TwinLife Technical Report Series

No. 11, June 2021

TwinLife ShortGuide

v1.1.0

by Kristina Krell*, Amelie Nikstat*, Anita Kottwitz, Myriam A. Baum, Eike F. Eifler, Anke Hufer-Thamm, Christoph H. Klatzka, Volker Lang, Mirko Ruks, Alexandra Starr, Lena Weigel

kristina.krell@uni-bielefeld.de

*These authors are joint first authors on this work.

UNIVERSITÄT BIELEFELD

UNIVERSITÄT DES SAARLANDES

Kristina Krell*, Amelie Nikstat*, Anita Kottwitz, Myriam A. Baum, Eike F. Eifler, Anke Hufer-Thamm, Christoph H. Klatzka, Volker Lang, Mirko Ruks, Alexandra Starr, Lena Weigel

***TwinLife* ShortGuide v1.1.0**

TwinLife Technical Report Series No. 11
Project TwinLife “Genetic and social causes of life chances”
Bielefeld, June 2021

TwinLife Technical Report Series
General Editors: Martin Diewald, Rainer Riemann and Frank M. Spinath
ISSN 2512-403X

This publication has been funded by the German Research Foundation (DFG).

TwinLife “Genetic and social causes of life chances”
University of Bielefeld
Faculty of Sociology
PO Box 100131
D-33501 Bielefeld
Germany

Phone: +49 (0)521 106-4309
Email: martin.diewald@uni-bielefeld.de
Web: https://www.twin-life.de

ShortGuide v1.1.0

Krell, K.*; Nikstat, A.*; Kottwitz, A.; Baum, M. A.; Eifler, E. F.; Hufer, A.; Klatzka, C.H.; Lang, V.; Mönkediek, B.; Ruks, M.; Starr, A.; Weigel, L.

May 2021

Contact: kristina.krell@uni-bielefeld.de

*These authors are joint first authors on this work.

The ShortGuide is intended to give both an overview of the longitudinal twin family study TwinLife and a short instruction on how to use the TwinLife data.
It corresponds to the contents of the data documentation website of TwinLife (www.twin-life.de/documentation).
The following pages contain information about the project and links to various helpful documents that should facilitate the first steps into working with the TwinLife data. For a very quick description of everything you need to get started with the TwinLife Dataset, see 'getting started' section below.

Table of content

0. Getting started

TwinLife – A genetically informative, longitudinal study about the development of social inequality

Design

- An overview of the theoretical and empirical background, the study design and content as well as the implementation of the study can be found in the TwinLife reference paper (Hahn et al., 2016).

Data

- The TwinLife data are described and archived in the GESIS data catalogue.

- To get access to the TwinLife data, please fill in the Data Use Agreement which you can find under 'Actions' in the GESIS data catalogue. All data sets are available with English and German labels in SPSS and Stata formats. You can also follow the direct link to the Data Use Agreement.

Variables

- In the Codebooks for each data collection (ZA6701_cod_wid$.pdf), you can find a list of the complete set of variables with names, variable and value labels as well as the distributions of frequencies. Furthermore, it documents question texts, filter conditions, and references of all variables. You can find the codebooks at the TwinLife data documentation website.

- A list of all variables (including their frequencies), their corresponding question text, answer options, and filter conditions can be found in the Datasets and Instrument files at the metadata documentation platform paneldata.org.

Questionnaires

- The questionnaire files include the implemented questionnaires. These files are only available in the original study language (German). Original questionnaires are available on the Downloads section on the TwinLife data documentation website.

- The Technical Reports Series are scientific contributions dealing with technical aspects of data presentation (e.g. sampling design and data collection, data documentations, field reports, etc.) as well as methodical questions. Furthermore, they contain various reports on data handling for a selection of constructs and variables.

Citation

Please acknowledge the use of the TwinLife data in your work by citing both the dataset (Diewald et al., 2019) and the reference paper (Hahn et al., 2016).

Change log

Since the last release of the ShortGuide (v1.0.0), the following changes have been made:

- Update to the release of the latest data collections CATI 2 and CoV 1 (v5-0-0) incl. correction of the release date
- Addition of Corona supplementary survey to the description of the data structure and data files (chapter 3.1) and correction of the file name structure
- Minor correction of SPSS syntax in chapter 3.9 (generalization of file path)
- Revision of chapter 5 "Generated variables and scales": shortening of the generation description, insertion of the generation description of the mig variables, insertion of a reference to the documentation website, linking of Technical Report No. 06
- Addition of the citation reference to use the current DOI

1. About TwinLife

TwinLife is a longitudinal, interdisciplinary twin family study on the development of social inequality. It takes a genetically informed life course perspective on social inequalities that acknowledges the importance of both genetic and social influences, social structure, and individual agency. The combination of genetically sensitive data, the survey design, multiple indicators of social success or failure, and a variety of environmental variables enables a fine-grained investigation of the complex interplay between nature and nurture concerning social inequality.

1.1 Basic Concept

TwinLife sets out to take a look at the **biological origin** as well as the **social origin** of social inequality (Figure 1). Using the data of identical and fraternal twins as well as their families, the impact of genetic differences on certain behavior can be determined.

Furthermore, environmental characteristics such as the socio-economic status (SES), family structure and home environment, relations among family members as well as characteristics of the neighborhood are observed.

