ÜGK/COFO/VeCoF 2017: Assessment of language skills

# Technical report:

# Student-questionnaire data

Giang Pham

Manuela Hauser

PH SG
Pädagogische Hochschule
St.Gallen

# 1 Introduction

In 2017, the second nationwide Assessment of the Achievement of Basic Educational Competences (ÜGK/COFO/VeCoF) was conducted to assess the achievement of the basic competencies of Swiss students at the end of the 8th year of compulsory schooling in the school language (**L1**) and in the first foreign language (**L2**) learned at school. A total of 20,177 students from all Swiss cantons participated in the survey (50% were female; the mean age of students at the time of survey was 12.5 and ranged from 8.8 to 16.9). Besides the language tests, the survey included a student questionnaire (**SQ**) which collected information on selected school and out-of-school influencing factors of student performance, including individual, family-, and school-related contextual characteristics. This technical documentation describes the data processing and scaling procedures of the questionnaire data in the survey. For a description of the student population, the student sample, and the sampling process, see Verner and Helbling (2019). The test design, the test development process, the data processing, and the scaling procedures of test data are described by Angelone and Keller (2019).

## 1.1 The student questionnaire

A scientific team consisting of researchers from five universities and educational institutions in Switzerland collaborated to construct the student questionnaire (**SQ**) in this survey. For an introduction to the theoretical constructs of the SQ, see Erzinger, et al. (2019). In brief, the SQ contained items and scales covering aspects from the following main domains:

- home-related factors (e.g., social background of the family)
- individual characteristics of the student (e.g., academic self-concept, learning strategies)
- teacher characteristics and teaching-related factors (e.g., using multimedia in instruction)
- other schooling effects (e.g., years of learning the first foreign language – **L2** – at school)

A full list of all constructs (items and scales) included in the SQ is shown in Table 1.

*Table 1: Constructs in the ÜGK 2017 SQ*

| Question Number | Construct | Construct_ ID | L2-specific | Domain | No. of Items | Source |
|---|---|---|---|---|---|---|
| **AA/AB** | Test motivation | testmot | No | Student | 3 | (BIFIE, 2015) |
| **Contributions from home** | | | | | | |
| **A01** | Family structure | hhmemb | No | Home | 9 | (OECD, 2002; OECD, 2014) |
| **A02** | Employment status mother | emplm | No | Home | 1 | (OECD, 2002; TREE, 2016; DAB , 2016) |
| **A03** | Employment status father | emplf | No | Home | 1 | (OECD, 2002; TREE, 2016; DAB , 2016) |

| | | | | | | |
|---|---|---|---|---|---|---|
| **A04m** | Occupation mother (Component of construct social background) | occupm | No | Home | 2 | (OECD, 2002) |
| **A05m** | Occupational status mother | soclm | No | Home | 1 | (DAB , 2016 adapted) |
| **A04f** | Occupation father (Component of construct social background) | occupf | No | Home | 2 | (OECD, 2002) |
| **A05f** | Occupational status father | soclf | No | Home | 1 | (DAB , 2016 adapted) |
| **A08m** | Mother's education level (Component of construct social background) | meduc | No | Home | 2 | (OECD, 2002 adapted; TREE, 2016; DAB, 2016) |
| **A08f** | Father's education level (Component of construct social background) | feduc | No | Home | 2 | (OECD, 2002 adapted; TREE, 2016; DAB, 2016) |
| **A10** | Country of birth (student/father/mother) (Primary variables of immigration background) | cob | No | Home | 3 | (OECD, 2002 adapted) |
| **A11** | Age when arrived in Switzerland, if not born here | immigage | No | Home | 1 | (OECD, 2013) |
| **A12a** | Main language spoken at home (Component of construct home language) | langhome | No | Home | 1 | (OECD, 2002; OECD, 2014) |
| **A13a/b** | Another language spoken at home (Component of construct home language) | langhome2 | No | Home | 2 | (OECD, 2014) |
| **A14** | Number of books at home (Component of construct social background) | books | No | Home | 1 | (OECD, 2002) |
| **A15a/b** | Wealth index | wealth | No | Home | 10 | (OECD, 2002; OECD, 2014; OECD, 2020) |
| **A15a** | Cultural possession index | cultposs | No | Home | 3 | (OECD, 2002; OECD, 2014) |
| **A16** | Family affluence scale | holyn | No | Home | 1 | James 2014 (adapted) according to Hupka-Brunner, et al. (2015) |
| **C01** | Family educational support | famedsup | No | Home | 3 | (OECD, 2002) |
| **C03** | Parental pressure for academic success | press | No | Home | 4 | (Böhm-Kasper, Bos, Körner, & Weishaupt, 2001) |
| **Schooling, teaching, learning and student related factors** | | | | | | |
| **B01** | Age of entry to kindergarten | kinderg | No | Student | 1 | (Hascher, 2004) |
| **B02** | Emotional integration | emoint | No | Student | 4 | (Venetz, Zurbriggen, Eckhart, Schwab, & Hessels, 2015) |
| **B02** | Social integration | socint | No | Student | 4 | (Venetz, Zurbriggen, Eckhart, Schwab, & Hessels, 2015) |
| **B02** | Academic self-concept (PIQ) | motint | No | Student | 4 | (Venetz, Zurbriggen, Eckhart, Schwab, & Hessels, 2015) |

