DigiPrim

UNIVERSITÄT BERN

Interfaculty Centre for Educational Research (ICER)

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

School Principal Questionnaire: Data Manual

Cooperation ÜGK / DigiPrim

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

October 2023, document version v1-0

Abstract: This data manual serves as a resource for researchers who plan to use the data from the school principal questionnaire of the ÜGK / COFO / VECOF 2024 (HarmoS 4) Field Trial 2022. This data manual offers a documentation of the data editing process for the scientific use file and provides guidance on how to use the scientific use file. The data manual, therefore, functions as documentation but also goes beyond the documentation purposes of the respective data file.

Keywords: data documentation, variables, ÜGKH4 field trial, survey data, principal questionnaire, large-scale assessment

Suggested citation: Herzing, Jessica M. E., Seiler, Simon, & Röhlke, Leo (2023). ÜGK / COFO / VECOF 2024 (HarmoS 4) Field Trial, School Principal Questionnaire: Data Manual. Version 1. Bern: University of Bern, Interfaculty Centre for Educational Research. DOI:10.48350/183643.

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 express their gratitude to 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 ... 3
1.1 About this manual ... 3
1.2 Overview of the data ... 3
1.3 Respect rules of data usage ... 4
1.4 Publications with ÜGK / DigiPrim data ... 4
2 General conventions ... 4
2.1 File names ... 4
2.2 Variable names ... 5
2.3 Missing values ... 7
2.4 Special conventions for variables from test data ... 8
3 Study design: Sampling ... 8
4 Data structure ... 9
5 Data processing ... 9
5.1 Unit nonresponse ... 9
5.2 Item nonresponse ... 9
5.3 Coding ... 10
5.4 Other open-ended answers ... 11
5.5 Generated data ... 13
5.6 Plausibilization ... 19
5.7 Further anonymization ... 19
5.8 Anonymization principles ... 20
6 Sample and nonresponse adjustments ... 21
6.1 Nonresponse adjusted sampling weight ... 21
6.2 Recommendations for addressing the complex survey design ... 22
7 Further information ... 22
7.1 Recommendations ... 22
References ... 24

1 Introduction

1.1 About this manual

This manual is designed to offer guidance to researchers and other interest groups in navigating the *school principal questionnaire* data of the ÜGK / COFO / VECOF 2024 study (HarmoS 4) Field Trial 2022, which is referred to here as the ÜGKH4 field trial. Its primary purpose is to provide information on key aspects of the data structure, the data processing (cleaning), and the survey and sampling design associated with the school principal questionnaire data. Additionally, the manual provides comprehensive insights into the coding of open answers and generated variables included in the scientific-use file (SUF).

Throughout this manual, other documents within the ÜGK / DigiPrim collaboration documentation are referenced. This documentation encompasses a study description of the DigiPrim project, a codebook, a technical report, a report detailing scales and concepts, along with further resources concerning the items and implementation of the school principal questionnaire. All these materials can be accessed via SWISSUbase.

Should you come across any mistakes or have suggestions to enhance the quality of this data manual or other related documents, please feel free to reach out to us at: data.icer@unibe.ch

1.2 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 school principal questionnaire only contains information resulting from the school principal questionnaire and some additional information from the sampling process (e.g., weights). Access to critical data, such as open-ended responses (text entries), unaggregated 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

1.3 Respect rules of data usage

Any usage of the ÜGK / COFO / VECOF (including the school principal 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).

1.4 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 school principal 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) of 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[16, 17, 18, 23, 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
[ts, sq, par, sch, para, pro]	Indicator for data source: sq = student questionnaire par = parent questionnaire sch = school principal questionnaire
[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 2 (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 3: 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 in 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 4: 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 field trial of ÜGKH4 was conducted. The aim of ÜGKH4 field trial is 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 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.

