DigiPrim

UNIVERSITÄT BERN

Interfaculty Centre for Educational Research (ICER)

ÜGK / COFO / VECOF 2024 (HarmoS 4) Field Trial

Parent Questionnaire: Data Manual

Cooperation ÜGK / DigiPrim

Leo Röhlke, Jessica M. E. Herzing & Simon Seiler

October 2023, document version v1-0

Abstract: This data manual supports researchers who intend to use the data from the parent questionnaire of the ÜGK / COFO / VECOF 2024 (HarmoS 4) Field Trial 2022. This data manual contains documentation of the data editing for the scientific-use file and provides guides on how to use the scientific-use file. The data manual, therefore, functions as a documentation but goes beyond the documentation purposes of the respective data file.

Keywords: data documentation, survey data, data processing, large-scale assessment

Suggested Citation: Röhlke, Leo, Herzing, Jessica M. E., & Simon Seiler (2023). ÜGK / COFO / VECOF 2024 (HarmoS 4) Field Trial, Parent Questionnaire: Data Manual. Version 1-0. Bern: University of Bern, Interfaculty Centre for Educational Research. DOI: 10.48350/183639.

Acknowledgment: A special thanks goes to all cooperation partners, the ÜGK / COFO / VECOF project management teams data preparation, sampling, and context questionnaire, Dilan Cümen, Francesco Moser, and the ICER Team. Furthermore, thanks are extended to the ÜGK / COFO / VECOF steering group, participating parents, and school principals for their support and involvement.

Funding: This work was supported by the University of Bern in cooperation with BeLEARN, an initiative of the Canton of Bern, Switzerland. DigiPrim and the ICER thank the EDK (Swiss Conference of Cantonal Ministers of Education) for granting access to the ÜGK / COFO / VECOF sample.

Publisher: Interfaculty Centre for Educational Research (ICER)
Universität Bern
Fabrikstrasse 8
CH-3012 Bern

Web: https://www.icer.unibe.ch/
Contact: data.icer@unibe.ch

Contents

1 Introduction

This data manual is intended to assist researchers and other interest groups with the *parent questionnaire* data of the study ÜGK / COFO / VECOF 2024 (HarmoS 4) Field Trial 2022 (in the following referred to as ÜGKH4 field trial). The focus of this manual is to provide information on key aspects of the data structure, the data processing (cleaning), the survey, and the sampling design of the parent questionnaire. The manual further contains detailed information on recoding of missing values, coding procedures of open-ended answers, and generated variables.

On several occasions, the data manual refers to other files from the documentation of the ÜGK / DigiPrim cooperation. The documentation includes a general study description for the project DigiPrim, a codebook, a technical report, a report on scales and concepts, as well as further documentation on the items and implementation of the parent questionnaire. The documentation can be accessed via SWISSUbase.

If you encounter mistakes or if you have suggestions to improve the quality of this data manual or other documents, please do not hesitate to contact us at: data.icer@unibe.ch

1.1 Overview of the data

In general, the ÜGK / COFO / VECOF data consists of raw data, (partly) processed/cleaned data, and measurement instruments from the following sources (Data use concept ÜGK, p. 4):

- large-scale assessment data
- context data
- test items and tests
- questionnaires (e.g., students, parents, school principals)
- student lists and attendance lists
- sampling frame data
- test session protocols
- logfiles

The scientific-use file (SUF) of the parent questionnaire only contains information resulting from the parent questionnaire and some additional information from the sampling process (e.g., weights). Access to critical data, such as open-ended responses (text entries), un-aggregated data, as well as mode or item-specific studies, requires a special data usage request at Kosta HarmoS, the pertinent commission within the Swiss Conference of Cantonal Ministers of Education (EDK). For further information on the proceedings of special requests, please contact data.icer@unibe.ch

1.2 Respect rules of data usage

Any usage of the ÜGK / COFO / VECOF data, including the parent questionnaire data, requires the conclusion of a data usage contract. When working with any ÜGK / COFO / VECOF data, be aware of the data usage rules you have signed in the data usage contract. You are not allowed to publish any analyses that aim for or allow a direct comparison of cantons, schools, school principals, teachers, students, or parents. Any form of “rankings” using the ÜGK / COFO / VECOF data is strongly prohibited. Singling out school principals, teachers, parents, or students is not permitted as well.

For more information, please consult the data usage concept (Data use concept ÜGK) and the data usage agreement (when applying for the data at SWISSUbase).

1.3 Publications with ÜGK / DigiPrim data

When publishing results of any kind that are based on data from the ÜGKH4 field trial, it is required to give credit according to the data usage contract. Please identify the dataset used with its digital object identifier (DOI) in the data description section. In addition, any publications using data from the parent questionnaire must include the following acknowledgment:

“This paper uses data from the DigiPrim add-on study of the ÜGK / COFO / VECOF 2024 (HarmoS 4) Field Trial 2022 (DOI: [insert DOI of the data set here], further details can be found in Herzing et al., 2023).”

2 General conventions

The naming of data files and variables follows several conventions, which are described in the following sections.

2.1 File names

The file naming conventions are aimed at ensuring the consistency of file names within and across data sets. The naming of data files is summarized in Table 1.

For example, the file name “uegk24h4_ft_par_suf_v1-0.dta” indicates that the file contains data from the ÜGK / COFO / VECOF for the study year 2024 (24), conducted on the HarmoS 4 level (h4). The data is from the field trial (ft), from the parent questionnaire (par). The data was released as a scientific-use file (suf). It is the first main release of the data (v1), and there have been no minor updates of this first main data release (-0).