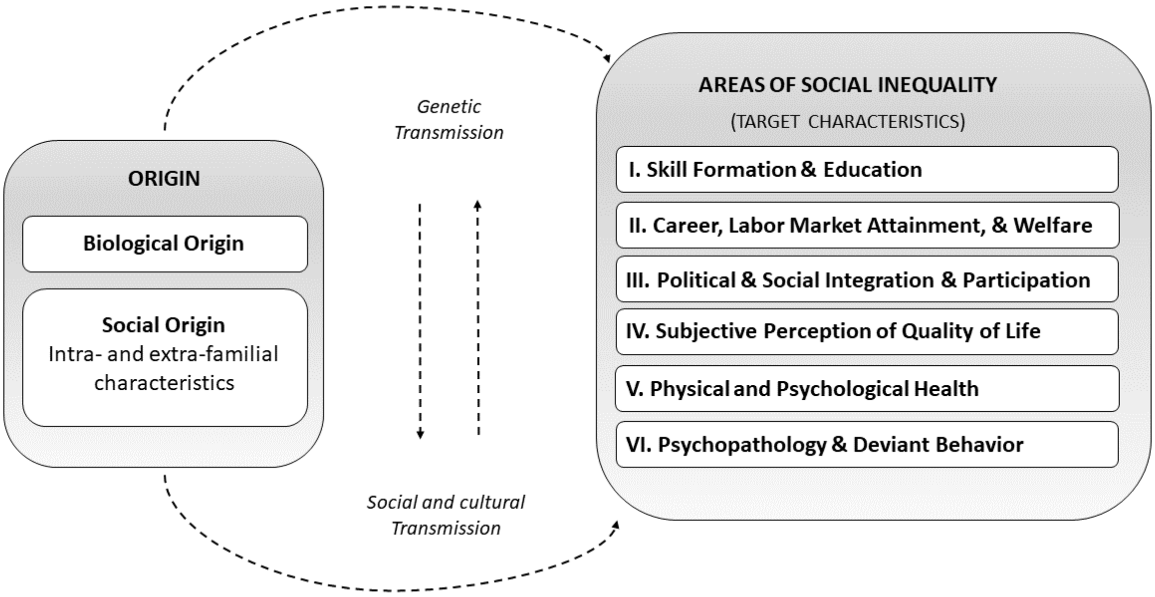
ORIGIN
Biological Origin
Social Origin
Intra- and extra-familial characteristics
Genetic Transmission
Social and cultural Transmission
AREAS OF SOCIAL INEQUALITY
(TARGET CHARACTERISTICS)
I. Skill Formation & Education
II. Career, Labor Market Attainment, & Welfare
III. Political & Social Integration & Participation
IV. Subjective Perception of Quality of Life
V. Physical and Psychological Health
VI. Psychopathology & Deviant Behavior

Figure 1: The basic concept and six domains of social inequality.

Social inequality is divided into six areas (see Figure 1) that reach from education and labor market attainment to health, psychopathology, and deviant behavior.

I. **Skill formation and education**
This area contains e.g., the level of education, educational aspirations, and achievement motivation in addition to cognitive abilities and their development.

II. **Career, labor market attainment, and welfare**
This domain includes the perceived security of the workplace, job satisfaction, commitment to work, current occupation, and current position as well as income or, if applicable, the receipt of social benefits.

III. **Political and social integration and participation**
The characteristics of the social environment of every person such as the support by family, friends, and spouses are covered in this area. Furthermore, social and political commitment and social resources are of particular interest.

IV. **Subjective perception of quality of life**
This category comprises, e.g., a person's self-esteem, global life satisfaction as well as their satisfaction concerning specific areas.

V. **Physical and psychological health**
Assessments of general health in terms of diseases, but also subjectively perceived impairments as well as information on health behaviors are covered in this area.

VI. **Psychopathology and deviant behavior**
Criminal or delinquent behavior as well as the degree of internalizing and externalizing problem behavior are surveyed in this domain.

1.2 Study Design and Sample Structure

Data collection began in 2014 with a population-based sample of 4,097 twin families. The cross-sequential survey design (see Figure 2) contains four twin birth cohorts with ~1,000 same-sex (both monozygotic and dizygotic) twin pairs.

Face-to-face interviews within the households take place every other year, and telephone interviews are conducted in the consecutive years. In the face-to-face interviews, data was collected using a mixed-mode design: Participants were surveyed by an interviewer (computer-assisted personal interview; *CAPI*), by means of questionnaires on a tablet or laptop (computer-assisted self-interview; *CASI*), via a paper-and-pencil interview (*PAPI*), and/or via online questionnaires (computer-assisted web interview; *CAWI*).

On the one hand, these mixed modes ensured the most suitable assessment strategy for each question type, i.e., more sensitive topics were covered via CASI. On the other hand, they allowed a certain degree of flexibility for the interviewer in order to minimize the total interview duration in the households.

School reports and developmental check-up reports were scanned as part of the CAPI and encoded afterwards.

