2022

DATA MANUAL, SCIENTIFIC USE FILE VERSION 1.0

RESEARCH CENTRE OF DEMOGRAPHIC CHANGE (FZDW)

HEALTH BEHAVIOR AND INJURIES IN SCHOOL AGE (GESUNDHEITSVERHALTEN UND UNFALLGESCHEHEN IM SCHULALTER, GUS)

DATA MANUAL

ROBERT LIPP

Author:
Robert Lipp

Suggestion for citation:
Lipp, R. (2022). *Data Manual: Health Behavior and Injuries in School Age, Scientific Use File Version 1.0*. Frankfurt am Main, Germany, Research Centre of Demographic Change (FZDW), Frankfurt University of Applied Sciences.

Carrying out the GUS study was a major collaborative effort at the FZDW. The preparation of the Scientific Use File was carried out in close coordination with the Research Data Center (RDC) at LIfBi and with the friendly support of the coordinators of the KonsortSWD project "RDM Grants" at LIfBi. The contribution of the following persons is gratefully acknowledged:

FZDW (Frankfurt am Main)	*FZDW (Frankfurt am Main)*	*LIfBi (Bamberg)*
Andreas Klocke	Christina Lenz-Bokhari	Christian Aßmann
Sven Stadtmüller	Ilona Kraus	Daniel Fuß
Andrea Giersiefen	Christine Leyhe	Tobias Koberg
Christina Wagner	Mario Englert	Katja Vogel
Sarah Maier	Carina Schwaderer	Friederike Schlücker
Jannik Track	Jessica Konschu	Clara Wolf

Contact:
Frankfurt University of Applied Sciences
Forschungszentrum Demografischer Wandel (FZDW)
Nibelungenplatz 1
60318 Frankfurt am Main
E-Mail: info@fzdw.de
Web: www.fzdw.de

Table of contents

Preamble 4
1. Introduction 4
1.1 About this manual 4
1.2 Further documentation 4
1.3 Data release strategy and data format 5
1.4 Data access 6
1.5 Publications with GUS data 6
1.6 Rules and recommendations 7
1.7 Contact to the data provider 8
2. Sampling and Survey Overview 8
2.1 Health behavior and injuries during secondary education 8
2.2 Sample design 8
2.3 Data collection and questionnaire content 9
3. General Conventions 10
3.1 Variables 11
3.2 Missing Values 13
3.3 Generated variables 14
4. Special Issues 15

Preamble

The present data package is the result of a collaboration between the Research Centre of Demographic Change (Forschungszentrum Demografischer Wandel, FZDW) at the Frankfurt University of Applied Sciences and the Research Data Center at the Leibniz Institute for Educational Trajectories (Leibniz-Institut für Bildungsverläufe, RDC at LIfBi) in Bamberg. As part of a project grant from the Consortium for the Social, Behavioural, Educational and Economic Sciences in the National Research Data Infrastructure (NFDI) to support researchers in preparing and making available relevant new data for secondary use, the data collected by the FZDW project team for the "Health Behavior and Injuries in School Age" study (Gesundheitsverhalten und Unfallgeschehen im Schulalter, GUS) were comprehensively edited and documented in a user-friendly manner. With the integration of the resulting Scientific Use File into the existing research data infrastructure of the RDC at LIfBi, the GUS data are now accessible to the scientific community free of charge for empirical analyses.

1. Introduction

1.1 About this manual

This manual is intended to facilitate your work with data from the study “Health Behavior and Injuries in School Age” (Gesundheitsverhalten und Unfallgeschehen im Schulalter, GUS). It serves both, as a first guide for getting started with the Scientific Use File data and as a reference book. The primary emphasis is on practical aspects such as sample development and data structure. The manual is neither complete nor exhaustive. However, several links to other resources are provided in the following sections.

This first chapter refers to further documentation material, notes on the release strategy and the file format, requirements for accessing the data, instructions for data citation, and some general rules for data handling. The second chapter provides a general overview of the GUS study describing its original purpose, the sampling procedures and the way it was carried out. This includes field times, realized case numbers and a short questionnaire description. The principles of Scientific Use File data-editing processes including the list of identifier variables as well as the conventions for naming variables and labeling missing values are explained in the third chapter. Some details on the generation of additional variables are also given there. The fourth chapter focuses on special issues to keep in mind when working with the data.