| | | | | | | |
|---|---|---|---|---|---|---|
| **B03a/b** | Grade retention | graderep | No | Student | 2 | (OECD, 2014) |
| **B04a** | Out-of-school learning support | coach | No | Home | 4 | (OECD, 2014 adapted) |
| **B04b** | School subject with out-of-school learning support | coachaim | No | Home | 4 | (OECD, 2014 adapted) |
| **B05a/b/c** | School marks/grades | mark | yes | Student | 2 | (OECD, 2002 adapted) |
| **B21/H17** | Positive attitude towards school | posatsc | No | Student | 7 | (Hascher, 2004; Hascher, Hagenauer, & Schaffer, 2011) |
| **B22/H18** | Joy in school | gladsc | No | Student | 6 | (Hascher, 2004; Hascher, Hagenauer, & Schaffer, 2011) |
| **B23/H19** | Academic self-esteem | selfe | No | Student | 4 | (Hascher, 2004; Hascher, Hagenauer, & Schaffer, 2011) |
| **B06** | Interest in reading | intrea | No | Student | 3 | (OECD, 2002) |
| **B06** | Academic self-concept (PISA) | scacad | No | Student | 3 | (OECD, 2002) |
| **B06** | Reading self-concept | scverb | No | Student | 3 | (OECD, 2002) |
| **B06** | L2 self-concept | l2con | yes | Student | 3 | (OECD, 2002) |
| **B07** | IT-battery mode effect CASI (EDK) | ict | No | Student | 2 | (ICILS, 2015 adapted) |
| **B08** | IT-battery mode effect CASI (EDK) | ictmot | No | Student | 8 | (ICILS, 2015 adapted) |
| **B09** | Language contact, exchange | langcont | yes | Schooling | 2 | (ESLC, 2011 adapted) |
| **B10** | L2 instruction: Using multimedia in instruction | teachlist | yes | Teaching | 1 | (ESLC, 2011 adapted) |
| **B11** | L2 instruction: Using the test language | teachuse | yes | Teaching, Learning | 4 | (ESLC, 2011 adapted) |
| **B12** | L2 instruction: Using multilingual didactics | teachmult | yes | Learning | 4 | (ESLC, 2011 adapted) |
| **B13** | L2 instructional quality | teachbeh | yes | Teaching | 5 | DESI (adapted and extended) according to Dörnyei (2009) & Heinzmann (2013) |
| **B13** | L2 motivational self-system | motselfs | yes | Student | 7 | New, Dörn JCI (adapted) according to Dörnyei (2009) & Heinzmann (2013) (combined) |
| **B14** | L2 motivational self-system | motselfs | yes | Student, Home | 5 | New, Dörn JCI (adapted) according to Dörnyei (2009) & Heinzmann (2013) (combined) |
| **B14** | L2 learning experience | learnexp | yes | Student, Teaching, Teaching material | 5 | New items, DESI, Dörn JCI (adapted) according to Dörnyei (2009) & Heinzmann (2013) (combined) |
| **B14** | L2 language anxiety | langanx | yes | Student | 1 | (Heinzmann, 2013) |

| | | | | | | |
|---|---|---|---|---|---|---|
| **B15** | Reading frequency | freq | No | Student | 1 | (OECD, 2011; OECD, 2020 adapted) |
| **B16** | Learning L1: Memorisation | memor | No | Learning | 4 | (OECD, 2011) |
| **B16** | Learning L1: Control strategies | cstrat | No | Learning | 5 | (OECD, 2011) |
| **B16** | Learning L1: Elaboration | elab | No | Learning | 4 | (OECD, 2011) |
| **B17** | Learning L1: Understanding and rememoration | undrem | No | Learning | 6 | (OECD, 2011; OECD, 2020) |
| **B18** | Learning L1: Synthesis | metasum | No | Learning | 5 | (OECD, 2011; OECD, 2020) |
| **B19** | Joy in reading L1 | joyread | No | Student | 11 | (OECD, 2011; OECD, 2020) |
| **B20** | Relevance of reading L1 | readrelev | No | Student | 6 | SALSA/EVOLIT according to Hascher, Brühwiler, Erzinger, Girnat , & Hagenauer (2016) |
| **AC** | Years of learning L2 at school | AC | yes | Schooling | 1 | New item |