Table 1: Naming convention of file names.

Element	Definitions
uegk[…, 24,…] h[4, 8, 11]	Name of study [Year of study] HarmoS level [level]
[ms, ft, pft]	Indicator for study type: ms = main study ft = field trail pft = pre-field trial
[sq, par, sch]	Indicator for data source: sq = student questionnaire par = parent questionnaire sch = school site
[suf, intern]	Type of data set: suf = scientific-use file intern = internal data set
v[##]-[##]	Version First digits denote the main release number. A change in this number indicates major updates that affect the data structure (e.g., the release of imputed datasets). Updating syntax files may be necessary. Second digits indicate minor updates, which affect the content of cells or labels but not the data structure. Updating syntax files is mostly not necessary.

2.2 Variable names

The variable naming conventions are aimed at ensuring the consistency of variable names across data sets as well as documenting the origin of information for each variable. A variable name can consist of up to three elements:

- Prefixes indicate the *data source* of the information contained in this variable.
- A *semantic name* of the variable.
- Suffixes give *further details* about the information the variable contains or whether the variable was generated as part of the data preparation process.

Table 2: Naming conventions of variable prefixes.

Prefixes	Definitions
meta_	metadata (test administrator information)
sf_	sampling frame information

Table 3 (continued).

pq_	information from the parent questionnaire
sq_	information from the student questionnaire
spq_	information from the school principal questionnaire
st_	student tracking form
para_	information from paradata
ta_	test administrator information
m##_, sl##_; sh##_	test data based on subjects
pv_	plausible values
tp_	information from the test protocol
sc_	school-level information from the sampling frame
smp_	information from the sampling process

Table 4: Naming conventions of variable suffixes.

Suffixes	**Definitions**
_g	generated variable (mainly numeric variables)
_othertext	string variable from the open-ended answer given to «other»
_a	variable from an experimental split, group a
_b	variable from an experimental split, group b
_c to _d	variable from an experimental split
_coded	numeric variable generated from a string variable; open answers were coded based on a coding scheme
_flag	information about observation-level problems
_imp	imputed version of the variable
_impflag	information about imputed values

2.3 Missing values

We provide different missing codes for different events causing missing values. In general, we distinguish between missing codes indicating item nonresponse, inapplicability, and edition missings. Incorrect handling of missing information is a frequent source of error in empirical analyses. When working with the data, make sure that those codes are handled correctly by your statistical package. Alphabetic codes are used in the Stata® datasets (.dta), whereas numeric missing codes are used in comma-separated values files (.csv).

Table 5: Value label convention of missing values.

Alphabetic code	Numeric code	Missing description
Item nonresponse (item not answered due to the participant not being individually able or willing to).		
.a	-99	I don't know (default answer option or indicated in the open text field; test item not answered)
.b	-98	implausible/invalid (open answer or indicated number is implausible)
.c	-97	no answer / refused (missing, but valid answer in one of the following items)
.d	-96	breakoff (item not reached due to breakoff; if missing and no valid answer in the following items; test item not reached)
Not applicable (item not answered due to the questionnaire (design) and / or the participant's external circumstances)		
.f	-89	missing by design (filter)
.g	-88	missing by design (experimental split)
.h	-87	not administered (short questionnaire)
.i	-86	does not apply (default answer option or indicated in the open text field)
.j	-85	unspecific missing (every missing not fitting into another category)
Edition missings (item answered, but information removed/recoded into missing by the data editors)		
.n	-79	anonymized (sensitive information removed; access may be granted upon request)
.o	-78	not determinable (insufficient information to generate the variable value)
.p	-77	partial information (due to missing filter information)

2.4 Special conventions for variables from test data

Naming variables and missing codes corresponding to test items follow an alternative nomenclature. Please consult the corresponding documentation for this data generated by students.

3 Study design: Sampling

In May and June 2022, the ÜGKH4 field trial was conducted. The aim of the ÜGKH4 field trial was to assure the quality of the main study (in 2024). The target population of the ÜGKH4 field trial included approximately 85,000 pupils and 3,700 school sites (on a language-regional level). All cantons in Switzerland except the canton of Zug and the Rhaeto-Romanic-speaking region of the canton of Grisons (Graubünden) participated in the study. Schools were excluded that either teach based on foreign programs or not in any Swiss national language. Most students in special needs schools were also excluded from the target population.

As a full population survey was not aimed for, school and student samples were drawn. A two-stage sampling procedure was used in all cantons. First, school sites were sampled, and then, students (of HarmoS level 4) in the selected schools were randomly selected. For further details on the sampling, please consult the DigiPrim study description.

The parent sample is based on the student sample. All parents of the selected students were invited to participate in an online survey. Parents were invited to participate in the study by means of an invitation post card that the students brought home to their parents after they had participated in the assessment. There was one reminder card for parents handed out to students who were on the list to have participated in the assessment.

Due to the complex field coordination, it is possible that some parents of sampled students who, for any reason, did not participate in the student assessment of the ÜGKH4 field trial, participated in the parent questionnaire. However, because the children of these parents (students) had originally been eligible for the test, the scientific-use file contains all participating parents irrespective of their children's participation status. Based on the survey participation status of the parents, a probability weight adjusting for nonresponse mechanisms (*smp_w_nrapqw*) as well as a variable indicating the primary sampling unit (*smp_psu*) and the strata (*smp_strata*) was generated and included in the SUF. Section 6 offers further guidance on how to use these variables in empirical analyses.