1.2 Further documentation

The data manual already covers a lot of relevant information about the study. However, some aspects deserve a more detailed description. Thus, additional reports and documentation materials are offered via the LIfBi website: https://www.lifbi.de/GUS-Study

Questionnaires

The questionnaires contain all questions, that (1) students, (2) heads of the respective schools, and (3) the interviewers were asked during the course of the study. They are documented by wave and target population. The questionnaires for the ***heads of the schools*** as well as for the ***interviewers*** are documented in their original paper versions. The questionnaires for the ***students*** were administered on

tablet devices. However, every interviewer had some copies of a paper version of the questionnaire, in case there were not enough tablets for all students in a class. This paper version is provided with the documentation. All questionnaires are available in German only. Please note: Many of the questions were developed by the GUS project team and fall under copyright protection. If you want to reuse items from the GUS study, make sure to request permission at: info@fzdw.de

Codebook

The codebook – provided in German and English language – lists all variables of the Scientific Use File together with the corresponding values and frequency distribution by waves.

Variable overview

In order to better visualize the contents of the GUS data, an overview of the variables in the Scientific Use File was created, grouping them by topics. It shows in which waves the corresponding questions were asked.

Constructs and generated variables

All generated variables in the GUS data are marked with the suffix _g1, _g2, … , g#. These variables combine information from two or more other variables. Furthermore, they may contain coded categories of open responses. Where possible, the categories were formed using standard classifications (such as ISO 3166-1 for country codes). Detailed information, necessary to understand how the generated variables were created, can be found in the provided ***Technical Report***. For general conventions regarding variable names, labels, etc., please consult chapter 3 of this manual.

1.3 Data release strategy and data format

GUS data are published in the form of a Scientific Use File named ***GUS_SUF_D_1-0***. It is provided free of charge to the scientific community for empirical analyses. The Scientific Use File consists of a single dataset named ***GUS_Data_D_1-0***. This dataset contains panel information from the repeated survey of students and principals (=heads of schools) as well as the interviewers present during the student survey in the schools (for further information see Chapter 4). The data format is “long”, which means that:

- a row in the dataset contains the information of one student from one survey wave (i.e., a filled-in questionnaire),
- a personal identifier and the wave variable are needed to uniquely identify a single row for the information of one single student in one single wave (see Section 3.1).
- Although not all variables were administered in each survey wave, the integrated structure of the dataset contains cells for all variables in all waves. If no data is available, e.g. because a variable was not queried in a particular wave, the corresponding cells are filled with a missing code (see Section 3.2).

File format

The data in the Scientific Use File are provided in ***Stata*** and ***SPSS*** formats with bilingual variable and value labels in German and English. In the SPSS format, there are separate data files for both languages. Data stored in Stata format contain both languages within one file; the switch is induced by the Stata command: *label language [de/en]*

Due to the change of encoding to "Unicode" in Stata14 and the fact that older Stata versions are not able to open such data files, the GUS Scientific Use File contains two Stata format data packages, namely Stata14 and Stata12.

Versioning and Digital Object Identifier

Each time the Scientific Use File is updated, a new two-digit version number is assigned. The first digit informs about major releases, while the second digit is used for minor updates. The version number is also part of the Digital Object Identifier (DOI) of the respective Scientific Use File.

Every release of the GUS data package is registered at da|ra – the DOI registration agency for social and economic data in Germany – and clearly labeled with a unique DOI. The current Scientific Use File release is indicated by ***doi:10.5157/GUS:SUF:1.0***.

1.4 Data access

Access to the GUS data is free of charge but limited to the purpose of research and to members of the scientific community. Granting the right to obtain the data requires the conclusion of a ***Data Use Agreement*** with the RDC at LIfBi as the data provider. The existence of a valid Data Use Agreement entitles to work with all current and future GUS Scientific Use Files, i.e., the full data portfolio is at the disposal of the data recipient and all additional persons included in the Data Use Agreement. The data package is available via download from: https://www.lifbi.de/GUS-Study

Application for data access

- Fill in the German or English online form for the GUS Data Use Agreement. Enter a title, the duration, and a short description of the intended research project. Make sure that all project participants with GUS data access are specified in the form and that these persons have signed the agreement by hand on the last page. Submit one copy of the completed form by e-mail, fax, or mail to the RDC at LIfBi.
- After approval by the RDC at LIfBi, the applicant and all involved project participants will receive a personal user name and an initial password.
- This initial password needs to be changed to a personal password in order to log in at the website for downloading the Scientific Use File.
- For changes of the Data Use Agreement regarding adding further project participants or extending the project duration, a second form is provided on the same website as the Data Use Agreement.
- Detailed instructions and all relevant forms are available at: https://www.lifbi.de/GUS-Study

1.5 Publications with GUS data

Referencing the use of data from the GUS study is essential for a good scientific practice as well as for revealing the scientific value of this study. The following citation rules apply to all publications based on GUS data. The German language versions are provided on the website and as an additional text file within the Scientific Use File.