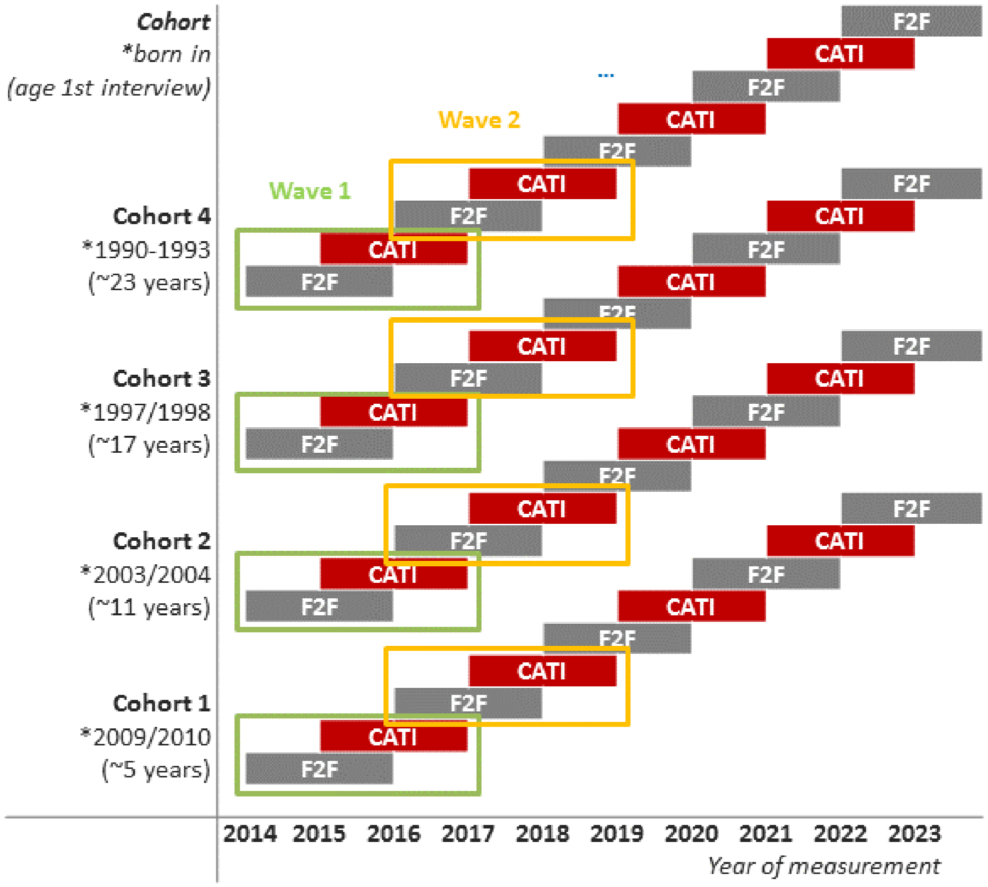

Figure 2: Cross-sequential survey design.

The cross-sequential structure is combined with an Extended Twin Family Design (ETFD, see Figure 3). The sample contains not only monozygotic and dizygotic twins, but also their biological, adoptive or foster parents, one biological, adoptive, or foster sibling (if available), as well as (if applicable) stepparents and partners of the twins. Thus, the ETFD captures both the biological family of the twins as well as the environment the twins live in.

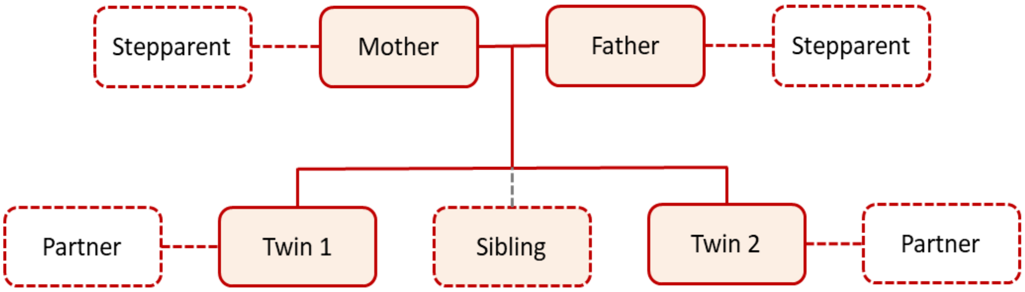

Figure 3: ETFD sample structure.

There are two characteristics of data collection and sample composition, which are briefly explained here.

First characteristic: Two subsamples

When the study was implemented, two subsamples of the initial age cohorts (consisting of twins aged about 5, 11, 17, and 23 years) were drawn, which were born in consecutive years (see Figure 2). This was necessary to achieve a sufficiently large sample size for TwinLife based on the target population of twin families in Germany.

As a result, the first subsample (Subsample a) consists of twins born in 1990/91, 1997, 2003 and 2009 and the second subsample (Subsample b) consists of twins born in 1992/93, 1998, 2004 and 2010. The subsample a) was interviewed for the first time in 2014 while subsample b) was interviewed first in 2015. Together, subsamples, a) and b) form the complete sample of TwinLife.

☛ Identifiers for the two subsamples are recorded in variable *wav0100*.

Second characteristic: Two data collections are one survey wave

Face-to-face interviews (F2F) are carried out biennially in TwinLife. In the year between, a part of the sample is surveyed via telephone interviews (CATI). The main purposes of the CATI interviews are: 1) keeping in touch with the participants, 2) updating the contact information of the families (tracking), 3) collecting some complementary data that could not be surveyed in the face-to-face interview due to time restrictions, and 4) keeping track of important life events and transitions. Therefore, each F2F data collection together with the consecutive CATI data collection including both subsamples a) and b) is defined as one survey wave.

☛ The identifiers for the survey waves are recorded in variable *wav0200* and the identifiers for the type of data collection (F2F or CATI) are recorded in variable *wav0300*. The consecutive numbering for each data collection is recorded in variable *wid* (see Table 1 for an overview).

wav0200	wav0300	wid	wav0100	(planned) release date
Wave 1	Face-to-Face 1 (F2F 1)	Data collection 1	Subsample a	Autumn 2017
			Subsample b	
	Telephone interview 1 (CATI1)	Data collection 2	Subsample a	January 2019
			Subsample b	
Wave 2	Face-to-Face 2 (F2F 2)	Data collection 3	Subsample a	Spring 2020
			Subsample b	
	Telephone interview 2 (CATI 2)	Data collection 4	Subsample a	Spring 2021
			Subsample b	
Wave 3	Face-to-Face 3 (F2F 3)	Data collection 5	Subsample a	Winter 2021
			Subsample b	
	Telephone interview 3 (CATI 3)	Data collection 6	Subsample a	Winter 2022
			Subsample b	
Wave 4	Face-to-Face 4 (F2F 4)	Data collection 7	Subsample a	Winter 2023
			Subsample b	
	Telephone interview 4 (CATI 4)	Data collection 8	Subsample a	Winter 2024
			Subsample b	
Wave 5	Face-to-Face 5 (F2F 5)	Data collection 9	Subsample a	Winter 2025
			Subsample b	

Table 1: Survey and sample structure.

Usually, there is one data release each year, including new data of either a F2F or a CATI data collection for both subsamples.

For further information about the TwinLife sample and sampling design, please take a look at the methodology report of the first wave (Brix et al., 2017) and the corresponding article (Lang & Kottwitz, 2020). Methodology reports of all data collections including a fieldwork description and outcomes can be found at the Downloads section of TwinLife's documentation website.