4 Data structure

The parent questionnaire data has a cross-sectional data structure with one row corresponding to one interviewed parent (identified by *mergeid_parents*). Only one parent / legal guardian per sampled student was allowed to participate in the survey (only one questionnaire access code was provided, however, the questionnaire could be accessed multiple times). Parents / legal guardians decided individually which parent / legal guardian would participate in the survey in the case of two-parent families (if questions were asked concerning this aspect, the test administrators and the support

hotline were instructed to recommend that the parent with whom the child spends the most time of the week should complete the questionnaire, and hence, this selection was nonrandom).

The first part of the main parent questionnaire included questions on demographics and the family's socioeconomic background, as well as on children's attendance of leisure groups. The second part of the questionnaire focused on aspects of digital media use, resulting from the collaboration between ÜGK / COFO / VECOF and the research project DigiPrim. These questions addressed different aspects of digital inequalities and media socialization, from device possession over children's digital media use to related parental attitudes.

Parents could select one out of sixteen languages prior to the start of the online questionnaire. Four language options were linked with the main (long) version of the questionnaire (German, French, Italian, and English), while the remaining twelve languages led to a short version of the questionnaire (covering only certain questions regarding the demographics and the family socioeconomic background). These additional languages were Spanish, Portuguese, Bosnian/Croatian/Serbian, Albanian, Arabian, Chinese (simple), Dari, Polish, Russian, Tamil, Tigrinese, and Turkish.

Most questions on family sociodemographic characteristics were posed identically in the student questionnaire. This applies to home language(s), generational status, number of siblings, potential birthday wishes, house type and surroundings, home facilities and endowment, parental employment, and occupation.

5 Data processing

Data processing of the parent questionnaire data was performed using the statistical software Stata (version 16.1; StataCorp, 2019). Data manipulations beyond usual data cleaning steps (e.g., removing obvious typos in numeric open-ended questions) are documented for each variable in this section.

5.1 Unit nonresponse

In general, the dataset contains all respondents who gave a valid answer to at least one question. Parents who did not participate (including those who accessed the online questionnaire but did not answer any questions) are not included in the SUF.

5.2 Item nonresponse

Item nonresponse in the parent questionnaire has four possible causes (see section 2.3). Skipping and, therefore, not answering individual questions was possible in the online questionnaire, with no warning message appearing. Furthermore, questionnaires did not have to be completed or submitted (using the "submit" button at the end of the questionnaire) in order to be included in the SUF. It was not possible to identify with certainty whether an individual item was seen or read by the respondents. Missing codes for variables, therefore, indicate consistently whether an item without any (valid or

invalid) answer was followed by any form of answer in the remainder of the questionnaire ("No answer / refused") or if a breakoff could be plausibly assumed because all following items were missing ("breakoff"). The variable *part_status* provides information on the participation status of each respondent according to the AAPOR standard (American Association for Public Opinion Research [AAPOR], 2017), indicating the (grouped) individual percentage of validly answered questions. For details on the coding of implausible answers, please refer to section 5.6.

5.3 Recoding of selected missing values

Because skipping questions in the online questionnaire was possible, it is not always certain whether an empty open-ended text field or an item that was not selected represents a case of item nonresponse or whether respondents attempted to convey information (e.g., leaving fields empty instead of entering "0"). In the following, we discuss these issues and the attempts to account for these uncertainties in the data processing individually for all affected variables in the SUF. The data processing for these cases is aimed at providing user-friendly, ready-to-use variables while avoiding unnecessarily strong assumptions.

hhmemb_moth, hhmemb_moth_raw, hhmemb_fath, hhmemb_fath_raw, hhmemb_parenttwo, hhmemb_parenttwo_raw, hhmemb_bro, hhmemb_bro_raw, hhmemb_sis, hhmemb_sis_raw, hhmemb_ofam, hhmemb_ofam_raw

Because children around the age of eight years cannot live without a parent or legal guardian, all observations without answers in any of the household member categories were treated as item nonresponse category "no answer / refused." In all remaining cases where at least one household member category was selected, the categories not selected were treated as "true" zeros instead of missing values. Data users who prefer a different strategy are provided with a "raw" version of the variables (suffix "_raw").

nsib_older, nsib_older_raw, nsib_younger, nsib_younger_raw, nsib_same, nsib_same_raw

Because the answer categories were implemented as open-text fields, some respondents filled in zeros, while others left certain fields empty. To provide data users with ready-to-use variables, some assumptions were made in the data processing:

1) empty answers were recoded as zeros if there was a valid answer in any of the other categories of the sibling question (e.g., a respondent indicated four older siblings and left the field for younger and same-age siblings empty: The latter were recoded as zeros).
2) if a respondent indicated brothers or sisters living with the child in the previous question, and all variables regarding the number of siblings were empty, these were treated as missing values instead of zeros.
3) if no answer was given regarding the number of siblings, and in the household members question, valid answers were given but no brothers or sisters in the household

were indicated, the “number of siblings” variables were all set to zero. The last assumption implies that it is more likely that parents skipped the question regarding the number of siblings after having indicated no in-household siblings than the scenario of no in-household siblings but out-of-household siblings, which were then not indicated in the respective question.