Firstly, it is obligatory to acknowledge the GUS study in general by including a phrase like this in your publication:

> This paper uses data from the study "Health Behavior and Injuries in School Age" (Gesundheitsverhalten und Unfallgeschehen im Schulalter, GUS). From 2014 to 2020, GUS data was collected by the Research Centre of Demographic Change (FZDW) at the Frankfurt University of Applied Sciences (Frankfurt UAS) as funded by the German National Accident Insurance (DGUV).

Secondly, the used data version of the Scientific Use File has to be indicated in the bibliography:

> Frankfurt University of Applied Sciences, & Leibniz Institute for Educational Trajectories. (2022). *Health Behavior and Injuries in School Age (GUS), Scientific Use File*. RDC LIfBi, Bamberg. https://doi.org/10.5157/GUS:SUF:1.0

Authors of any kind of publications based on GUS data are requested to notify the RDC at LIfBi about their articles and to provide an electronic version, special print, or copy. All reported publications are listed in the LIfBi Bibliography at: https://www.lifbi.de/Publications

Citing documentation

To refer to any of the documentation material published in the ***GUS Research Data Documentation Series*** (e.g., this manual), please make use of the following citation templates:

> Lipp, R. (2022). *Data Manual: Health Behavior and Injuries in School Age, Scientific Use File Version 1.0.* Frankfurt am Main, Germany, Research Centre of Demographic Change (FZDW), Frankfurt University of Applied Sciences.

If no author is given, please use the universal *FZDW* instead:

> FZDW (2018). *Fragebogen für Schülerinnen und Schüler – Gesundheitsverhalten und Unfallgeschehen im Schulalter (GUS), Welle 6*. Frankfurt am Main, Germany, Research Centre of Demographic Change (FZDW), Frankfurt University of Applied Sciences.

If a document has not been published in this series, please refer to the author(s) and the title as in the following citation of a sampling report by our survey design partner:

> Zins, S. (2014). *Stichproben Design*. Mannheim, Germany, GESIS – Leibniz Institute for the Social Sciences.

1.6 Rules and recommendations

Working with the GUS data is bound to a couple of rules that are codified in the Data Use Agreement. Each data user has to confirm these rules by his or her signature. The already mentioned obligation to cite the study and to indicate any kind of publication resulting from the use of GUS data (see Section 1.5) are just two of these rules. The major part of rules refers to issues of data privacy and the requirements of careful data handling.

- ***Avoidance of re-identification**:* Any action aimed at and suitable for re-identifying persons, or institutions is strictly forbidden. This also includes the combination of GUS data with other data that allow for a re-identification of persons or institutions. In case of any accidental re-identification, the RDC at LIfBi has to be informed immediately and all individual data gained therefrom have to be kept secret.
- ***Avoidance of data disclosure**:* GUS data are exclusively provided on the basis of a valid Data Use Agreement — for a defined purpose (research project) and to a defined group of persons

(data recipient and additional project members that are involved in the contract). Any use for commercial or other economic purposes is not permitted just as any transfer of the data to third parties. Please handle the provided GUS data with strict confidentiality.

- ***Regulations on using the federal-state label:*** For GUS data collected in connection with schools, it is not allowed to use federal-state-related information contained in the data directly or indirectly for analyses
  - aiming at direct comparisons of the German federal states (Bundesländer), or
  - aiming at direct conclusions to be drawn about a federal state, or
  - aiming at reconstruction of the concrete federal state affiliation of persons and institutions.
  - Any kind of ranking between the federal states based on GUS data is prohibited.

Please note that violation of these rules may lead to severe penalties as stated in the Data Use Agreement. If there is any doubt or question regarding the given regulations, please contact the RDC at LIfBi (see Section 1.7). The same applies in case of encountering any deficiencies in data quality or any security leaks with regard to GUS data protection and data security.