1.3 Where and how to get the Data

To receive the TwinLife data, please fill in the Data Use Agreement which you can find under 'Actions' in the GESIS data catalogue. All data sets are available in both English and German and in SPSS and Stata formats.

2. Documentation of the Study

There are various ways of getting informed about the data sets and variables. A first overview of the TwinLife data can be found and downloaded here.

2.1 Data Documentation Website and ShortGuide

The central documentation website of TwinLife gives the user a quick overview of the study, its structure and documentation and provides information on how to handle the data. It largely corresponds to the contents of the ShortGuide. In addition to the ShortGuide, the documentation website provides all documents mentioned in the ShortGuide in the Downloads section, including questionnaires, methodology reports as well as helpful syntax files, wave-specific and longitudinal overviews (e.g., ZA6701_longitudinal_overview_v4-0-0.xlsx). The tables contained in the latter describe specific facets of the main study domains surveyed in each year. They also include information on the family members and age groups that were questioned as well as information on the sources of the items and scales. Furthermore, the download section contains documentation of changes between the different data releases (e. g. ZA6701_changes_in_v4-0-0.pdf).

2.2 Documentation within the Data Sets

Stata users can find several information on variables within the notes attached to the dataset. Use the command -note *varname*- in order to get the following information:

- filter condition in Stata language
- question and item text in German and English
- the data collections in which the variable has been surveyed

2.3 paneldata.org

The platform paneldata.org documents the data sets and meta data of various German panel studies. At https://paneldata.org/twinlife you can find a brief study description of TwinLife, information on the survey methods, topics and the data access. Furthermore, detailed information on variables and instruments are available: Use the keyword search or the topic list in order to find your variable of interest. By clicking on the variable, you receive information on

- labels
- categories
- frequencies
- related instrument(s) and variables

At https://paneldata.org/twinlife you can also find all questionnaires and the question wording for each variable including the input filter of the question in English and in German.

2.4 Codebooks

The Codebooks give an overview of the variables included in the current data release of the person files of the data collections (for the structure of the delivered data, see chapter 3.1) and their frequencies. The variables are sorted alphabetically within content-related chapters. The codebooks document the following information for these variables:

- variable name
- variable label (also includes information on age groups and person types surveyed)
- question text
- categories (including the value labels) and frequencies
- the waves the variable was surveyed in
- filter conditions based on the instruments used for the variable
- sources for additional information

The codebooks for all data collections can be found in the Downloads section of the TwinLife Data Documentation website.

2.5 Technical Report Series, Methodology Reports, and Working Paper Series

The TwinLife Technical Reports (ISSN: 2512-403X) are scientific contributions dealing with both technical aspects of data presentation (e.g., sampling design and data collection, data documentations, field reports, etc.) and methodological questions. The version number of a technical report consists of three digits separated by dots (e.g., *v1.0.0*). The first digit marks significant changes (major release), the second digit designates extensions and additions (minor release), and the 3rd digit contains only bug fixes (revision level).

The Methodology Reports (first-authored by the survey institutes) contain detailed information on the data collection and the instruments included in each survey wave and are part of the Technical Report Series.

The TwinLife Working Papers (ISSN 2512-4048) are refereed scholarly papers that provide a forum for presenting work in progress.

All Technical Reports (including the Methodology Reports) and Working Papers are available here.

☛ **Are you using the TwinLife data?** The TwinLife Working Paper Series and the TwinLife Technical Report Series are open for your submission! Contributions are reviewed by the editors before a decision on publication is made. Please submit your manuscript to submission.twinlife@uni-bielefeld.de.

3. Data Structure

The following chapter describes the data sets, the data format, and the data structure itself.

3.1 Data Formats and Data Files

The TwinLife data is available free of charge and can be accessed via GESIS after filling in the Data Use Agreement. The data delivery consists of a certain number of data files in SPSS and Stata format:

- Master data (ZA6701_master_v$): Includes information on the gross sample, such as consistency checked variables that are stable over time (sex, year of birth, relation to the twins, zygosity, migration background) and wave-specific variables (person type, response status, family composition) about all individuals included in TwinLife in each wave.
- Survey data in person format (ZA6701_person_wid$_v$): There is one data set for each data collection (F2F 1, CATI 1, F2F 2, CATI 2). Each surveyed person has one data row. The data collection identifier is the variable *wid*.
- Data of covid supplemental surveys (ZA6701_person_cov$_v$): There is one data set for each covid supplemental survey. Each surveyed person has one data row. The data collection identifier is the variable *cov*.
- Survey data in family format (ZA6701_family_wide_wid$_v$): There is one data set for each data collection (F2F 1, CATI 1, F2F 2, CATI 2). Each family has one data row with information of each participating person in the family being stored in separate variables/columns). See chapter 3.3 for more detailed information. Person format and family format data sets contain the same data using different structures.
- Twin zygosity assessment (ZA6701_zygosity_v$): A data file with the information of the twin zygosity assessment in F2F 1.
- Unadjusted data of all variables collected in the PAPI survey mode (ZA6701_person_unadj_wid$_v$): One data file for each data collection with data unadjusted for filter errors for all constructs/variables that were at least partly surveyed in the PAPI mode (as of data release v4-1-0 in autumn 2020). See chapter 4.2 for further details.

All data is provided with English and German variable descriptions. In Stata, these languages are included in one data set while in SPSS, these are separate data files. The data are checked for inconsistencies and adjusted for filter errors (for details see chapter 4).

3.2 Person Types

According to the Extended Twin Family Design (ETFD), multiple persons in the family are surveyed. Each type of person in the family has a specific letter and a code that are displayed in the variable names of proxy-questions and the variable *ptyp*. The characters are specified in Table 2.