Data users who prefer another data cleaning approach are provided with a “raw” version of the “number of siblings” variables (suffix “_raw”).

empstat_moth_full, empstat_moth_part, empstat_moth_trai, empstat_moth_home, empstat_moth_retir, empstat_moth_noemp, empstat_moth_jseek, empstat_fath_full, empstat_fath_part, empstat_fath_trai, empstat_fath_home, empstat_fath_retir, empstat_fath_noemp, empstat_fath_jseek

If at least one status was selected, the remaining missing items were recoded as zeros.

leisure_sportsclub, leisure_danceschool, leisure_scouts, leisure_musicschool, leisure_theatergroup, leisure_craftgroup, leisure_churchgroup

As opposed to having no household members, doing none of the activities in the item battery is plausible. Therefore, if none of the leisure activity items were selected, but either the preceding or following question contained at least one (valid or invalid) answer, the leisure activities were all set to zero. Only if the previous and the following questions were both empty, and no leisure activity was indicated, the latter were set to missing.

devicetype_smartphone, devicetype_laptop, devicetype_pc, devicetype_radio, devicetype_printer, devicetype_player, devicetype_fixconsole, devicetype_mobileconsole, devicetype_digicam, devicetype_tablet, devicetype_kidspc, devicetype_streaming, devicetype_assist, devicetype_vr, devicetype_musicbox, devicetype_3dprinter, devicetype_pi, devicetype_mediaplayer, devicetype_fitness, devicetype_smartwatch, devicetype_ereader, devicetype_scanner

Several parents exclusively chose the "1 (yes)" option without ever selecting the "0 (no)" option. To minimize redundant missing values in variables, any missing values were transformed to "0 (no)" if none of the devices were assigned a value of 0 by the respective parent. This assumption is based on the notion that these values indicate non-possession rather than authentic missing entries. If "0 (no)" had been selected at least once within this item battery, missing values were retained. Data users may consider reassigning some of these missing values as zeros.

5.4 Open-ended answers

langhome_other, langtwowhich_other

Open answers were internally categorized in order to accurately define the generated variables *homelang_g* and *langhome_g*. The main categories of the original language variables

were not affected by the coding. Information on languages other than the main categories resulting from open answers is not included in the SUF.

hhmemb_other

If parents manually indicated household members that were covered by the regular item categories, these categories were recoded as “1 (yes)”. Most open answers provided irrelevant information, e.g., regarding pets or occasional guests. In some cases, respondents indicated that other household members were present only part-time or that the household situation was more complex regarding further individuals. This type of information was considered to be of minor relevance and is therefore not included in the SUF.

educstring_moth, educstring_fath

The short questionnaire was targeted at parents with a high probability of having completed formal education abroad. Therefore, educational levels were surveyed via open-ended text fields only. The resulting strings (n = 113 for mothers and n = 114 for fathers) were translated into German and then coded manually to International Standard Classification of Education (ISCED) 2011 levels by a student assistant. ISCED2011 levels were then collapsed in accordance with the scale used in the long version of the questionnaire (*edu_moth / edu_fath*): ISCED levels 1 and 2 correspond to scale value 1 (lower secondary education), ISCED levels 3 to 8 correspond to the respective scale values, and ISCED level 0 was recoded as missing value “does not apply.” The resulting variables combining information from the long and short questionnaires are *edu_moth_g* and *edu_fath_g*.

job_fath, job_moth

Occupations were surveyed using open-ended text fields with an autocomplete function when entering text into the text field (drop-down text box). The open answers were then recoded into occupations according to the International Standard Classification of Occupations 2008 (ISCO-08) whenever possible automatically. The manual coding was performed separately for regular occupational strings, obvious missing values, and answers that did not contain occupations but mentioned locations, brands, industries, or activity descriptions. The latter were regularly coded and included in the main ISCO-08 variables (*job_fath_isco08_g, job_moth_isco08_g*). However, the respective values are flagged by the generated variable *job_fath_flag_g* (*job_moth_flag_g*). Furthermore, there was an attempt to manually code as many responses as possible by coding locations, brands, industries, or activity descriptions (*job_fath_isco08_exp_g, job_moth_isco08_exp_g*). Detailed documentation of the coding procedure can be found in the documentation for the parent questionnaire on SWISSUbase.

leisure_othergroup

Two external coders from one of the collaborating institutions independently categorized the open answers (n = 228; Cohen’s kappa κ = 0.79). The final codes were drawn as a random

mixture from both coders. If parents manually indicated leisure groups that were covered by the regular item categories, these categories were recoded as “1 (yes)”. The generated variable *leisure_othergroup_g* contains the remaining open answers indicating any groups other than the ones covered by the regular item categories (e.g., language classes or physical activities).

devicetype_othertext

If parents manually indicated possession of devices that were considered by the data editors to be equivalent to one of the main categories, the respective categories were recoded as “1 (yes)”. Three new variables were generated, capturing information from the remaining open answers (see section 5.5).

intuse_othertext

If parents manually indicated using the Internet in a way that was considered by the data editors to be equivalent to one of the main categories, the respective categories were recoded as “1 (yes)”. Three new variables were generated, capturing information from the remaining open answers (see section 5.5).

5.5 Generated data

Several variables in the dataset do not represent original parental responses to questions but are provided by the data editors after manipulating information from existing variables. In most cases, this is the result of anonymization concerns, where the original information was considered sensitive, e.g., open answers. Generated variables are marked with the suffix “_g”.

age_g

This variable contains the age of students on the day of their ÜGK test session in full years (additional months and days were truncated). It was generated using parent-reported information on the date of birth and test dates provided in the student tracking form. A plausibilization procedure was applied (see the section on plausibilization) because of deanonymization concerns regarding children who are either very young (younger than 7 years) or very old (10 years or older) compared to other children at the school level HarmoS 4, the age variable was top- and bottom-coded.