Each student was assigned a SQ with instructions in his/her school language after having completed the language tests. All items and scales – except for the L2-specific constructs and items of the subscale “Academic self-concept” from the scale “Perceptions of inclusion questionnaire” (PIQ, Venetz et al., 2015) – were assigned to all students (*common part* of the SQ). The items of the subscale “Academic self-concept” from the scale “Perceptions of inclusion questionnaire” (PIQ, Venetz et al., 2015) were included only in the version of the questionnaire for students participating in the canton of Grisons (students with Italian as L1 or L2 at school) for pilot purposes. Students with English or French as L2 at school also answered L2-specific questions (L2-specific part of the SQ). This L2-specific part was not included in the SQ version for students with German or Italian as L2 at school. In total, there were six versions of the questionnaire, corresponding to six student subsamples based on the combination of their L1 and L2 at school (see Table 2).

*Table 2: Subsamples and questionnaire versions*

| SQ version | Subsample L1 | Subsample L2 | SQ Common part | SQ L2-specific part | “Motivation integration”-items | Sample size |
|---|---|---|---|---|---|---|
| **1** | German | French | Included | Included | Not included | 4430 |
| **2** | German | English | Included | Included | Not included | 8699 |
| **3** | German | Italian | Included | Not included | Included | 741 |
| **4** | French | German | Included | Included | Not included | 5484 |
| **5** | Italian | French | Included | Included | Not included | 744 |
| **6** | Italian | German | Included | Not included | Included | 79 |

All questionnaire items are listed in three school languages (German, French, and Italian) in Appendix A.

## 1.2 Aims of the data processing and scaling procedures

The data processing and scaling procedures aimed at 1) preparing the context variables for the national report (Konsortium ÜGK, 2019) and 2) preparing datasets for the research community for secondary analyses.

To accomplish these aims, the following main processes were implemented:

- dealing with non-responses/missing data;
- calculating construct scores and the corresponding psychometric quality criteria (e.g., descriptive statistics, reliability indices, model fits, comparability of scale scores with regard to different instruction languages).

The process of dealing with missing data and the calculation of the context variables used in the national report (social background, home language, and immigration background) are described in the Technical Appendices of the national report (see Pham et al., 2019). The missing data were multiply imputed. The process of calculating the scale scores of other constructs in the SQ is described in the following section.

# 2 Construct scores and scaling models

In the SQ, there are constructs with only one indicator (item), two indicators, and three or more indicators (items).

## 2.1 Single-item constructs

The variable names and construct scores of the single-item constructs are equal to the variable names and values of the items (for the value labels, see columns `categories` and `cat_label` in Appendix B). They are:

- `A02`: Employment status mother
- `A03`: Employment status father
- `A05m`: Occupational status mother
- `A05f`: Occupational status father
- `A11`: Age when arrived in Switzerland, if not born here
- `A16`: Family affluence scale
- `B01`: Age of entry to kindergarten
- `B03a`: Grade retention
- `B10`: Instruction: Using multimedia in instruction
- `B14`: L2 language anxiety
- `B15`: Reading frequency
- `AC`: Years of learning L2 at school

## 2.2 Constructs with two indicators

For constructs with two indicators, the arithmetic mean scores were calculated and treated as construct scores. The variable names of these construct scores were created by combining the construct IDs (see Table 1) and the suffix "`_M`" (M stands for mean score). The names are:

ict_M (IT battery mode effect, CASI), langcont_M (Language contact, exchange), and TestMot_M (Test motivation[1]).

## 2.3 Some indices

"famstruct" is the index for family structure, which was calculated based on the values of the first four items of this construct (A01hhmemb1 – A01hhmemb4); it has four categories according to the PISA coding guide (OECD, 2005): 1 = a single parent family (students living with only one of the following: mother, female guardian, father, male guardian); 2 = a nuclear family (students living with a father and a mother); 3 = a mixed family (a father and a guardian, a mother and a guardian, or two guardians); 4 = another structure.