1.7 Contact to the data provider

The RDC at LIfBi accounts for archiving the GUS data as well as for data dissemination and preservation. For questions, comments, requests, and suggestions regarding the data or documentation, please contact the RDC staff:

E-mail: fdz@lifbi.de
Web: https://www.lifbi.de/Institute/Organization/Research-Data-Center
Phone: +49 951 863 3511

2. Sampling and Survey Overview

2.1 Health behavior and injuries during secondary education

The data originates from the panel study “Health Behavior and Injuries during School Age”, which started on an annual cycle in the school year 2014/15 in Germany. Its goal was to analyze the causes for injuries occurring at school, on the way to school or during school activities. The study initially surveyed 5th grade students (10 to 12-year-olds) and tried to re-survey them every year until they reached the 10th grade in the school year 2019/20.

2.2 Sample design

The target population consists of students in 5th grade, who were enrolled in German secondary education public schools in the academic year 2014/15. As there is no list of the individual children, the selection was made on the school level. In order to achieve an adequate representation of all federal states in Germany as well as to account for the state-specific distribution of school types, a ***stratified random sample*** was drawn. The layers in the stratified random sample represent a combination of the federal state, school type, school size and level of urbanity. Some layers, especially those belonging to small federal states, received a higher selection probability, resulting in a ***disproportionate stratified sample***. For pragmatic reasons, all the children in the respective grade of the selected schools were

surveyed (cluster sampling). This sample design was developed with the kind support of GESIS – Leibniz Institute for the Social Sciences.

The gross sample of the first wave included 854 schools in eleven participating federal states. Almost a fifth of the schools contacted (17.3%) participated in the survey (net sample). Since primary school in three federal states (namely Berlin, Brandenburg, and Mecklenburg-West Pomerania) lasts for six years, schools in these states were surveyed from wave 3 (2016/17) onwards. In total, 14 out of 16 federal states in Germany are represented in the study (not included: Hamburg and Bavaria).

Since names, addresses, or other direct personal information were not collected from the students, it was not possible to implement any tracking mechanisms for participants from previous waves. This means that students who left the school or had to repeat a grade after a particular wave, were not surveyed in the subsequent years ("final dropouts"). Since the entire grade was surveyed in each year, students who skipped a grade could join the panel at a later time. The same is true for students who changed to a surveyed school after 5th (or 7th) grade. Students who were part of the original sample but did not participate in a wave (due to illness, lack of informed consent, etc.) could re-enter the panel as "temporary dropouts" in the subsequent wave(s).

The Scientific Use File contains all cases that were surveyed in waves 4, 5 and 6. For waves 1, 2, and 3, however, only those cases of students who participated at least in one of the subsequent waves (4, 5, or 6) could be included for legal reasons. Therefore, the provided number of cases in waves 1 to 3 is smaller than the number of students actually surveyed in these waves. Table 1 gives an overview of the differences between the numbers of surveyed and published cases.

Table 1: Number of surveyed and published cases of students

	wave 1	**wave 2**	**wave 3**	**wave 4**	**wave 5**	**wave 6**	**total**
surveyed cases	10,611	10,086	9,974	9,119	8,424	4,237	52,451
published cases	6,151	6,988	7,804	9,119	8,424	4,237	42,723
discrepancy	*-4,460*	*-3,098*	*-2,170*	*0*	*0*	*0*	*-9,728*

Table 2 provides an overview of the structure of cases included in the Scientific Use File.

Table 2: Number of published cases of schools, classes, and students

	wave 1 ***2014/15***	**wave 2** ***2015/16***	**wave 3** ***2016/17***	**wave 4** ***2017/18***	**wave 5** ***2018/19***	**wave 6** ***2019/20***	**total**
number of schools	115	121	132	133	124	69	135
number of classes	459	490	526	525	489	252	573
number of students	6,151	6,988	7,804	9,119	8,424	4,237	12,313

2.3 Data collection and questionnaire content

The students were interviewed in their classes within a school hour of 45 minutes by means of a ***self-administered questionnaire on a tablet PC*** (offline classroom survey). Incentives for survey participation were offered only at the school level, i.e., principals received a separate results report after each wave with statistics on various topics, trends over time, and comparisons of their own school means

to the means of the overall sample. Students were not given incentives. Prior to the survey days, students were given consent forms to be completed by their parents/guardians. On the survey day, before the questionnaires were processed, the consent forms were collected and checked by a trained interviewer, who was later also present when students filled out the questionnaires. The interviewers were also responsible for introducing the survey, guiding students through the process of the ID-code generation, offering support in dealing with the tablet PCs and responding to questions concerning the contents of the questionnaire. In the first part, the children were interviewed in depth about possible injuries. Subsequently, the children's exercise routines and nutritional behavior were assessed as well as their physical and mental health. In addition, data were collected concerning sociodemographic characteristics, their perception of the schooling situation, their family situation, and their personality.