age_month_g

This variable contains the exact age of students on the day of their ÜGK test session, measured in months. It was generated using parent-reported information regarding the date of birth and test dates as provided in the student tracking form. The same plausibilization and anonymization procedures applied to variable *age_g* were also applied to this variable. Moreover, *age_month_g* was anonymized if it fell below 84 months (equivalent to 7 years) or

exceeded 120 months (equivalent to 10 years). The necessity for more stringent anonymization compared to age_g (where these extreme values were top-/bottom-coded instead) was driven by the smaller cell sizes associated with this variable. Data users have the option to utilize the variable *age_g* to identify cases that have undergone additional anonymization, enabling them to tailor individual solutions according to their specific data analysis needs.

langhome_g

This variable contains the first language spoken at home, combining information from the variables *langhome* and *langhome_other.* Open answers in *langhome_other* were coded manually. Languages other than Swiss German, Standard German, French, Italian, and "Ticinese and Grisons dialect" were summarized under "other" as part of the anonymization efforts (including Rhaeto-Romanic due to low numbers).

homelang_g

This variable contains information about the home languages in relation to the school language as defined in PISA 2015 (and ÜGK 2016), combining information from the variables *smp_region, langhome, langhome_other, langtwo, langtwowhich*, and *langtwowhich_other.* Open answers were coded manually. Contrary to the procedure in ÜGK2016, missing values were not imputed. Basic coding rules and theoretical reasoning can be obtained under https://www.ÜGK / COFO / VECOF-schweiz.ch/wp-content/uploads/2019/06/ÜGK / COFO / VECOF16__Technical-appendices.pdf (retrieved August 2023), section 1.2.

language_CHD, language_D, language_F, language_IT, language_ticinoGrisons, language_rhaeto, language_PT, language_ALB, language_other

These variables contain information on whether a specific language is spoken at home according to the respondent, irrespective of whether the language is spoken as a first or second language. The variables combine all available information from the variables *langhome, langhome_other, langtwo,* and *langtwowhich_other,* including manually coded open-text answers. The variables *language_rhaeto, language_PT* and *language_ALB* were anonymized in the SUF due to low case numbers (risk of de-anonymization). The variable *language_other* indicates these cases without specifying the exact language.

immig_pisa_g

This variable contains the immigration status as defined in PISA 2015 (and ÜGK 2016), combining information from the variables *cob_child, cob_moth*, and *cob_fath*. Contrary to the procedure in ÜGK 2016, missing values were not imputed. Basic coding rules and theoretical reasoning can be obtained under https://www.ÜGK / COFO / VECOF-schweiz.ch/wp-content/uploads/2019/06/ÜGK / COFO / VECOF16__Technical-appendices.pdf, section 1.3. The variable was generated for cases in which information was provided for at least one parent (missing information for the second parent / legal guardian did in general not lead to a missing value in *immig_pisa_g*). If no information on the child was provided regarding birth

in Switzerland (*cob_child*), this only caused a missing value in those cases when no parents were known to being born in Switzerland, since the information on the child was necessary to distinguish between first- and second-generation immigrants. If one parent was known to be born in Switzerland, missing values in the child variable (*cob_child*) did not lead to a missing value, because these cases are always considered as “native Swiss”, regardless of the child’s birthplace.

enrolyear_g

This variable contains the year of the child’s initial school enrolment in Switzerland. Because of low cell sizes and resulting risks of deanonymization, the variable was top- and bottom-coded.

nsib_older_g, nsib_younger_g, nsib_same_g

These variables contain the number of the child’s older, younger, and same-age siblings. The variables were top-coded because a very high number of siblings can enable re-identification.

nsib_total_g

This variable was generated by summing up the total number of siblings reported in the variables *nsib_younger, nsib_older, and nsib_same*. It was not directly asked in the questionnaire to reduce the length of the questionnaire. Missing values in one of the variables were counted as zeros. The variable was top-coded because a very high number of siblings can enable reidentification.

yard_g, trash_g, grafi_g

These variables result from experimental survey items which aimed at assessing an alternative way to measure parental socioeconomic background based on interviews with young children (in the student questionnaire). The original items (*housetype, htype_**) presented stylized images of different types of houses, which both students and parents were asked to compare to the house in which they live and decide which image came closest to their own house. The variables listed here were generated by combining information from several variables from the parent questionnaire (identical questions, only the type and size of houses on the images varied). The variable *yard_g* is the combination of the variables *htype_onefam_ya, htype_bigfam_ya, htype_mfam_ya, htype_row_ya, htype_block_ya,* and *htype_mstory_ya.* The variable *trash_g* is the combination of *htype_onefam_tr, htype_bigfam_tr, htype_mfam_tr, htype_row_tr, htype_block_tr,* and *htype_mstory_tr*, and the variable *grafi_g* is generated from the variables *htype_onefam_gr, htype_bigfam_gr, htype_mfam_gr, htype_row_gr, htype_block_gr,* and *htype_mstory_gr.*

empstat_moth_g, empstat_fath_g

These variables contain the aggregated version of the variables *empstat_*_full, empstat_*_part, empstat_*_trai, empstat_*_home, empstat_*_retir, empstat_*_noemp,* and

empstat_*_jseek (the asterisk represents either “moth” or “fath”). Some of these employment situations are very rare (e.g., retirement), which increases the risk of deanonymization. A hierarchy was generated to deal with multiple answers (e.g., respondents or their partners could indicate to be part-time employed while participating in further training): Full-time employment (*empstat_*_full*) was prioritized over part-time employment (*empstat_*_part*) and part-time over non-employment (*empstat_*_trai, empstat_*_home, empstat_*_retir, empstat_*_noemp, empstat_*_jseek*).

edu_moth_g, edu_fath_g

These variables harmonize parental responses from the open-ended question in the short questionnaire version (*educstring_moth* and *educstring_fath*: manually categorized) and the closed-ended question in the long questionnaire (*edu_moth* and *edu_fath*).