"coach" is the index which lists whether the student received paid-for learning support in the last year of primary school (e.g., private tuition, private coaching) in any school subject (0 = no, 1 = yes). This was recorded based on the values of four corresponding items: B04acoachlang1, B04acoachmath, B04acoachlang2, and B04acoachother. The other four items regarding the aims of the paid-for learning support B04bcoachaim1 – B04bcoachaim4 were treated and regarded as single items.

## 2.4 Constructs with three or more indicators

Differences between studies/projects – even within one study/project over different measurement points – have been observed regarding the relationship between one construct and its measures. Similarly, different estimates of construct scores have been applied in different studies. For instance, the relationship between the construct "Academic self-concept" and its items was modelled differently in the following studies/projects:

- Shavelson et al. (1976) and Marsh (1993) applied the confirmatory factor analysis (CFA) approach to investigate the dimensionality and different validity coefficients of this construct. They used the sum scores as estimates of construct scores. By doing this, interval-observed data were required/assumed, and all items had uniform nominal weights/slopes in the estimation of the construct scores.
- Item response theory (IRT) approaches were applied to model the relationship between "Academic self-concept" and its measures and to estimate the construct scores in PISA. However, different IRT models were applied in different PISA cycles: The partial credit model (PCM) (Master & Wright, 1997) was applied in PISA 2012 (OECD, 2014); the generalized partial credit model (GPCM) (Muraki, 1992) was applied in PISA 2015 (OECD, 2017). In both approaches, interval response data are not required/assumed. However, while items have uniform item slopes in the first approach, item slopes can vary in the second approach. In both PISA cycles, model-based weighted Warm likelihood estimates – WLEs (Warm, 1989) – were estimated as construct scores (OECD 2014, 2017).

[1] Originally, the construct "Test motivation" (AA/AB) had three indicators. However, one item had low discrimination ($r_{it}$ = –.05, see Section 3.1). This item (ABSQ002) was therefore eliminated from all subsequent data processing and analysis procedures.

- In ÜGK 2016 in Switzerland, the CFA model for ordered categorical data[2] was applied to model the relationship between "Academic self-concept" and its items, and model-based empirical Bayes means were estimated as construct scores (Sacchi & Oesch, 2017).

Given this observation, we computed not only one but three different estimates of scores for each construct with three or more indicators using different scaling models with different underlying assumptions. These estimates are arithmetic mean scores (variables with suffix "_M"), WLEs based on the PCM approach (variables with suffix "_PCM"), and WLEs based on the GPCM approach (variables with suffix "_GPCM"). In the case of the two constructs "Wealth index" and "Cultural possession index" with binary item responses, the Rasch model (Rasch, 1960) and the two-parameter logistic model – 2PL model (Birnbaum, 1968) – were applied. Due to the low reliability of the model-based WLEs for the construct "Cultural possession index" (see Appendix D), expected a posteriori (EAP) ability estimates (Bock & Mislevy, 1982) were computed instead of WLEs as individual scores for this construct. The item calibration, scaling process, and estimation of construct scores (WLEs/EAPs) were performed in the statistic environment R using the package **TAM** (Robitzsch, Kiefer, & Wu, 2018).

These procedures were applied for the following constructs:

- emoint: Emotional integration (EI)
- socint: Social integration (SI)
- posatsc: Positive attitude towards school
- gladsc: Joy in school
- selfe: Academic self-esteem
- scacad: Academic self-concept
- scverb: Reading self-concept
- intrea: Interest in reading
- l2con: L2 self-concept
- ictmot: IT-battery mode effect CASI (EDK)
- teachuse: Instruction: Using the test language
- teachmult: Instruction: Using multilingual didactics
- teachbeh: Teacher's instruction
- motselfs: L2 motivational self-system
- learnexp: L2 learning experience
- memor: Learning: Memorisation
- cstrat: Learning: Control strategies
- elab: Learning: Elaboration
- undrem: Learning: Understanding and rememoration
- metasum: Learning: Synthesis

[2] Takane and de Leeuw (1987) showed the mathematical equivalence between the CFA for the ordered categorical data model (Muthén, 1984) and the graded response model (Samejima, 1969). Furthermore, Maydeu-Olivares et al. (1994) showed that the graded response model and the GPCM (Muraki, 1992) produced very similar results. Hence, results based on the CFA model for ordered categorical data and based on the GPCM are similar. However, the construct scores were not directly comparable between the two ÜGK surveys in 2016 and 2017. The two student samples (ÜGK 2016: students of the 11th year of compulsory schooling; ÜGK 2017: students of the 8th year of compulsory schooling) were different; the metrics of the item and person parameters were estimated independently and were not aligned.