Letter in variable name	**Person**	**Code in variable *ptyp***
t	twin 1	1
u	twin 2	2
s	surveyed sibling - full, half or step; for details see variable fpr0107	200
p	partner of twin 1	110
q	partner of twin 2	120
m	mother of the twins - biological, adoptive, or foster; for details see variable fpr0107	300
n	partner of the twins' father / 'stepmother'	600
f	father of the twins - biological, adoptive, or foster; for details see variable fpr0107	400
g	partner of the twins' mother / 'stepfather'	500

Table 2: Person codes. Please note: Additionally, non-participating siblings (ptyp 201 – 2xx) and other household-members (ptyp 700 – 7xx, 800) are coded in the variable 'ptyp'.

By using the variable *ptyp*, the data can be restricted to a certain sample of person types that are of interest for the analysis. The variable *fpr0107* describes the relation to the twins in more detail, e.g. it differentiates full siblings, half siblings and stepsiblings, etc.

Please note: The person type of twins in *ptyp* is always identical to the person type in the first data collection and always corresponds with the last three digits of the person ID *pid*. If the information on the relation to the twins was corrected in one of the most recent surveys, the variable fpr0107 in the master data set is adjusted accordingly. Thus, the information differs between *ptyp* and *fpr0107* in very rare cases. For instance, if one twin was declared as the 'first-born twin' in the first data collection, but this was corrected to 'second-born twin' in a more recent data collection, the twin is longitudinal-consistently treated as the first-born twin in *ptyp*, but files as second-born twin in *fpr0107*. Thus the birth order of the twins is documented in *fpr0107*, not in *ptyp*. Please keep this in mind when using the variables for sample selection or analyses.

3.3 System of Variable Names

Generally, the variable name is a composite of the variable stem (three letters), the item block (two digits), and the item number (two digits). Furthermore, depending on the person- or family-format of the data sets, person code letters in variable names indicate that proxy information is given from and/or about a certain family member (see chapter 3.2 for person codes).

Examples

1. What information is in the variable name *edu0100* in the 'person format' ('long' dataset)?

> ☛ edu is the variable stem (type of education – self-report), 0100 is 'school attendance' within the facet 'type of education'.

2. What information is in the variable name *pas0100m* in the 'person format' ('long' dataset)?

> ☛ pas is the variable stem (parental style – child report), the four digits 0100 are the item number within the construct. In most cases, for scale variables the first two digits are the item block within the construct (01, parental style) and the last two digits are the item number (00, shows affection); the variable extension m is the person code (statement about the mother). Thus, this item assesses the child's rating of parental style, more precisely on how much affection the parent, here the mother, shows.

3. What further information is in the variable names in the 'family format' ('wide' dataset), e.g. *pas0100m_s_1*?

> ☛ _s is the person code of the person who provides the information (sibling), _1 is the data collection code (first data collection). Thus, this item assesses the child's rating of parental style, more precisely on how much affection the parent (in this case mother) shows – rated by the sibling in the first data collection.

In the family format, two particularities have to be mentioned:

First, information on the households get the code *_1* to *_x* (x= maximum number of households within families) instead of a person code.

- For example, *ptyp_hq_1_1* provides information on the type of respondent for the household questionnaire in data collection one for the first household which belongs to the family.
- The variable *pih_$_$* shows the household number of a person (e. g. *pih_t_1* is the household number of the first-born twin in data collection 1).

Second, information given by a parent about their children is not filed under the type of person of the interviewed parent. Instead, the letter suffix reflects for which person information was given.

- For example, the variable name eca0300t_1 shows that the statement was made for the first-born twin, but provides no information about the parent who filled in the questionnaire.
- Instead, the variable *ptyp_cp_1* indicates which parent completed the questionnaire about their children in data collection 1.

Variable labels contain a short description of the variable content and information about filter conditions (e.g., which person types and which age groups have answered the question). Generated variables are marked with '(gen)'; labels of variables with differing age filters over time include the minimum age of respondents across data collections.

3.4 ID Variables, Wave and Data Collection Identifiers

In TwinLife, various ID variables/identifiers are available. Each person belongs to a family that has a unique family ID (*fid*) and to a household that has a unique household ID which is wave-specific (*hid*; a composite of the family ID and an indicator for the household).[1] Additionally, each person also has a unique person ID (*pid*, which is a composite of the family ID and the person type). Although the person types except for the twins can change (i.e. '700 - other person' might change to '110 - partner of twin', or '200 - surveyed sibling' might change to '201 - non-surveyed sibling'), the person ID is invariable over time.

The family ID consists of six digits: the first digit indicates the twin birth cohort (e.g., 1 for the first cohort; note that information about birth cohort is also coded in variable *cgr*); the other five digits are assigned randomly. ID variables are particularly important when different data files have to be combined. To match data of different survey waves in the family-wide-format, the variable *fid* needs to be used; to match the master data set with the person-format, the variable *pid* has to be used. Please note that time variable information in the master data set need to be reshaped into the long format in order to match the data with the person-format of the survey data. Before matching the master data set with the family format, the master data set has to be restructured to family format.

Furthermore, the variables with variable stem *wav* describe exactly in which survey wave (wav0200, wav0300) and subsample (wav0100) the data was assessed. The variable *wid* is the data collection identifier (wid == 1 stands for the first face-to-face household survey (F2F1), wid == 2 for the first telephone survey (CATI1), and so on).

[1] The indicator for the household and therefore the household ID itself are linked to the information whether the twins live in this household or not. It is possible that in two consecutive years two households with different household compositions have the same household ID. Therefore, *hid* should not be used for longitudinal analyses.

3.5 Missing Types

In TwinLife, missing values are delivered in a differentiated way. Table 3 gives an overview of the standard missing codes and a short explanation of their meaning.