job_moth_isco08_g, job_fath_isco08_g

These variables result from the occupational coding procedure described above. They contain 4-digit codes corresponding to the ISCO-08 standard.

job_moth_isco08_str_g, job_fath_isco08_str_g

These variables contain a string version of the variables *job_moth_isco08_g* and *job_fath_isco08_g* as a service for data users not interested in the value labels. Using them may help avoid problems with ISCO-08 major group “0” in the numeric version of the variables.

job_moth_isei_g, job_fath_isei_g

These variables were automatically generated from the respective ISCO-08 variables (*job_moth_isco08_g* and *job_fath_isco08_g*). They represent the International Socio-Economic Index of Occupational Status (ISEI-08; Ganzeboom, 2010b) of the father or mother. They were generated using the “isei()” command from the Stata package “iscogen”, version 1.0.6 (Jann, 2019). The transformation is based on Ganzeboom’s (2010a) file.

job_moth_isei2_g, job_fath_isei2_g

These variables were automatically generated from the respective ISCO-08 variables (*job_moth_isco08_g* and *job_fath_isco08_g*). They represent the International Socio-Economic Index of Occupational Status (ISEI-08; Ganzeboom, 2010b) of the father or mother. They were generated using the “iseisps()” command from the Stata package “iscogen,” version 1.0.6 (Jann, 2019). The transformation is based on Ganzeboom’s SPSS syntax “isqoisei08.sps” (publicly accessible under http://www.harryganzeboom.nl/isco08/).

job_moth_flag_g, job_fath_flag_g

These variables indicate whether ISCO-08 and ISEI values in variables *job_moth_isco08_g, job_moth_isco08_str_g, job_moth_isei_g, job_moth_isei2_g, job_fath_isco08_str_g,*

job_fath_isco08_g, job_fath_isei_g, and *job_fath_isei2_g* are based on open answers referring to locations, activities or industries instead of occupations. The average reliability of flagged values is lower compared to the rest of the values. For more information, see section 5.4.

job_moth_isco08_exp_g, job_fath_isco08_exp_g

These variables contain additional ISCO-08 codes resulting from an experimental coding procedure developed as an additional branch of the main procedure. The experimental coding aimed at extracting as much information from the open answers as possible. Open answers that were coded as missing values in the main procedure were coded again by the same coders. The resulting variables are included in the SUF to support imputation routines. Using the variables directly for analyses is not advised.

job_moth_isco08_str_exp_g, job_fath_isco08_str_exp_g

These variables contain a string version of the variables *job_moth_isco08_exp_g* and *job_fath_isco08_exp_g* as a service for data users not interested in the value labels. Using them may help avoid problems with ISCO-08 major group “0” in the numeric version of the variables.

job_moth_isei_exp_g, job_fath_isei_exp_g

These variables were automatically generated from the respective ISCO-08 variables (*job_moth_isco08_exp_g* and *job_fath_isco08_exp_g*). The transformation was based on Ganzeboom’s (2010a) file. Please note that the variables result from an experimental occupational coding procedure (see description of *job_moth_isco08_exp_g* and *job_fath_isco08_exp_g*).

job_moth_isei2_exp_g, job_fath_isei2_exp_g

These variables were automatically generated from the respective ISCO-08 variables (*job_moth_isco08_exp_g* and *job_fath_isco08_exp_g*). The transformation was based on Ganzeboom’s (2010) SPSS syntax “isqoisei08.sps” (publicly accessible under http://www.harryganzeboom.nl/isco08/). Please note that the variables result from an experimental occupational coding procedure (see description of *job_moth_isco08_exp_g* and *job_fath_isco08_exp_g*).

leisure_othergroup_g

This variable indicates child attendance in leisure groups other than the ones covered by the regular items. It results from the manual coding procedure described in the section “Open-ended answers”.

devicetype_kidsaudio_g

This variable indicates household possession of an audio player specifically designed for children. This kind of device was not part of the closed-ended answer options but reported

by a substantial number of parents in the open text field “Other,” which is why the information is provided in the SUF as a generated variable.

devicetype_kidswatch_g

This variable indicates household possession of a smartwatch specifically designed for children. This kind of device was not part of the closed-ended answer options but reported by a substantial number of parents in the open text field “Other,” which is why the information is provided in the SUF as a generated variable.

devicetype_other_g

This variable indicates household possession of any kind of digital device other than the ones specified in any of the other variables. A substantial number of parents reported a variety of devices in the open text field “Other.” To allow data users to include this information when calculating, e.g., the number of different digital devices in the home, the information is provided in the SUF as a generated variable.

intuse_jobrelated_g

This variable indicates parental (respondents’) Internet use in the context of their job, e.g., virtual meetings. A substantial number of parents reported a variety of ways to use the Internet for work in the open text field “Other,” which is why the information is provided in the SUF as a generated variable.

intuse_furtheredu_g

This variable indicates parental (respondents’) Internet use in the context of further education, e.g., online language classes. A substantial number of parents reported a variety of ways to use the Internet for education in the open text field “Other,” which is why the information is provided in the SUF as a generated variable.

intuse_other_g

This variable indicates parental (respondents’) Internet use for any purpose other than the ones specified in the other variables. To allow data users to include this information when calculating, e.g., the number of different ways of using the internet per individual, the information is provided in the SUF as a generated variable.

eval_jobflag_g

This flag variable indicates that the respondent considered his or her employment situation to be not adequately covered by the categories offered in the variable *empstat_**. Hence, data users may consider setting the observation(s) to missing when working with the employment status variable. This statement was given in the open-ended text field for further comments at the end of the survey, which is not part of the SUF.

eval_famflag_g

This flag variable indicates that the respondent considered his or her family constellation to be not adequately covered by the categories offered in the questionnaire (not further specified). This statement was given in the open-ended text field for further comments at the end of the survey, which is not part of the SUF.

eval_unknflag_g

This flag variable indicates that the respondent cautioned data users regarding the validity of his or her statements about his or her partner (the father or mother of the child, e.g., his or her income and occupation). Hence, data users may consider setting the respective information to missing for this observation. This statement was given in the open-ended text field for further comments at the end of the survey, which is not part of the SUF.