- `joyread:` Joy in reading
- `readrelev:` Relevance of reading
- `famedsup:` Family educational support
- `press:` Parental pressure for academic success
- `wealth:` Wealth index
- `cultposs:` Cultural possession index

For the secondary analyses, each researcher/research group could individually determine the assumptions, scaling model, and, correspondingly, the estimates of the construct scores they want to use. Of course, other estimates of construct scores based on other statistical models are also valid, providing they are justifiable.

# 3 Item- and model-fit statistics

Before the data imputation process, the item- and model-fit statistics were computed. During the data handling and scaling process, we worked together with the researchers who helped design the SQ (Erzinger, et al., 2019) to identify dodgy items and eliminated them from further analyses[3]. These statistics can also help researchers to judge whether the items and construct scores are good enough to be used in secondary analyses.

## 3.1 Item-fit statistics

At the item level, the following item-fit statistics were computed (cf. Itzlinger-Bruneforth, Kuhn, & Kiefer, 2016; Trendtel, Pham, & Yanagida, 2016; OECD, 2017, see Appendix C):

- Item difficulty index ($p$ = item mean score / item maximal score): items with a too high ($p > 0.95$) or too low ($p < 0.05$) difficulty index were identified as dodgy and eliminated from further analyses. Based on this criterion, two items `A15awealth3` ($p = 0.99$) and `A15awealth5` ($p = 0.98$), both from the construct “Wealth index”, were eliminated.
- Item discrimination ($r_{it}$ = item whole correlation for this item against the scale without this item, analyses based on polychoric correlation matrix between items): items with low discrimination ($r_{it} < 0.2$) were identified as dodgy items and eliminated from further analyses. Based on this criterion, the item `ABSQ002_r` ($r_{it} = -0.05$) from the construct “Test motivation” was eliminated.
- Model-based item infit statistics (infit = weighted mean standardized quadratic residuals between observed values and model-based expected values): identified items had an infit statistic higher than 1.1 (in analogy to PISA). However, this consideration is purely statistical and does not take the validity of the construct into account (Baghaei, 2008; Robitzsch, 2016). Hence, no items were eliminated only because they had a poor infit statistic (larger than 1.1).
- Differential item functioning (DIF) with regard to different language versions of the SQ: high DIF occurs when the students’ probability of choosing item responses not

[3] Five items `B13teach10`, `B13motselfs5`, `B13motselfs6`, `B13motselfs7`, `B14motselfs8` had already been identified as dodgy items based on the results of the pilot study. Therefore, they were neither included in the estimation of the construct scores nor in the imputation model.

only depends on the construct scores, but also depends on the language version of the SQ they received. For constructs with items with more than two answer categories, DIF was analyzed using ordinal logistic regression techniques (Crane, Gibbons, Jolley, & van Belle, 2006) with the R package **lordif** (Choi, with contributions from Gibbons, & Crane, 2016). To identify items with high uniform DIF, the proportional beta change criterion (with 0.05 as cut-off value) was applied. For constructs/scales with binary-coded items, logistic regression analysis (Zumbo, 1999) was performed to estimate item DIF using the R package **sirt** (Robitzsch, 2018). To flag items with high uniform DIF, the ETS classification system (Longford, Holland, & Thayer, 1993) was applied. An item with high DIF should only be eliminated from the measurement model of a construct if the reasons are believed to be not relevant to the construct (but to other language-relevant factors, cf. Trendtel, Schwabe, & Fellinger, 2016). No items were eliminated in the data handling process due to high DIF, since no construct-irrelevant reasons were identified[4].

Besides these fit statistics, the PCM and GPCM item parameters (cf. Trendtel, Pham, & Yanagida, 2016) can also be found in Appendix C.

Together with the item descriptive statistics and DIF indicators, conventional reliability estimates of the measures of the ÜGK 2017 SQ constructs were used to identify possible "poorly" fitting items and to decide whether an item should be eliminated from the calculation of the construct scores.