A two-stage sampling procedure was used in all cantons. First, school sites were sampled and then, analogous to the single-stage sampling procedure, students in the selected schools were randomly selected. For further details on the sampling, please consult the DigiPrim study description (Herzing et al., 2023).

The goal of the school principal questionnaire is to gather information on the school context of the children in the study. Consequently, the school sites on the first stage of the sampling process represent the target population for the school principal questionnaire. School principals (N = 240) of all sampled school sites (N = 275) were invited to participate in an online questionnaire via an invitation and a reminder email. Because some principals were principals of more than one of the sampled school sites, the number of invited principals is lower than the number of school sites. The respective school principals received one invitation email asking for participation, with different personalized online links for every sampled school site. In such cases, the school principal was assigned one "main" questionnaire for a randomly selected school site, and a short questionnaire for each additional school site. The short questionnaire only included a selected number of questions, focusing on the school context and not on the school principal him- or herself.

Based on the survey participation status of the school sites represented by the school principals, a probability weight adjusting for nonresponse mechanisms (*smp_nra_sqw*) as well as a variable indicating the strata (*smp_strata*) was generated and included in the scientific-use file (SUF). Section 6 offers further guidance on how to use these variables in empirical analyses. It is important to keep in mind when using the data, especially when weights are applied, that the target population of the study are *school sites* and students, not school principals. It is impossible to infer valid statements about any population of school principals based on the sample underlying the school principal questionnaire dataset.

4 Data structure

The school principal questionnaire data has a cross-sectional data structure with one row corresponding to one school site (not one school principal), identified by the variable *mergeid_school_princquest*. Per sampled school (site), only one school principal questionnaire could be submitted, which was ensured by assigning personalized links to the online questionnaire. As described in the study design section, school principals are often responsible for multiple school sites. Therefore, the sample included several school sites for which the same school principal was responsible. In such cases, the school principal received one invitation to the "main" questionnaire and an additional invitation to a short questionnaire for each additional school site.

Even if a school site was covered by a short questionnaire, the respective data row in the SUF contains responses (if available) to all questions posed in the main questionnaire. This was achieved by duplicating the responses from the school's responsible school principal given in the main questionnaire. Only responses to questions not included in the short questionnaire were duplicated. Hence, data users are provided with full information on all school sites in the SUF (except for cases of item nonresponse), even if a school site was covered by a short questionnaire only.

The first part of the main principal questionnaire includes general questions on the person of the school principal and the internal school context. The second part features questions from a collaborating project on teaching outside the classroom (SILVIVA). The third part addresses the digitalization in primary schools, with questions from the DigiPrim project. Questions in this part cover digital resources, the position of the ICT coordinator, the digital school culture, and school principals' attitudes towards ICTs. The last part includes the questionnaire evaluation and questions regarding the willingness to be contacted for further research projects.

5 Data processing

Data processing of the school principal 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.

5.2 Item nonresponse

Item nonresponse in the school principal questionnaire has four possible reasons (see section 2.3). Skipping and therefore not answering individual items was possible in the online survey, with no warning message appearing. Furthermore, questionnaires did not have to be completed 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 can be plausibly assumed because all following items are missing (breakoff). For details on coding of implausible answers, please refer to section 5.6 (Plausibilization).

5.3 Coding

furtheredu_cert, furtheredu_events, furtheredu_educoff, furtheredu_private, furtheredu_mentor, furtheredu_conf, furtheredu_network, furtheredu_read

Some respondents selected only the “1 (yes)”, but never the “0 (no)” option. In order to provide data users with variables with the lowest possible number of unnecessary missing values, missing values were recoded as “0 (no)” if 0 had been selected for none of the educational opportunities by the respective school principal, assuming that these values represent non-attendance rather than “true” missing values. If 0 had been selected at least once in this item battery, missing values were retained. Data users may consider recoding some of these missing values as zeros.