Value	Value label	Explanation
-99:	not specified (refused to answer)	The respondent has explicitly refused to answer the question.
-98:	don‘t know	The respondent has explicitly stated not to know what to answer.
-96:	mixed missing values – e.g., don‘t know / not specified	Here, the instrument has not differentiated between two or more types of missings. This is mostly the case in the paper and pencil questionnaires of the first survey wave.
-95:	doesn't apply (screened out)	It indicates that the question was not intended for the respondent based on the filter conditions, for instance, due to the respondent’s age.
-94:	technical error	In most of these cases, the filter condition of the question was programmed incorrectly so that respondents have falsely received or not received a certain question.
-93:	unclear classification of system missing (only for paper-and-pencil questionnaires)	Here, it is not possible to determine what the reason for a missing value is (whether it does not apply, the respondent does not know or whether they do not want to reply).
-92:	no participation in survey module	The respondent has not participated in the survey module / questionnaire, either because the questionnaire did not apply for the respondent or they refused to participate.
-90:	no participation in data collection	The person has participated in past data collections or may participate in future data collections, but not in the current one (all variables have the missing -90 in this wave for this person).
-87:	multiple answers	The person has given multiple answers in questions where only one answer is possible (usually in paper and pencil questionnaires).
other -80s:		… have different meanings and can contain valid information (see below)

Table 3: Missing codes.

Please note that the missing codes -80 to -89 (except for -87) have special meanings that can differ between variables and can be of importance for analyses. Generally, they contain ‘valid’ information but the answers are not direct answers to the question and are therefore coded as negative missings.

One example is the variable *emp0800* in wid == 3 (data collection F2F 2) “When did you leave your last job?” – The answer that the interviewee *has not been employed until now* is not a direct answer to the question but contains valid information and is therefore coded as a negative missing (-81).

In contrast, the answer “I already have children” on the question “Having children: How likely do you think you are to achieve this goal in your lifetime?” (*lgd0204*) in wid == 3 (data collection F2F 2) is a valid and analyzable answer to the question and is therefore coded with a positive code outside the actual answering scale (96).

You should check all variables for positive and negative meaningful missing values before using them for analyses.

3.6 Delivered Para Data

Para data are administrative data about the survey that are collected besides the actual survey data. In TwinLife, mainly month and year, in which an interview was conducted, is delivered as para data. Time stamps of questions or thematic blocks are recorded as well, but not delivered in the Scientific Use File. Under certain circumstances, it is possible for researchers to get access to the time stamps (e.g., for methodological research). Please contact info@twin-life.de if you are interested in using these data.

3.7 Weights

Currently (released data version v5-0-0), no weights are delivered.

3.8 Peculiarities of Data

Peculiarities concerning families or persons are coded in the variable *pec* in the master data set. It records cases which in some way deviate from the sampling design, and which should, therefore, be considered when analyzing the data. For instance, this is the case when the twins are in fact part of triplets. Other examples are that the twins are orphans or, in very rare cases, the twins do not belong to the target birth cohorts defined within the sample design.

3.9 How to match the Data Files

For longitudinal studies, the data sets of different survey data collections need to be combined.

The single data sets can easily be appended as variable names and categories have already been harmonized across all data collections.

For the person long format, different matching strategies can be chosen, depending on the desired data structure of the combined dataset ('long': several rows per person (one for each data collection) and one column per variable vs. 'wide': one row per person and a column for each data collection of variables).

In the following we provide syntax for Stata and SPSS for both cases. Especially for the family wide format, it is strongly recommended to only use and merge the variables that are needed for the analyses in order to limit the size of the final data set.

Matching data files in Stata

1. Person long format, ‘long’ (one row per data collection for each person and one column per variable). The following example for combining the two face-to-face data collections F2F1 and F2F2 can be customized:

```
cd “path” // navigate into the folder were the data is stored
use ZA6701_person_wid1_v4-0-0.dta // fill in the name of the data set in the
version you are using
append using ZA6701_person_wid3_v4-0-0
append using … // optionally append further files of all data collections you
want to use for longitudinal analysis
```

   You can also limit the data to the variables you want to analyze by using the command

```
use varlist using ZA6701_person_wid1_v4-0-0.dta // replace ‘varlist’ by the list
of variables you want to use for your analyses
```

2. Person long format, ‘wide’ (one row per person over all data collections and one column for each data collection and variable). Append the data of the data collections you want to analyze using the procedure described in 1):

```
cd “path”
use ZA6701_person_wid1_v4-0-0.dta
append using ZA6701_person_wid3_v4-0-0
```

   Use the -reshape- command in order to get the person-wide format:

```
local varselect "varlist" // select list of variables that need to be converted
from long to wide form
rename (`varselect') =_  // add suffix
reshape wide *_, i(pid) j(wid) // convert selected variables from long to wide
form
```

3. Family wide format, ‘wide’ (one row per family over all data collections and separate columns for variables per person and data collection).

For analyses using the family wide format with Stata, use the -merge- command with the family identifier *fid*.

```
cd “path” // navigate into the folder where the data is stored
use varlist using ZA6701_family_wide_wid1_v4-0-0.dta // replace ‘varlist’ by
the list of variables you want to use for your analyses
merge 1:1 fid using ZA6701_family_wide_wid3_v4-0-0.dta, keepusing(varlist) //
replace ‘varlist’ by the list of variables you want to use for your analyses
```

Matching data files in SPSS

1. Person long format, ‘long’ (one row per data collection for each person and one column per variable). The following example for combining the two face-to-face data collections F2F1 and F2F2 can be customized:

```
add files
/file= 'path\SUF_4-0-0_beta_04052020\ZA6701_en_person_wid1_v4-0-0.sav'
/file= 'path\SUF_4-0-0_beta_04052020\ZA6701_en_person_wid3_v4-0-0.sav'.
save outfile= 'path\SUF_4-0-0_beta_04052020\en_person_wid13_match.sav'.
exe.
```

2. Person long format and Family wide format, ‘wide’ (one row per person/family and one column for each data collection and variable; one row per family over all data collections and separate columns for variables per person and data collection).