## 3.2 Model-fit statistics

The established model-fit statistics regarding the (uni)dimensionality of the construct, the reliability of the measurement, the model-fit statistics of the PCM and GPCM, and the reliability of the WLEs/EAPs can be found in Appendix D:

- Dimensionality: To examine whether a construct is unidimensional, we conducted a two-step analysis:
    - First, an explorative factor analysis based on a tetrachoric/polychoric correlation matrix was executed. The eigen values of the tetrachoric/polychoric matrix were calculated. A comparison of the scree slope of the factors of the observed data with that of the random data matrix of the same size as the original revealed the number of essential factors underlying the items (parallel analysis technique, Revelle & Rocklin, 1979). This was performed using the R package **psych** (Revelle, 2018).
    - Adopting the number of factors suggested in the first step, two strength indices including the McDonald's *omega hierarchical* ($omegaH$, Zinbarg et al., 2005) and the *Explained Common Variance* ($ECV$, Reise et al., 2013) were then computed. $omegaH$ is an estimate of the general factor saturation of a test. Zinbarg et al. (2005) compared McDonald's $omegaH$ to Cronbach's $alpha$ and Revelle's $beta$ and concluded that $omegaH$ is the best estimate for this purpose. The $ECV$ is the ratio of the general factor eigen value to the sum of all of the

[4] For some L2-specific constructs, DIF effects at item level were expected. Thus, this criterion was not applied to eliminate items of the L2-specific constructs. This should be taken into account seriously by comparing the L2-specific construct scores of students with different school languages. In addition, DIF analyses were not possible for constructs with three or less items.

eigen values. It is a good indicator of unidimensionality. These statistics were computed using the R package **psych** (Revelle, 2018).

- Reliability estimates of the measurements based on a tetrachoric/polychoric correlation matrix between items: Cronbach's Alpha (Cronbach, 1951) and Guttman's Lambda 6 (G6, Guttman, 1945). These statistics were computed using the R package **psych** (Revelle, 2018).
- Model fit statistics including the absolute fit statistics AIC, BIC, cAIC, AICc, SRMR, and SRMSR (Maydeu-Olivares, 2013; Robitzsch, Kiefer, & Wu, 2018). An SRMSR of zero indicates perfect fit; models with $SRMSR \leq .05$ have substantively negligible amount of misfit (Maydeu-Olivares, 2013). These fit statistics were obtained using the R package **TAM** (Robitzsch, Kiefer, & Wu, 2018).
- The reliability of the WLE and EAP estimates. These fit statistics were obtained using the R package TAM (Robitzsch, Kiefer, & Wu, 2018).

# 4 Construct scores in the raw dataset and in multiply imputed datasets

## 4.1 Construct scores in the raw dataset

Students who did not respond to all items of a construct had no construct scores in the raw dataset. To obtain construct scores based on raw data (with missing values), PCM-based and GPCM-based WLEs/EAPs were computed for all students with at least one valid response per construct. To calculate mean scores of constructs with missing values, the MCMC-based two-way imputation method of Van Ginkel et al. (2007) was applied to take into account the differences in item difficulties. The missing values were imputed ten times using this method; then, mean scores per construct for each person were calculated and then averaged.

## 4.2 Construct scores in the 20 imputed datasets

In each of the 20 imputed datasets, all missing response data were replaced by estimated values. After the data imputation process, the mean scores were computed in a straightforward manner within each imputed dataset. To obtain the PCM- and GPCM-based WLEs/EAPs, the calibration (estimation of item parameters) and scaling process (estimation of individual construct scores) were performed separately. In order to obtain the same item parameter metric across the 20 imputed datasets, the item calibration process was performed simultaneously using the data of all 20 datasets, by combining the 20 datasets and restructuring them into a long dataset (the 20 imputed datasets were placed below each other). After the calibration process, the individual construct scores were obtained via a scaling model with fixed item parameters from the first step. The item- and model-fit statistics were calculated per scaling model (each scaling model used data from only one imputed dataset), then pooled according to the Rubin's rule (Rubin, 1987) in order to report the results (Appendix E and F). Sample weights (see Verner & Helbling, 2019) were taken into account in both the calibration and scaling processes to account for the complex sample design and to yield unbiased population estimates.

The descriptive statistics for all items and scales based on the 20 imputed datasets can be found in Appendix B (columns `M_imp`, `SD_imp`, `M_SE_imp`).

### 4.3 Score equivalence between students with different school languages

To allow a comparison between the construct scores of students with different school languages, it is important to assess the score equivalence of students with different language versions of the SQ. For this purpose, the coefficient of determination (CD, see Sacchi & Oesch, 2017) was calculated. It is the correlation between the individual scores in the released datasets and the scores obtained from a scaling model using only data from one student subsample – students who share the same school language. The score equivalence with regard to school language is given for the PCM-based construct scores (see Appendix G, CD = 1 for all constructs). The CD for the GPCM-based construct scores is also high for all constructs ($CD \geq 0.96$) except for the "Cultural possession index" ($CD = 0.81$).

## 5 Some practical guidelines

For guidelines on how to deal appropriately with multiply imputed datasets, see Pham et al. (2019). A short R manual with guidance on how to read and analyze the 20 imputed datasets in the statistic environment R can be found in Appendix H.

