{i,k}}^{b}$	${d_{i,k}}^{c}$	
3	Base	a	${d_{i,k}}^{b}$	${d_{i,k}}^{c}$	
4	Base	a	${d_{i,k}}^{b}$	${d_{i,k}}^{c}$	
5	Base + 1st refreshment	NA	${d_{i,k}}^{b}$	${d_{i,k}}^{c}$	
6	Base + 1st refreshment	NA	${d_{i,k}}^{b}$	${d_{i,k}}^{c}$	
7	Base + 1st refreshment	NA	${d_{i,k}}^{b}$	${d_{i,k}}^{c}$	
8	Base + 1st refreshment +2nd refreshment	NA	${d_{i,k}}^{b}$	${d_{i,k}}^{c}$	
9	Base + 1st refreshment +2nd refreshment	NA	NA	${d_{i,k}}^{c}$	
10	Base + 1st refreshment +2nd refreshment	NA	NA	${d_{i,k}}^{c}$	
…	…	NA	NA	(NA)	…

If one wants to analyze the data from the most recent wave (including all cohorts), weights ${d_{i,k}}^{c}$ should be employed. If analyzing data from a previous wave, that was conducted before the second refreshment was added, weights ${d_{i,k}}^{b}$ should be used and if one wants to analyze data that has been gathered before any of the refreshments occurred, weights a should be used. Due to the self-weighted design of the base cohort, a is a constant that can be calculated by the population size in 2013 divided by the sample size of the recruitment survey.

Design weights of the latest waves are computed by including all cohorts. In case, some of the cohorts should not be included in the analyses, the latest design weights of persons in the remaining cohorts cannot be used and need to be altered. By doing so, different design weights result dependent on the included cohorts. Since this may result in a large number of resulted design weights and due to reasons of clarity, only the design weights of all cohorts are considered and published. In case one needs design weights for special cohort combinations please contact info@gesis-panel.org.

References

Stadtmüller, S., Silber, H., Daikeler, J., Martin, S., Sand, M., Schmich, P., Schröder, J., Struminskaya, B., Weyandt, K. W., & Zabal, A. (2019). Adaptation of the AAPOR Final Disposition Codes for the German Survey Context. Mannheim, GESIS – Leibniz-Institute for the Social Sciences (GESIS – Survey Guidelines). doi: https://doi.org/10.15465/gesis-sg_en_026

Hippel, P. T. von, Scarpino, S. V., & Holas, I. (2016). Robust Estimation of Inequality from Binned Incomes. *Sociological Methodology*, *46*(1), 212–251. https://doi.org/10.1177/0081175015599807

Jargowsky, P. A., & Wheeler, C. A. (2018). Estimating Income Statistics from Grouped Data: Mean-constrained Integration over Brackets. *Sociological Methodology*, *48*(1), 337–374. https://doi.org/10.1177/0081175018782579

Rizzi, S., Gampe, J., & Eilers, P. H. C. (2015). Efficient estimation of smooth distributions from coarsely grouped data. *American Journal of Epidemiology*, *182*(2), 138–147. https://doi.org/10.1093/aje/kwv020

GESIS Panel Data Manual
GESIS – Leibniz-Institut für Sozialwissenschaften
Survey Design and Methodology
GESIS Panel
Postfach12 21 55
68072 Mannheim
E-Mail: info@gesis-panel.org
www.gesis.org/gesis-panel
Last update with Wave ia, Version 43-0-0 (2022-05-30)