schoolstruct_steer, schoolstruct_level, schoolstruct_subj, schoolstruct_parent, schoolstruct_health, schoolstruct_culture, schoolstruct_lesson, schoolstruct_students, classtype_regular, classtype_mixedage, classtype_special, driverict_proj, driverict_manag, driverict_coll, driverict_author, driverict_comp, driverict_curr, driverict_mat, deci_recom, deci_finan, deci_licen, deci_digistrat, deci_compat, deci_host, deci_server, deci_edulog, deci_sourcecode

If respondents selected none of the response categories but gave at least one valid answer in the preceding or following question, missings were treated as zeros, assuming that the question had been seen. Data users may consider recoding some of these zeros as missing values.

teambody_totalfull, teambody_totalpart, teambody_recogfull, teambody_recogpart, teambody_norecogfull, teambody_norecogpart

The answer categories were implemented as open text fields. Some respondents filled in zeros, while others left certain fields empty. Empty values were recoded as zeros in the variables *teambody_recogfull, teambody_recogpart, teambody_norecogfull* and *teambody_norecogpart* if mathematically necessary (e.g., if *teambody_totalfull* = *teambody_recogfull* and *teambody_norecogfull* = missing). It was assumed that these cases represent true zeros rather than missing values. In several cases, the values from the different categories do not add up (e.g., more teachers with a diploma than total teachers). Because the *teambody* variables are provided in anonymized form in the SUF, users are provided with

the flag variable *teambody_flag_g* indicating inconsistencies not resulting from missing values (see section 5.5). Data users are advised to use these variables with some caution and possibly conduct sensitivity analyses with and without flagged observations.

teamage#, num_board, num_beamer, num_pcteacher, num_vis, num_lab, num_mobile, num_tablet, num_h4_mobile, num_h4_tablets

The answer categories were implemented as open text fields. Some respondents filled in zeros, while others left certain fields empty. Empty categories were recoded as zero if respondents had filled in at least one of the categories with a value larger than zero but had not used zeros in any of the categories. It was assumed that these cases represent true zeros rather than missing values.

schoolsteps#

Because having no school levels is impossible, all cases in which respondents had not selected any school level were recoded as missing values.

ictc_teach_ict, ictc_teach_sub

In the rare case that respondents who selected “Yes” for the category “No teaching of students” (variable *ict_teach_noteach*) and left one of these two variables referring to the teaching of students empty, it can be assumed with high certainty that these represent a zero (“no"), not a true missing value. Hence, these cases were set to zero instead of missing.

5.4 Other open-ended answers

motive_othertext1, motive_othertext2

Open answers were categorized manually, resulting in the generated variables *motive_othertext1_g* and *motive_othertext2_g*. If responses referred to already existing categories in their open answers, the respective categories were recoded as 1 (named).

schoolmodel_othertext

Open answers were categorized manually, resulting in the generated variable *schoolmodel_othertext_g*. If responses referred to already existing categories in their open answers, the respective categories were recoded as 1 (yes).

addsupport_othertext

Open answers in this variable did not refer to existing categories and neither yielded meaningful new categories. The presence of open answers is documented in the variable *addsupport_othertext_g*.

workload_othertext

Open answers were categorized manually, resulting in the generated variable *workload_othertext_g*.

schoolstruct_othertext

Open answers were categorized manually, resulting in the generated variable *schoolstruct_othertext_g*. If responses referred to already existing categories in their open answers, the respective categories were recoded as 1 (yes).

classtype_othertext

Open answers were categorized manually. All open answers in this variable referred to already existing categories of class types (regular, mixed age, or special classes). The corresponding variables were recoded as 1 (yes).

spec_othertext

Open answers in all languages were categorized by a collaborating expert on the subject of special education. If respondents referred to already existing categories in their open answers, the corresponding value (number of students) from the variable *spec_other* was added to the value of the respective category. Several answers were recoded as invalid according to the expert's suggestion. For plausibility checks, see section 5.6.