## 6 References

Angelone, D., & Keller, F. (2019). *Überprüfung der Grundkompetenzen (ÜGK) in den Fächern Schulsprache und erste Fremdsprache im 8. Schuljahr. Technische Dokumentation zur Testentwicklung und Skalierung.* Aarau: Geschäftsstelle der Aufgabendatenbank EDK (ADB). Retrieved Sep. 21, 2020, from http://www.vecof-svizzera.ch/wp-content/uploads/2019/06/%C3%9CGK2017_Technischer-Bericht_ADB.pdf

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transactions, 22*, S. 1145–1146.

BIFIE. (2015). *Schülerfragebogen. Standardüberprüfung Deutsch 4. Schulstufe 2015.* Salzburg: BIFIE. Retrieved Sep. 21, 2020, from https://www.bifie.at/wp-content/uploads/2017/05/BIST_UE_D4_SFB.pdf

Birnbaum, A. (1968). Some Latent Trait Models and Their Use in Inferring an Examinee's Ability. In F. M. Lord, & M. R. Novick, *Statistical Theories of Mental Test Scores* (pp. 397–479). Reading/Mass: Addison-Wesley.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a micro computer environment. *Applied psychological measurement, 6*, pp. 431–444.

Böhm-Kasper, O., Bos, W., Körner, S., & Weishaupt, H. (2001). EBI - Das Erfurter Belastungs-Inventar zur Erfassung von Belastung und Beanspruchung von Lehrern und Schülern am Gymnasium: Darstellung des Instruments und erste Ergebnisse aus dem Pretest. In H. Merkens, & H. Weishaupt, *Schulforschung und Schulentwicklung. Aktuelle Forschungsbeiträge (Erfurter Studien zur Entwicklung des Bildungswesens, Bd. 14)* (pp. 35–66). Erfurt: Universität Erfurt.

Choi, S. W., with contributions from Gibbons, L. E., & Crane, P. K. (2016). lordif: Logistic Ordinal Regression Differential Item Functioning using. *R package version 0.3-3*. Retrieved Jul. 05, 2018, from https://CRAN.R-project.org/package=lordif

Crane, P. K., Gibbons, L. E., Jolley, L., & van Belle, G. (November 2006). Differential Item Functioning Analysis With Ordinal Logistic Regression Techniques. DIFdetect and difwithpar. *Medical Care, 44*(11 Suppl 3), S. 115–123.

Cronbach, L. J. (1951). Coefficient Alpha and the internal structure of tests. *Psychometrika, 16*(3), pp. 297–334.

DAB . (2016). *Codebuch Schülerinnen und Schüler DAB-Panelstudie Konzepte und Skalen Befragungswelle 1: Januar/Februar 2012.*

Dörnyei, Z. (2009). The L2 motivational self system. In Z. Dörnyei, & E. Ushioda, *Motivation, language identity and the L2 self* (pp. 9–42). Clevedon: Multilingual Matters.

Erzinger, A., Hauser, M., Dutrevis, M., Hascher, T., Keller, R., Lenz, P., & Soucis, A. (2019). *Erläuterungen zu den Skalen des Kontextfragebogens der ÜGK Sprachen 2017: Theoretischer Hintergrund, Inhalte und Konstrukte.* Bern & St. Gallen: Universität Bern, Pädagogische Hochschule St. Gallen, Service de la recherche en éducation (SRED), Universität Fribourg. Retrieved Sep. 21, 2020, from http://uegk-schweiz.ch/wp-content/uploads/2019/09/%C3%9CGK2017_Sprachen_KFB_def.pdf

European Survey on Language Competencies (ESLC). (2011). Student questionnaire for the European Survey on Language Competences. 2011 main study. Retrieved Sep. 24, 2020, from https://ec.europa.eu/education/

Guttman, L. (1945). A basis for analyzing test-retest reliability. *Psychometrika, 10*(4), pp. 255–282.

Hascher, T. (2004). *Wohlbefinden in der Schule.* Münster: Waxmann.

Hascher, T., Brühwiler, C., Erzinger, A., Girnat , B., & Hagenauer, G. (2016). *Erläuterungen zu den Skalen des Kontextfragebogens Mathematikteil: Theoretischer Hintergrund und Forschungsinteressen.* Bern, St. Gallen, Basel: Unveröffentlichter Projektbericht.

Hascher, T., Hagenauer, G., & Schaffer, A. (2011). Wohlbefinden in der Grundschule. *Erziehung und Unterricht, 161*(3-4), pp. 381–392.