5.6 Plausibilization

age_g, age_month_g

If the student age indicated by the parents fell below 6 years (72 months) or exceeded 10 years (120 months), the parent response was replaced with information from the sampling frame. In multiple of these implausible cases, the parent response differed from the sampling frame information only in terms of the birth year, suggesting a parental data input error. By using the sampling frame information in this way, the most extreme outliers were corrected. For inconsistencies between parent and sampling frame information within the plausible range, it was uncertain which of the two values was correct, so parental responses were retained.

enrolyear, enrolyear_g

Some parents indicated a year of school enrolment in Switzerland which was earlier than the child's year of birth. In other cases, the supposed school enrolment occurred when the child was extremely young according to the parent-reported child's day of birth. Because this inconsistency can result both from errors in the age of the child and the enrollment year variables and it is difficult to identify an accurate "plausibility threshold", these inconsistencies were not corrected by the data editors. Data users are advised to address this problem individually when using the enrolment year variable in their analyses.

5.7 Anonymization principles

The ÜGK / COFO / VECOF surveys cover information about students, context persons (e.g., parents, teachers, or school principals), and institutions (e.g., school forms). Based on the data use concept (Data use concept ÜGK), the data processing and data securing infrastructure are obliged to handle individual (or personal) data with the utmost consideration. In this regard, careful practice must be applied when ÜGK / COFO / VECOF data is distributed, as malicious handling may have legal consequences. Therefore, the disseminated ÜGKH4 field trial data needs to be anonymized without

preventing certain research projects or education monitoring from being conducted. The aim is to reduce the likelihood of three disclosive scenarios: spontaneous recognition, singling out unusual cases, and re-identification by matching information.

To ensure the protection of the data, we have established an approach with interlocks five different security measures to guarantee the best possible protection for the data:

1. Institutional (data is only disseminated to researchers through a Swiss data archive or the institution which oversees the data protection and there is differential privacy of data sets)
2. Legal (data users must sign a data use agreement)
3. Informational (users are sensibilized to data protection and misuses through the documentation)
4. Technical (more sensible data is only available under special request and restrictions)
5. Statistical (modifying data and render them to make them more anonymous)

The institutional security measures of the data archive only allow authenticated researchers to use the data, e.g., researchers must be affiliated with a research institution, must describe the use of the data, and delete the data at the end of the license period. By signing the data use agreement, researchers/data users are agreeing to the institutional security measures and to use the data only for research purposes. Within the data use agreement as well as in the documentation, the data users are informed about data protection issues and data misuse.

When downloading the data from the data archive to the researcher's local workstation, the data leaves the secured area of the archive and hence, the reduction of physical protection needs to be balanced out with additional anonymization procedures. For this purpose, a statistical security measure approach is used meaning that the data is modified by aggregating, recoding, or removing some information (e.g., variables or deleting whole observations).

There are automized procedures to anonymize data (e.g., noise addition or permutation to name a few), however, for the current data only non-perturbative (altering data values by adding noise, see Oganian, 2011), global modification methods (looking at the distribution of a variable for all observations, whereas a local method would look at the values of individuals, see Hundepool et al., 2012) were used. Consequently, data suppression involving removing or suppressing certain variables or attributes from the dataset that could be used to re-identify individuals, such as names, social security numbers, or other unique identifiers was used. The following data adjustments were performed for quasi-identifiers considering the size of the sample but also the size of the total population (a threshold of a sampling fraction of <.05 was used).

Aggregating values can be differentiated between global recoding and local recoding. Global recoding means aggregating whole answer categories, so the information given in the variable is reduced. For example, the specific country of birth, e.g., "Ghana," is aggregated into a global category such as "Foreign Country," and so forth. Local recoding would mean replacing one or more values within an unsafe combination with a missing value.

Topcoding values are one form of aggregation where the top end of a scale is affected. For example, instead of providing the exact number, the top numbers are truncated. This could be the number of students attending a certain school (e.g., 50, 100, 150, …, 300 and more).

Bottomcoding values are one form of aggregation where the bottom end of a scale is affected. For example, instead of providing the exact number, the bottom numbers are truncated. This could be the year of birth of students (2011 and earlier, 2012, 2013, 2014).

Aggregating and top- / bottom-coding in combination can also be used; for example, year of birth can be aggregated to age groups with an open highest and lowest category size (i.e., 7 years and younger, 8, 9, 10 years and older).

Transforming values and removing total values by, e.g., generating a new variable which is generated by means of two later removed variables (i.e., by percentualizing). For example, instead of offering the total amount of lessons provided for children with special needs per school, the number of lessons is divided by the total amount of students, generating the proportion of lessons per student.

Removing values because the information provided is too sensible for the SUF. For example, regional data on a postcode level was removed from the data, as it would make the identification of individuals easier, especially in combination with other characteristics.

The *k-anonymity* was calculated before and after the manual inspection of the data. After the data anonymization, the k-anonymity did not indicate any sensitive cases.