ictc_name

Open answers in all languages were categorized by a collaborating expert on the subject of ICT (information and communication technology) coordination and support in schools. The resulting variable *ictc_name_g* can be found in the SUF.

deci_othertext

Several respondents indicated not to be responsible for any decisions in relation to software purchases. In these cases, all variables from this item battery were recoded to "does not apply". If responses referred to already existing categories in their open answers, the respective categories were recoded as "yes". A new variable (*deci_other_g*) was generated to capture the remaining answers which did not fit into any of the existing categories. Due to a low number of "other" answers, this variable does not further specify which other decision criteria were described.

driverict_othertext

If respondents referred to already existing categories in their open answers, the respective categories were recoded as 1 (named). A new variable (*driverict_other_g*) was generated to cover the remaining answers which did not fit into any of the existing categories. Due to a low number of "other" answers, this variable does not further specify which other drivers were named in particular.

pro_othertext

A new variable (*pro_other_g*) was generated to capture open answers which did not fit into any of the existing categories. One new category emerged called "General program, local

level", describing ICT-related programs or collaborations on the local level, e.g., the municipality. The remaining open answer were labeled as "any other".

risk_othertext

A new variable (*risk_screentime_g*) was generated to capture answers relating to "Screen time" as a risk not covered by the existing categories. The variable *risk_other* covers the remaining open answers.

chance_othertext

The very few open answers to this question are treated as unspecific "other" answers and are covered by the variable *chance_other* without further data manipulation.

5.5 Generated data

age_g

This variable contains the categorized age of the respondent, obtained by subtracting the self-reported birthyear of the principal from the year 2022. The categories are based on the categories used in the "Schulleitungsmonitor Schweiz 2021" (Tulowitzki et al., 2023). The categorization was made for anonymization purposes.

motive_othertext1_g, motive_othertext2_g

These variables indicate motives for becoming a school principal other than the ones covered by the regular items. They contain the categorized open answers from variables *motive_othertext1* and *motive_othertext2*. Answers which referred to existing categories of motives were recoded as "does not apply".

schoolmodel_othertext_g

This variable contains the categorized open answers from variable *schoolmodel_othertext*. Answers which referred to existing categories of the variable *schoolmodel* were recoded as "does not apply".

addsupport_othertext_g

This variable indicated the presence of open answers in the variable *addsupport_othertext*.

workload_othertext_g

This variable contains the categorized open answers from the variable *workload_othertext*. The values in variable *workload_other* refer to these categories. Data users should be aware that some of these categories are identical to the original categories, and the data editors did not remove possibly contradictory information.

workload_other_g

This variable contains values from the variable *workload_other*. If open answers in *workload_othertext* referred to original categories in any of the workload variables, the respective

value in *workload_g* was recoded as “does not apply”. If respondents indicated more than one category of workload in *workload_othertext*, the value in *workload_g* was recoded as “implausible / invalid”. The remaining values are unchanged.

schoolstruct_othertext_g

This variable contains the categorized open answers from the variable *schoolstruct_othertext*. Answers which referred to existing categories of school structures were recoded as “does not apply”.

fulltime_ratio_total_g

This variable contains the ratio of full-time teachers (variable *teambody_totalfull*) to the total number of teachers (the sum of the variables *teambody_totalfull* and *teambody_totalpart*).

fulltime_ratio_recog_g

This variable contains the ratio of full-time teachers with a recognized diploma (variable *teambody_recogfull*) to the total number of teachers with a recognized diploma (the sum of the variables *teambody_recogfull* and *teambody_recogpart*). Data users should be careful when analyzing variables with information on diploma status because respondents often gave inconsistent answers.

fulltime_ratio_norecog_g

This variable contains the ratio of full-time teachers without a recognized diploma (variable *teambody_norecogfull*) to the total number of teachers with a recognized diploma (the sum of the variables *teambody_norecogfull* and *teambody_norecogpart*).