The dataset was enriched with some structural features of the participating schools (e.g., type of school, federal state) as well as with information from the questionnaires filled in by the interviewers and the heads of the schools. A comprehensive overview of the topics and the questions that were asked in each of the survey instruments, can be found amongst the documentation materials on the website. The field times in each panel wave went from November to May of the subsequent year (e.g., November 2014 to May 2015), with a few exceptions.[1]

3. General Conventions

The compilation of the GUS Scientific Use File follows two general paradigms on how to edit the source data (i.e., the data collected via the tablet- or paper-and-pencil-questionnaires in the respective waves). There may be exceptions to these principles that are explicitly noted in the respective documentation materials.

The first and foremost paradigm in creating the GUS Scientific Use File is the one of ***unaltered data***. Wherever possible, the data editing procedures do neither change nor destruct the content of the original data. We consider this to be the basis for preserving the full research potential of the collected data. As a consequence, this means that the data in the Scientific Use File may contain implausible values, unless corresponding controls were already provided in the survey instrument. Only in rare cases, in which a variable requires the removal of clearly implausible information, these values are replaced by the special missing code *implausible value* (".c"). Whenever original data was edited (e.g., to harmonize questions or to construct variables with information from several questions), new variables were generated and the original information was left unchanged. Information on how the generated variables were constructed can be found in the respective Technical Report.

The second paradigm is to ***integrate the data*** as much as possible. The underlying assumption is that for a vast majority of data users it is far more comfortable to reduce already integrated data for a specific analysis as opposed to correctly compile the relevant information from scattered source data themselves. Therefore, the GUS dataset contains information from all the different questionnaires (students, heads of schools, interviewers) and from all panel waves. Data collected on a higher level

[1] In waves 4 and 5, appointments with schools could only be realized from December on. Furthermore, in certain years, appointments with some schools had to be postponed to June or July. Finally, the field phase in wave 6 (2019/20) had to be stopped in March 2020 due to the nationwide school closures related to the Corona virus (SARS-CoV-2) outbreak.

than that of the individual student was duplicated for all individuals in the respective group. As data from the interviewers was collected on the class level, the information was duplicated for all participants in the class. Similarly, data from the heads of the schools was collected on the school level and the information was duplicated for all participants in the school.

In addition to these basic principles, several other conventions were implemented during the data editing process to maximize data usability and understandability. These conventions are explained in detail in the following sections.

3.1 Variables

The naming conventions for variables in the GUS Scientific Use File aim to ensure maximum of consistency between the panel waves. The names also refer to different characteristics and thus allow the data user an orientation regarding the contents of the variables.

ID variables

- ***ID_s*** = school ID, five digits
- ***ID_c*** = class ID, five digits, last digit matches the class in the order they were interviewed in, classes are NOT consistent across panel waves[2]
- ***ID_t_g1*** = numerical expression of the original target ID as provided by the subject generated ID code (SGIC), 13 digits
- ***ID_t_g2*** = numerical expression of the target ID after record linkage procedures[3] were applied, 13 digits, we recommend using this as your target ID variable
- ***wave*** = number of panel wave, one digit

Prefixes

- ***s_*** = information about the school added from public sources and the sampling frame
- ***i_*** = data from interviewer questionnaire
- ***h_*** = data from head of school questionnaire
- ***t_*** = data from student questionnaire (target)

Sets of variables from the interviewer questionnaire

- ***i_np_...*** = variables about students, who did not participate in the survey (non-participants)
- ***i_ef_...*** = variables about students, who finished the questionnaire before the designated end of the survey time (early finish)

[2] For example, suppose that class "5b" of school 22720 was temporally the first class in this school to be surveyed in wave 1. Thus, it received the class code 22721. If the same class (now "6b") was surveyed second within the school in the following wave (because "6a" was now first on the schedule), it received class code 22722 in wave 2. Efforts were made to harmonize the class assignment, but this proved difficult and was often ambiguous because of actual changes in class structures. Therefore, the data were left unchanged.