   If the combined data needs to be in wide format, it is important that all variables (except for the matching variables) in every dataset have a data collection-specific suffix. In the person format, this suffix has to be created for all variables except *pid* before matching. In the family format, wave-specific suffixes are already provided (except for the variables *wav0100*, *cgr* and *zyg0102,* which are time stable and therefore identical in all waves). Variable suffixes can be easily created using the python plugin. The following code can be customized to do this:

```
begin program.
variables = 'all' # define the variables which should get a suffix, you can use
e.g. 'all', 'x, y, z'; 'x to y'.
suffix ='_1' # enter the chosen suffix.
import spss,spssaux
oldnames = spssaux.VariableDict().expand(variables)
newnames = [varnam + suffix for varnam in oldnames]
spss.Submit('rename variables
(%s=%s).'%('\n'.join(oldnames),'\n'.join(newnames)))
end program.
```

When each dataset has data collection-specific suffixes, all datasets must be sorted by the matching variable; datasets in person format by the *pid*, datasets in family format by the *fid* (see chapter 3.5). To finally combine two data sets, the following code can be customized:

```
sort cases by pid.
match files
/file= 'path\SUF_4-0-0_beta_04052020\ZA6701_en_person_wid1_v4-0-0.sav'
/file= 'path\SUF_4-0-0_beta_04052020\ZA6701_en_person_wid2_v4-0-0.sav'
/by pid.
save outfile= 'path\SUF_4-0-0_beta_04052020\en_person_wid12_match.sav'.
exe.
```

4. Check Routines and Data Adjustment

There is a range of quality check routines that were carried out by the survey institute and the TwinLife team in order to ensure the consistency and plausibility of the data. Furthermore, the data are adjusted for filter inconsistencies.

4.1 Check Routines

The following types of checks are carried out by the TwinLife team or the survey institute (detailed information about the procedures will be provided in a Technical Report):

1. Longitudinal consistency
    - variable pid, sex identical to prior waves
    - body height and weight
    - variable coding longitudinally consistent
2. Identity checks
    - match of sex, first name, date of birth, relationship to twins between preload and survey
    - checks for duplicates (persons included several times in the sample?)
    - change of surveyed sibling identified and corrected
    - household composition
3. Logical relations
    - age of twins vs. cohort
    - age vs. ptyp
    - age and sex of (biological) parents
    - twins' age and sex identical
    - body height and weight
4. Completeness
    - Starting sample vs. realized sample (on personal, household, and family level)
    - Instrument/questionnaires vs. delivered variables
5. Instrument
    - Completeness and correctness of preload variables and generated filter variables
    - filter conditions
6. Variables
    - correctness of variable labels and value labels
    - correctness of missing values

4.2 Data Adjustment

The TwinLife data have been adjusted for some filtering inconsistencies that occur when a respondent answers the filter entry question and the following question(s) in an inconsistent manner with regard to the previous answer. This is mainly a phenomenon in the paper-and-pencil questionnaires (PAPI) that are filled by the respondents without interviewer assistance. In rare cases, there can be programming errors in the CAPI/CASI modules as well which can lead to filter inconsistencies.

The TwinLife data are adjusted for this kind of inconsistency, which means that entries or answers not meeting the filter conditions are deleted. This procedure assumes that the filter entry question was answered correctly whereas the following questions were answered incorrectly.

The constructs that are mainly affected by the adjustment are discrimination (dis), dia (diagnoses), hbe and doc (health-related behavior), spa (academic self-concept), del (delinquent behavior), net (social networks), imo (motivation), sat (domains of life satisfaction), sop (social participation) and mus (cultural capital) because they were at least partly surveyed in a paper-and-pencil questionnaire and included (more or less complex) filter conditions. With the interim data release v4-1-0, which will be provided in autumn 2020, it is planned to release the unadjusted data for all constructs/variables that were at least partly surveyed in the PAPI mode. Users should carefully review whether and which data they use in which way for their analyses. The TwinLife Technical Report 07 proposes a way how to treat the unadjusted variables that belong to the discrimination construct which is particularly affected by filter inconsistencies.

Please note: The adjustment was not carried out for the igf-variables (intelligence test) in the second face-to-face interview (F2F2), where some of the participants were falsely treated as new members of the sample and took the test a second time. Therefore, the igf-variables of the F2F2 data contain values for participants who should not have filled in the test again. Please consider this when analyzing the igf-variables.

5. Generated Variables and Scales

In the following, it is briefly described which variables are generated and provided within the data preparation process as well as which and how scales can be generated by the users themselves. Perspectively, details about the generation processes of different variables will be published in Technical Reports. A more detailed description of the generated variables and scales is provided in the online documentation of the survey.

5.1 Generated Variables

In the course of data processing, some variables were generated based on the raw variables that are surveyed in the field.

Household and personal income manipulation (inc0110, inc0111, inc0401)

In order to guarantee the twin families' anonymity, we applied a bottom- and top-coding strategy for the (household and personal) income variables inc0100, inc0101 and inc0400 following the recommendations of Wirth (1992)[2].

Net equivalent household income (inc0411)

We generate and provide the net equivalent household income using the modified OECD scale. Here, the net household income is weighted according to household size and age composition taking into account rising scale effects for bigger and younger households. For more information about the concept, see OECD (2011).[3]

ISCED-1997 (eca0106)

Drawing on the highest reported educational qualification, the International Standard Classification of Education (ISCED) measures the individual education, ranging from 0 (pre-primary education) to 6 (second stage of tertiary education). For more information about the ISCED classification see OECD (1999).[4]

Classifications of occupation and occupational activity: ISCO-08 (eca0205, emp0503, emp0513, emp0553), KldB-2010 (eca0205, emp0501), SIOPS (eca0208, emp0506), ISEI (eca0207, emp0505), EGP (emp0507)

The International Standard Classification of Occupations (ISCO) from 2008 classifies occupations based on the required skill level and the degree of skill specialization. TwinLife delivers the two-digit ISCO codes. For more information about the ISCO codes visit the ILO website.

SIOPS (Standard Index of Occupational Prestige Scala) is a classification for a prestige ranking of occupations ranging from 0 to 100 based on the ISCO-88 classification. For more information see Ganzeboom & Treiman (1996).[5]

ISEI (International Socio-Economic Index of Occupational Status) is a measure for the socio-economic status of a person ranging from 12 to 90 based on the ISCO-88 classification. For more information see Ganzeboom & Treiman (1996).