diploma_ratio_total_g

This variable contains the ratio of teachers with a recognized diploma (the sum of the variables *teambody_recogfull* and *teambody_recogpart*) to the total number of teachers (the sum of the variables *teambody_recogfull, teambody_recogpart, teambody_norecogfull* and *teambody_norecogpart*).

diploma_ratio_full_g

This variable contains the ratio of full-time teachers with a recognized diploma (variable *teambody_recogfull*) to the total number of full-time teachers (the sum of the variables *teambody_recogfull* and *teambody_norecogfull*).

diploma_ratio_part_g

This variable contains the ratio of part-time teachers with a recognized diploma (variable *teambody_recogpart*) to the total number of part-time teachers (the sum of the variables *teambody_norecogpart* and *teambody_recogpart*).

flag_teambody_g

This flag variable indicates whether mathematically inconsistent values are present. A value is flagged as inconsistent if the sum of the variables *teambody_recogpart* and *teambody_norecogpart* does not equal the value of *teambody_totalpart* or if the sum of the variables *teambody_recogfull* and *teambody_norecogfull* does not equal the value of *teambody_totalfull*. Data users may consider recoding the values of the respective ratios to missing values and conduct sensitivity analyses.

spq_teamage#_ratio_g

These variables contain the ratio of teachers in different age groups (variable *teamage#*) to the total number of teachers (measured as the sum over all *teamage#* variables. Hence, the ratios over all age groups add up to 1.

schoolsize_g

This variable contains the size of each school in terms of the number of students enrolled in cycles 1 and 2, based on the *studnumber* variable. Because defining substantive thresholds is difficult and the original values considerably increase deanonymization risks, the original values were cut into quartiles. Hence, there is no substantive interpretation to each category. The variable indicates the relative size of each school compared to the other schools in the SUF. Data users should be cautious when using this variable, as there are some indications (DigiPrim-internal evaluations) that school principals referred to different frames of reference when reporting the number of students, e.g., some school principals may have reported the total number of students over different school sites.

total_spec_g

This variable contains the sum of the variables *spec_basicspec* and *spec_reinfspec*. As with all absolute numbers, it was completely anonymized.

quote_spec_g

This variable contains the sum of the variables *spec_basicspec* and *spec_reinfspec* divided by the number of students as reported in the variable studnumber, multiplied by 100. This is the support rate at the school level (Klemm, 2014).

quote_spec_basicspec_g

This variable contains the ratio of the variable *spec_basicspec* to the number of students as reported in the variable *studnumber*, multiplied by 100.

quote_spec_reinfspec_g

This variable contains the ratio of the variable *spec_reinfspec* to the number of students as reported in the variable *studnumber*, multiplied by 100.

quote_spec_gifted_g

This variable contains the ratio of the variable *spec_gifted* to the number of students as reported in the variable *studnumber*, multiplied by 100.

quote_spec_germansec_g

This variable contains the ratio of the variable *spec_germansec* to the number of students as reported in the variable *studnumber*, multiplied by 100.

quote_spec_logo_g

This variable contains the ratio of the variable *spec_logo* to the number of students as reported in the variable *studnumber*, multiplied by 100.

quote_spec_psychomot_g

This variable contains the ratio of the variable *spec_psychomot* to the number of students as reported in the variable *studnumber*, multiplied by 100.

spec_stud_les_basic_g

This variable contains the ratio of the variable *hfhnumber* to the number of students receiving basic educational measures as reported in the variable *spec_basicspec*.

spec_stud_les_reinforced_g

This variable contains the ratio of the variable *savnumber* to the number of students receiving basic educational measures as reported in the variable *spec_reinfspec*.