[3] These procedures corrected errors in the ID codes due to memory problems or typos. With this technique, 5% of previously unmatched cases could be merged with no loss of precision. For more information on the procedures see Lipp, R., Stadtmüller, S., Giersiefen, A., Wacker, C. & Klocke, A. (2021). Using Record Linkage to improve matching rates of subject-generated ID-codes – A practical example from a panel study in schools. *Survey Methods: Insights from the Field*. https://doi.org/10.13094/SMIF-2021-00006

Modules in the student questionnaire

- ***as*** = accidents in school
- ***al*** = accidents during leisure time
- ***h*** = health
- ***s*** = school
- ***m*** = mobbing/bullying
- ***f*** = family
- ***c*** = social context
- ***l*** = leisure time
- ***p*** = personality

Please note: Socio-demographical variables have speaking variable names (e.g., year of birth = t_birth) to make them easy to identify. Furthermore, some generated variables also have speaking names, especially if they are constructs using information from multiple other variables (e.g., t_sleep_g1).

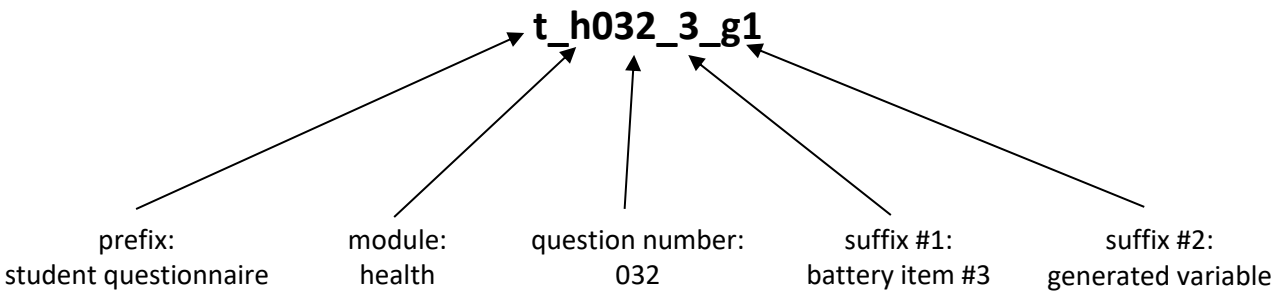
Suffixes

- ***_1, _2, … , _#*** = is used if variable is part of an item battery
- ***_to, _from*** = is used to distinguish between the way to school (_to) and the way home from school (_from).
- ***_w1, _w2, … , _w6*** = is used if a question changed during the course of the study; the number represents the wave in which the question was first asked in a particular way
- ***_h*** = indicates that it is a harmonized variable
- ***_s*** = indicates that it is a string variable
- ***_g1, _g2, … , _g#*** = indicates that it is a generated variable

Please note: Variables can have more than one suffix.

Example

The variable ***t_h032_3_g1*** contains the recoded information (values 0 = "no", 1 = "yes") on the third item ("type of appointments: tutoring") of a battery of eight items ("activities with fixed appointments in a normal week") from the student questionnaire of the sixth survey wave. The battery is the 32nd question of the health module (not corresponding to the question numbers in the questionnaire). In the Scientific Use File, it can be seen that a recoding of the values of the original variable t_h032_3 was made in two cases where the respondents did not report any tutoring appointments in the closed categories but stated to take English courses (“Englischkurs”) in the residual category.



3.2 Missing Values

The GUS dataset contains various missing codes to differentiate between different reasons for missing values. We distinguish three types of missing codes, which are summarized in Table 3 and described in more detail below.

Item nonresponse: The first type of missing codes occurs when a person has not (validly) replied to a question.

- The most common case of item nonresponse is *no answer* (.e). It is used when the respondent did not answer a question although he/she should have answered it.
- Some questions also allowed for active item nonresponse. Thus, participants were able to actively choose the option *don't know* (.d) or *can't assess* (.i). The latter was only used in the interviewer questionnaire.

Not applicable: The second type of missing codes occurs when an item does not apply to a respondent.