EGP class typology (Erikson-Goldthorpe-Portocarero classes) is a scheme of social classes. For more information, see Ganzeboom & Treiman (1996).

Housing conditions and household type (liv0210, liv0410)

The housing conditions were surveyed on the household level. The variable liv0210 provides information about the housing conditions from the twins' perspective on a personal level. The variable liv0410 provides information about the household type on a personal level.

[2] Wirth, H. (1992). Die faktische Anonymität von Mikrodaten: Ergebnisse und Konsequenzen eines Forschungsprojektes. [The factual anonymity of microdata: results and consequences of a research project]. ZUMA Nachrichten 16, 30, 7 - 65.

[3] OECD (2011). What Are Equivalence Scales? OECD Project on Income Distribution and Poverty.

[4] OECD (1999). Classifying educational programmes: Manual for ISCED-97 implementation in OECD countries. Organisation for Economic Co-operation and Development.

[5] Ganzeboom, H. B. G., & Treiman, D. J. (1996). Internationally comparable measures of occupational status for the 1988 International Standard Classification of Occupations. Social Science Research, 25 (3), 201-239.

Regional variables (ewi, gkpol_r)

Due to data privacy reasons, regional information are only provided in an aggregated form. Whether a household lives in East or West Germany is stored in the variable ewi, while the variable gkpol provides information about the (political) community size.

Country of birth (mig2000, mig2100, mig2200), Born in the GDR (mig2001, mig2101, mig2201), German citizenship (mig0520)

The information about the country of birth was collected by providing a list as well as the opportunity to give an open answer. For data privacy reasons the answers were recoded into country groups and stored in generated variables, where the self-reports and the proxy-reports of other persons were integrated into the final country of birth variables.

Whether a person was born in the GDR is generated on the basis of the country of birth and considers whether the person was born during the existence of the GDR between 7/10/1949 and 2/10/1990.

The variable mig0520 includes all available information on the personal level and displays whether the individual has the German citizenship or not, using self-reports and proxy information.

Report cards / certificates (cer variables)

School performance data based on photographs of the children's report cards are stored in the *cer* variables. For more information, please see the TwinLife Technical Report No. 04.

5.2 Generated Scales

For some constructs, the Twinlife data includes generated scales (e.g., intelligence, BMI, or zygosity), taking into account possible pitfalls of the data. The recoding process is documented in the TwinLife Technical Report No. 06. For more information about other commonly used scales and constructs in the social sciences included in the TwinLife data, we refer to the scales manual.

6. Publications and Citation

6.1 Publications and Literature Database

The TwinLife team puts great effort in preparing the data and making it available for the open science community. Every publication provides important insights into the forming of social inequalities and is highly appreciated.

Publications based on the TwinLife data are documented and can be found on the TwinLife website. Those lists are updated regularly and cover the following sections:

The TwinLife Bibliography contains the bibliographic information on all known publications related to TwinLife. All researchers who publish results based on the TwinLife data are kindly asked to provide information about their publication to info@twinlife.de so that their publications can be listed. We appreciate your compliance.

The Conference contributions contain a list of already presented contributions to national and international conferences.

TwinLife Working Papers are refereed scholarly papers that provide a forum for presenting work in progress. Submissions can be sent to submission.twinlife@uni-bielefeld.de and are reviewed by the general editors before a final decision on publication is made.

TwinLife Technical Reports are scientific reports on technical data presentation (e.g., data documentation, field reports, etc.) and methodical issues.

6.2 Citation

When preparing a publication based on TwinLife data, we kindly request you to mention our study in your publication. Please acknowledge our work by citing both the reference paper (Hahn et al., 2016) and the dataset itself (Diewald et al., 2019). Please keep in mind that the citation of the dataset may change as the study progresses and newer versions become available. To check the latest citation version, see the study's entry in the GESIS Data Catalogue. Older versions of the data and the related DOI's can also be found there.

☛ **Correct citation of the TwinLife reference paper (APA):**

Hahn, E., Gottschling, J., Bleidorn, W., Kandler, C., Spengler, M., Kornadt, A. E., ... & Spinath, F. M. (2016). What drives the development of social inequality over the life course? The German TwinLife Study. Twin Research and Human Genetics, 19(6), 659-672. doi:10.1017/thg.2016.76

☛ **Correct citation of the dataset, e.g. version 4.0.0 (APA):**

Diewald, M., Riemann, R., Spinath, F. M., Gottschling, J., Hahn, E., Kornadt, A. E., … & Weigel, L. (2020). TwinLife. GESIS Data Archive, Cologne. ZA6701 Data file Version 4.0.0, doi:10.4232/1.13539 (please insert the DOI of the data version you are using)

7. Useful Links

- TwinLife Data Documentation:
  https://www.twin-life.de/documentation/
- Download of all additional documents mentioned in the ShortGuide and on the documentation website:
  https://www.twin-life.de/documentation/downloads
- Meta data documentation via paneldata.org:
  https://paneldata.org/twinlife
- TwinLife reference paper (Hahn et al., 2016):
  https://pub.uni-bielefeld.de/publication/2906305
- Technical Report on sampling design and socio-demographic structure of the first wave of the TwinLife (Lang & Kottwitz, 2017):
  https://pub.uni-bielefeld.de/record/2913250
- Dataset citation:
  https://search.gesis.org/research_data/ZA6701
- GESIS data catalogue:
  https://dbk.gesis.org/dbksearch/sdesc2.asp?no=6701&db=e&doi=10.4232/1.13208
- Data Use Agreement:
  https://www.twin-life.de/documentation/images/TwinLife/Downloads/Data_Use_Agreement_TwinLife.pdf
- TwinLife Technical Reports and TwinLife Working Paper Series:
  https://www.twin-life.de/twinlife-series
- TwinLife Bibliography:
  https://www.twin-life.de/publikationen
- TwinLife conference contributions:
  https://www.twin-life.de/konferenzbeitraege