student_vze_spec_g

The variable contains the number of students divided by the number of full-time special education teaching units. The average number of weekly lessons of a full-time employed teacher in Switzerland is 28. The ratio was calculated using the following formula:

$$\frac{studnumber}{\left(\frac{hfhnumber + savnumber}{28}\right)}$$

spec_other_bandu_g

A new variable (*spec_other_bandu_g*) was generated to capture the number of students receiving advice and support units (a type of special support called “Beratung und Unterstützung (B&U)” in the Swiss context). This variable was generated according to a suggestion by the expert (see section 5.4).

silviva_school_g

This variable contains the mean of the variables *silviva_schoolyard, silviva_garden, silviva_campus* and *silviva_outsclass*, omitting respondents with missing values in at least one of the variables.

silviva_natural_g

This variable contains the mean of the variables *silviva_park, silviva_forest, silviva_lawn, silviva_water, silviva_hedge* and *silviva_reserve*, omitting respondents with missing values in at least one of the variables.

silviva_cultural_g

This variable contains the mean of the variables *silviva_agricult* and *silviva_infrastruct*, omitting respondents with missing values in at least one of the variables.

board_stud_ratio_g

This variable contains the ratio of the number of whiteboards (variable *num_board*) to the total number of students (variable *studnumber*), both referring to school cycles 1 and 2. A cautionary use of this variable is advised: Data users should be aware that the denominator variable *studnumber* stems from principals' responses and may therefore be inaccurate in some cases, when principals potentially refer to their school (site) as a whole.

beamer_stud_ratio_g

This variable contains the ratio of the number of projectors (variable *num_beamer*) to the total number of students (variable *studnumber*), both referring to school cycles 1 and 2. A cautionary use of this variable is advised: Data users should be aware that the denominator variable *studnumber* stems from principals' responses and may therefore be inaccurate in some cases, when principals potentially refer to their school (site) as a whole.

vis_stud_ratio_g

This variable contains the ratio of the number of visualizers (variable *num_vis*) to the total number of students (variable *studnumber*), both referring to school cycles 1 and 2. A cautionary use of this variable is advised: Data users should be aware that the denominator variable *studnumber* stems from principals' responses and may therefore be inaccurate in some cases, when principals potentially refer to their school (site) as a whole.

lab_stud_ratio_g

This variable contains the ratio of the number of stationary computers (variable *num_lab*) to the total number of students (variable *studnumber*), both referring to school cycles 1 and 2. A cautionary use of this variable is advised: Data users should be aware that the denominator variable *studnumber* stems from principals' responses and may therefore be inaccurate in some cases, when principals potentially refer to their school (site) as a whole.

mobile_stud_ratio_g

This variable contains the ratio of the number of laptops (variable *num_mobile*) to the total number of students (variable *studnumber*), both referring to school cycles 1 and 2. A cautionary use of this variable is advised: Data users should be aware that the denominator variable *studnumber* stems from principals' responses and may therefore be inaccurate in some cases, when principals potentially refer to their school (site) as a whole.

tablet_stud_ratio_g

This variable contains the ratio of the number of tablets (variable *num_tablet*) to the total number of students (variable *studnumber*), both referring to school cycles 1 and 2. A cautionary use of this variable is advised: Data users should be aware that the denominator variable *studnumber* stems from principals' responses and may therefore be inaccurate in some cases, when principals potentially refer to their school (site) as a whole.

mobile_h4_stud_g, tablet_h4_stud_g

These variables indicate the device availability of laptops and tablets for students on school level HarmoS 4. The variables are based on a comparison between the number of laptops (tablets) available for all students on cycles 1 and 2 (*num_mobile, num_tablets*) and the number of the same devices available only for HarmoS 4 students (*num_h4_mobile, num_h4_tablets*). Because the number of students on the school level HarmoS 4 was not surveyed, calculating ratios for HarmoS 4 as for cycles 1 and 2 was not feasible. The four resulting categories indicate whether there are no devices for any students on cycles 1 and 2, whether there are no devices solely for HarmoS 4 students, whether the number indicated for both was the same, or whether it was smaller for HarmoS 4 students. In some instances, principals reported a higher number of devices only for HarmoS 4 students. This was considered implausible, because HarmoS 4 students represent a subset of all students on cycles 1 and 2. The overall validity of the HarmoS 4 device measure is unclear, the resulting variables should be treated with caution.