For the generation of the SUF a so-called purging approach was used, meaning, that both the original and the modified variables are kept in the SUF version. The values of the original variables are overwritten with a specific missing code and hence, the original variable is purged. Consequently, all variables can be accessed in the SUF, but their content is not published in the SUF.

6 Sample and nonresponse adjustments

6.1 Nonresponse adjusted sampling weight

Student sampling was based on a two-stage sampling procedure as described in the DigiPrim study description. Schools educating students on HarmoS level 4 represent the primary sampling unit (PSU) on the first stage. Within these PSUs, eligible students visiting classes in HarmoS level 4 were randomly selected (second stage). School sampling in the first stage was stratified along several criteria (e.g., school size, language region).

In addition, parental participation in the study was voluntary, resulting in potentially selective unit nonresponse. The SUF of the parent questionnaire data, therefore, contains a nonresponse adjusted sampling weight (*smp_w_nrapqw:* parent questionnaire weight, nonresponse adjusted). These weights were calculated in an equivalent approach as described by Verner and Helbling (2019). Please note that only a limited set of nonresponse predictors was available to calculate the nonresponse adjustment factor.

6.2 Recommendations for addressing the complex survey design

Data users are generally advised to utilize the non-response adjusted sampling weight regardless of the type of analysis, to account for the unequal sampling probabilities. However, relying solely on the sampling weight slightly underestimates the standard errors. To address this problem, it is advisable to include two additional variables that reflect the complex survey design in analyses where inferential statistics are calculated: The SUF includes a clustering variable (*smp_psu*) indicating the primary sampling units and a stratum variable (*smp_strata*), which is equivalent to the language region (*smp_region*). Accounting for these two variables in addition to the sampling weight corrects for the underestimation of standard errors when inferential statistics are calculated, yielding more accurate, conservative estimates of the standard errors. For example, in the statistical software Stata (v16.1; StataCorp, 2019), the complex survey design can be addressed by means of Stata's suite of survey data commands (svy).

7 Further information

Please visit our web portal for further information and comprehensive documentation resources, such as:

- the study description,
- the questionnaires (in numerous languages and with screenshots of the implemented questions),
- conceptual reports (e.g., on constructs used in the surveys and the test development),
- technical reports (e.g., on survey methodology, unit and item nonresponse),
- and the codebooks (containing information on all variables).

For further support, please contact data.icer@unibe.ch

7.1 Recommendations

Please examine the data critically when you work with it. While the different project teams invested a lot to ensure the integrity of the provided data, the latter cannot be guaranteed. To achieve a precise assessment of the information contained in the variables, please consult the questionnaire documentation and keep in mind how the data was generated and processed.

Finally, when working with the data, please ...

- recode missing values adequately to your statistical software.
- read the documentation materials that can be downloaded.
- do not report unweighted distributions as substantial findings.
- note that selective unit nonresponse of parents can lead to biases that may not be mitigated by the weights.

- if you encounter problems or errors in the data, please contact data.icer@unibe.ch.

7.2 Further resources

A scale for parental socioeconomic status that was used in the context of ÜGK / VECOF / COFO in 2016, can be conveniently generated using the software Stata and the publicly available syntax under: https://boris.unibe.ch/152698/1/pisa-inspired-ses-imputation-1-2.do. It is based on the variables HISEI08 (based on the highest parental ISEI-08 available in the SUF), the number of books in the household (variable books5) and the highest level of parental education (based on the parental education variables available in the SUF). The scale is inspired by a scale that was developed in the context of PISA 2015. The syntax file applies a single imputation procedure to address missing values.

References

American Association for Public Opinion Research. (2017). *Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys: 9th edition.* AAPOR.

Ganzeboom, H. B. (2010a). *International standard classification of occupations ISCO-08 with ISEI-08 scores*. http://www.harryganzeboom.nl/ISCO08/isco08_with_isei.pdf

Ganzeboom, H. B. (2010b). *A new International Socio-Economic Index (ISEI) of occupational status for the International Standard Classification of Occupation 2008 (ISCO-08) constructed with data from the ISSP 2002–2007: With an analysis of quality of occupational measurement in ISSP*. Paper presented at Annual Conference of International Social Survey Programme. Lisbon.

Herzing, J. M. E., Röhlke, L., & Erzinger, A. B. (2023). *DigiPrim – Digitalization in Swiss schools and its impact on educational trajectories: Study Description*. Version 1-0. Bern. University of Bern, Interfaculty Centre for Educational Research. https://doi.org/10.48350/183647

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., & Wolf, P.-P. de. (2012). *Statistical Disclosure Control*. John Wiley & Sons.

Jann, B. (2019). *ISCOGEN: Stata module to translate ISCO codes*. http://ideas.repec.org/c/boc/bocode/s458665.html

Oganian, A. (2011). Multiplicative noise for masking numerical microdata with constraints. *SORT-Statistics and Operations Research Transactions*, 99–112.

StataCorp. (2019). *Stata Statistical Software: Release 16* (Version 16.1) [Computer software]. StataCorp LLC. College Station, TX.

Verner, M., & Helbling, L. A. (2019). *Sampling ÜGK 2016.: Technischer Bericht zu Stichprobendesign, Gewichtung und Varianzschätzung bei der Überprüfung des Erreichens der Grundkompetenzen 2016.* Institut für Bildungsevaluation, assoziiertes Institut der Universität Zürich. https://www.uegk-schweiz.ch/wp-content/uploads/2019/05/%C3%9CGK2016_Verner_Helbling_2019_-Sampling-%C3%9CGK-2016.pdf