- The code *missing by design* (.b) is assigned if respondents in a subsample have not been asked the respective questions. This is usually the case if the respective question or instrument was not included in the survey in a particular wave.
- The code *missing by filter* (.a) is assigned if a question was not asked due to previous responses that made an answer obsolete (=filtered out questions).
- The code *not participated* (.h) is specific to the questionnaires for the heads of the schools. It is assigned if no questionnaire was filled in by the head of the school, even though the students participated in the survey in the respective wave.

Edition missings: The third type of missing codes is defined in the process of data preparation for the Scientific Use File.

- In general, generated variables are created using a specific logic or coding scheme. If the original variables from which the generated variable was derived did not contain enough information to generate a correct value, the code *not determinable* (.g) was used instead.
- In rare cases, variables also contain *implausible value* (.c). This is the case if the answers given have an impossible outcome (e.g., grades higher than 6). This could not be ruled out through a technical solution for all the variables. Unrealistically high values (e.g., 15 younger siblings) were not defined as implausible. It is up to the researchers to define the respective thresholds.

Table 3: Overview of missing codes

Code	Meaning
Item nonresponse	
.e	no answer
.d	don't know
.i	can't assess
Not applicable	
.b	missing by design
.a	missing by filter
.h	not participated
Edition missing	
.g	not determinable
.c	implausible value

3.3 Generated variables

Coding and recoding of open responses

At some points in the GUS survey instruments, there are so-called open questions where respondents can or should enter their answers as text. One example is information about the country of origin of their parents. The open text format allows respondents to specify anything they want. A practical way to deal with the resulting string information is to code and recode the information for further processing and later analyses. In general, ***coding*** describes the process of assigning one or more codes from selected category schemes to the string information, e.g. the classification of countries according to ISO 3166 numeric. The term ***recoding*** is used here to describe the process of assigning a code from an already presented closed answer scheme. This usually applies to semi-open question formats where respondents enter a text under the category "other", but which can be assigned ad hoc to one of the given closed answer categories. Therefore, the recoding does not define any new codes; the presented answer scheme of the respective question is not extended.

Derived scales and convenience variables

The GUS dataset also includes additionally generated variables which have been included for the convenience of researchers working with the data. They are marked with suffixes *_g1, …, _g#* and serve different functions:

- ***Structural variables***: They provide information on structural aspects of the dataset or subsample groups; e.g. the proportion of item nonresponse relative to the total number of questions in each wave or the number of surveyed students per class.
- ***Derived scales***: These variables combine information from multiple other variables into indices or scales; e.g. the *Family Affluence Scale* (FAS) or the social capital index.
- ***Harmonized variables***: Whenever a question or item battery underwent changes during the course of the study (even if only a slight ones), new variables were created with suffixes *_w1, …, _w6* to account for those differences. However, in order for researchers to still be able to easily analyze the data, harmonized versions containing information from all versions of the questions were integrated into the dataset as well. The resulting variables are marked with the suffix *_h*.
- ***Categorized variables***: During the course of the study, the responsible researchers developed useful categorizations for some of the metric variables in the dataset. For your convenience, these categorized variables were kept in the published dataset as well.
- ***Imputed variables***: In general, missing values are not imputed in the GUS dataset. However, some variables are included which contain imputed values from other observations. This is especially the case for time constant variables (like gender or year of birth), for which the values from previous (or subsequent) waves were used to reduce the number of missing values without making assumptions based on other students. Furthermore, when students were asked about information concerning the whole class or grade (like the duration of a school lesson), missing values were imputed based on the class mode of the answers of these questions.
- ***Calculated variables***: To provide further assistance when working with the dataset, some variables containing ready-made calculations are also included in the dataset. This is especially true for questions concerning time frames, e.g. the time students go to bed at night and the

time they wake up in the morning. The answers to these questions were pre-calculated into duration variables, e.g. sleep duration.

All generated variables are described in detail in the provided Technical Report on generated variables.

4. Special Issues

This last chapter contains notes on some issues that have occurred during the collection or editing of the GUS data and that should be taken into account when handling the data:

1. For data protection reasons, some of the questions asked in the study and also documented in the questionnaires, can neither be found in the dataset nor in the codebook. If the respective variables are of interest to you, please contact the FZDW. We will try to find other solutions for you to analyze these variables.

2. In wave 6, the questionnaire on the tablet PC contained a filter error that caused the variables t_s015_1, t_s015_2, and t_s015_3 to have misleading information or incorrect missing values. Therefore, these variables were coded as “missing by design” in this wave.