pcteacher_ratio_g

This variable contains the ratio of the number of computers for teachers (variable *num_pcteacher*) to the total number of teachers (the sum of the variables *teambody_totalfull* and *teambody_totalpart*), both referring to school cycles 1 and 2.

ictc_name_g

This variable indicates the categorized name of the ICT coordinator. It is the product of the external coding procedure described in section 5.4, covering various terms for the position of those who are mainly responsible for ICT affairs within Swiss schools.

deci_other_g

See section 5.4 on open answers.

driverict_other_g

See section 5.4 on open answers.

pro_other_g

See section 5.4 on open answers.

risk_screentime_g

See section 5.4 on open answers.

5.6 Plausibilization

student_vze_spec_g

Values beyond 1000 students per full-time unit were considered as implausible by an expert in the field and therefore recoded as missing.

quote_spec_g, quote_spec_basicspec_g, quote_spec_g_flag, quote_spec_basicspec_flag

Values beyond 100% were capped in variables *quote_spec_g and quote_spec_basicspec_g*. In some cases, when all students receive basic special educational measures, and some students receive reinforced measures, the sum can exceed the total number of students enrolled in the school. The flag variables enable data users to recode the respective values in case they disagree with this step.

quote_spec_gifted_g, quote_spec_reinfspec_g, quote_spec_germansec_g, quote_spec_logo_g, quote_spec_psychomot_g

Values beyond 100% were considered as implausible and recoded to a missing value.

total_spec_g, quote_spec_g, spec_basicspec, spec_reinfspec, quote_spec_basicspec_g, quote_spec_reinfspec_g, spec_stud_les_basic_g, spec_stud_les_reinforced_g

For four school sites, the expert considered all values with regard to basic or reinforced measures as implausible and therefore, these were recoded to missing values.

schoolsteps#

Several school principals selected only other school levels than HarmoS 4 as their responsibility. Others indicated to only be responsible for one school level (HarmoS 4). While these responses are not in line with the recruitment of participants or highly implausible, they were retained in order to give data users the opportunity to decide for themselves on how to deal with these cases. Given that school principals were supposed to be selected based on their responsibility for HarmoS 4, it is possible that the question was misunderstood by some respondents.

5.7 Further anonymization

schoolsince

The variable was top- and bottomcoded because a very high (more than 20 years) or very low (less than one year) number of years as school principal increases the risk of school identification.

teamnumber

The variable was topcoded at six or more members of the school management team to avoid re-identification, because large school management teams are an exception.

5.8 Anonymization principles

The ÜGK / COFO / VECOF surveys encompass data on individuals, context persons (such as parents, teachers, or school principals), and institutions (such as school forms). Adhering to the data use concept (Data use concept ÜGK) , it is imperative that individual (or personal) data is treated with the utmost care by the data processing and data security infrastructure. In this regard, careful practice must be applied, when ÜGK / COFO / VECOF data is distributed, as mishandling could lead to legal repercussions. Therefore, the disseminated data needs to be anonymized without preventing certain research projects or education monitoring from being conducted.

To safeguard the data, we have implemented a multi-layered security approach comprising five distinct measures to ensure the best possible protection for the data:

1. Institutional: Data is exclusively disseminated through a Swiss data archive or the overseeing institution, with differential privacy applied to data sets.
2. Legal: Data users are required to sign a data use agreement, committing to comply with institutional security measures and using the data solely for research purposes.
3. Informational: Users are educated on data protection and potential misuse through documentation.
4. Technical: Sensitive data is available only upon special request and subject to restrictions.
5. Statistical: Data is modified to enhance anonymity, involving techniques such as aggregation, recoding, and data suppression to prevent re-identification

To further emphasize the importance of data protection, it is worth noting that when data is downloaded from the archive to a researcher's local workstation, it moves out of the secure archive area. Therefore, additional anonymization procedures are necessary.

For this purpose, a statistical security approach is employed, involving techniques like aggregation, recoding, and the removal of certain variables or attributes that could facilitate re-identification. These adjustments were made for quasi-identifiers, considering both the sample size and the total population.

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 age of a school principal, e.g., “53 years old”, is aggregated into a global category such as 50-60 years old”, and so forth. Local recoding would mean replacing one or more values within an unsafe combination with a missing value.

Topcoding values is 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 is 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 (e.g., 0.5 years). This could be the number of years as school principal (e.g., less than one year).

Aggregating and top- / bottomcoding 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 the 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. 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 (see section 2.3) and hence, the original variable is purged. Consequently, all variables can be accessed in the SUF, but their content is not published.

6 Sample and nonresponse adjustments

6.1 Nonresponse adjusted sampling weight

School sites educating students on HarmoS level 4 represent the primary sampling unit (PSU) on the first stage of the ÜGKH4 field trial two-stage student sampling. Within these PSUs, eligible students visiting classes on HarmoS level 4 were randomly selected (second stage). School sampling on the first stage was stratified along several criteria (e.g., school size, language region). In addition, some sampled school sites did not participate in the survey. The SUF of the principal questionnaire data, therefore, contains a nonresponse adjusted sampling weight (*smp_nra_sqw*), which simultaneously corrects for unequal sampling probabilities and nonresponse of school sites. These weights were calculated in an equivalent approach as described by Verner and Helbling (2019). Please note when applying the weights that the nonresponse adjustment was based on a limited set of predictors which are unlikely to mitigate all selective unit nonresponse.

An alternative weight named *smp_stuw_sch* contains the sum of student weights on the level of schools in the SUF of the school principal questionnaire. The resulting weight does not account for selective nonresponse in the school principal questionnaire. All sampled students eligible for participation in the ÜGKH4 field trial were included in the calculation of the sum of student weights. Similar school weights based on student weights have been used in the past in the context of PISA (Programme for International Student Assessment; OECD, 2019a, 2019b).

6.2 Recommendations for addressing the complex survey design

Data users are generally advised to make use of this sampling weight regardless of the type of analysis. In addition, the SUF includes a stratum variable (*smp_strata*) which is equivalent to the language region (*smp_region*). Accounting for stratification together with the sampling weight yields more accurate standard errors when inferential statistics are calculated. For example, in the statistical software Stata (v16.1), the complex survey design can be conveniently 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.

- be aware that the target population are school sites (not school principals) when interpreting your results.
- note that the relatively high unit response rate 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.

References

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

Klemm, K. (2014). Auf dem Weg zur inklusiven Schule: Versuch einer bildungsstatistischen Zwischenbilanz. *Zeitschrift Für Erziehungswissenschaft*, *4*(17), 625–637.

OECD. (2019a). *PISA 2018 Technical Report: Chapter 8: Survey Weighting and the Calculation of Sampling Variance*. https://www.oecd.org/pisa/data/pisa2018technicalreport/PISA2018-TecReport-Ch-01-Programme-for-International-Student-Assessment-An-Overview.pdf

OECD. (2019b). *What school life means for students' lives: Annex A3. Technical notes on analyses in this volume*. *PISA 2018 results / OECD: volume 3*. OECD Publishing. https://www.oecd-ilibrary.org//sites/ff99ab98-en/index.html?itemId=/content/component/ff99ab98-en#

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

Tulowitzki, P., Pietsch, M., Cometti, N., Sposato, G. G., & Schweinberger, K. (2023). *Schulleitungsmonitor Schweiz 2021 – Skalendokumentation.* https://doi.org/10.26041/fhnw-4832

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