Heinzmann, S. (2013). *Young language learners' motivation and attitudes: longitudinal, comparative and explanatory perspectives.* London: Bloomsbury.

Hupka-Brunner, S., Jann, B., Meier, T., Imdorf, C., Sacchi, S., Müller , B., . . . Becker, R. (2015). *Erläuterungen zum Kontextfragebogen der ÜGK 2016: Allgemeiner Teil.* Bern: Unveröffentlichter Projektbericht.

ICILS. (2015). *International Association for the Evaluation of Educational Achievement (IEA), ICILS 2013 User Guide for the International Database.* Amsterdam: IEA Secretariat.

Itzlinger-Bruneforth, U., Kuhn, J.-T., & Kiefer, T. (2016). Testkonstruktion. In C. Schreiner, & S. Breit, *Large-Scale Assessment mit R: Methodische Grundlagen der österreichischen Bildungsstandard-Überprüfung* (pp. 21–50). Vienna: UTB.

Konsortium ÜGK. (2019). *Überprüfung der Grundkompetenzen. Nationaler Bericht der ÜGK 2017: Sprachen 8. Schuljahr.* Bern & Geneve: EDK & SRED. doi:10.18747/PHSGcoll3/id/385

Longford, N. T., Holland, P. W., & Thayer, D. T. (1993). Stability of the MH D-DIF statistics across populations. In P. W. Holland, & H. Wainer, *Differential Item Functioning* (S. 171–196). Hillsdale, NJ: Erlbaum.

Marsh, H. W. (1993). The multidimensional structure of academic self-concept: Invariance over gender and age. *American Educational Research Journal, 30*(4), pp. 841–860.

Master, G. N., & Wright, B. D. (1997). The Partial Credit Model. In W. J. van der Linden, & R. K. Hambleton, *Handbook of Modern Item Response Theory* (S. 101–121). New York: Springer.

Maydeu-Olivares, A. (2013). Goodness-of-fit assessment of item response theory models (with discussion). *Measurement: Interdisciplinary Research and Perspectives, 11*, pp. 71–137. doi:10.1080/15366367.2013.831680

Maydeu-Olivares, A., Drasgow, F., & Mead, A. D. (1994). Distinguishing among parametric item response models for polychotomous ordered data. *Applied Psychological Measurement, 18*(13), S. 245–256.

Muraki, E. (1992). A generalized partial credit model: Application of an EM algorithm. *Applied Psychological Measurement, 16*, S. 159–176. doi:10.1177/014662169201600206

Muthén, B. (1984). A general structural equation model with dichotomous, ordered categorical, and continuous latent variable indicators. *Psychometrika,, 49*, pp. 115–132.

OECD. (2002). *Programme for International Student Assessment (PISA): Manual for the PISA 2000 Database.* Paris: OECD Publishing.

OECD. (2005). *PISA 2003 Technical Report.* Paris: OECD. Retrieved Sep. 21, 2020, from http://www.oecd.org/education/school/programmeforinternationalstudentassessmentpisa/35188570.pdf

OECD. (2011). PISA 2009 Results: Learning to Learn – Student Engagement, Strategies and Practices (Vol. III). Retrieved Sep. 24, 2020, from https://www.oecd-ilibrary.org/education/pisa-2009-results-learning-to-learn_9789264083943-en

OECD. (2013). *PISA 2012 Assessment and Analytical Framework: Mathematics, Reading, Science, Problem Solving and Financial Literacy.* Paris: OECD Publishing.

OECD. (2014). *PISA 2012 Technical report.* Paris: OECD Publishing.

OECD. (2017). *PISA 2015 Technical report.* Paris: OECD Publishing.

OECD. (2020). PISA 2018 Technical Report, OECD Website (Draft). Retrieved Sep. 24, 2020, from https://www.oecd.org/pisa/pisa-for-development/pisafordevelopment2018technicalreport

Pham, G., Helbling, L., Verner, M., & Ambrosetti, A. (2019). *ÜGK – COFO – VeCoF 2017 results: Technical appendices.* St.Gallen & Geneva: St.Gallen University of Teacher Education & Service de la recherche en éducation (SRED).

Rasch, G. (1960). *Studies in mathematical psychology: I. Probabilistic models for some intelligence and attainment tests.* Oxford, England: Nielsen & Lydiche.

Reise, S. P., Scheines, R., Widaman, K. F., & Haviland, M. G. (2013). Multidimensionality and Structural Coefficient Bias in Structural Equation Modeling: A Bifactor Perspective.

*Educational and Psychological Measurement, 73*(1), pp. 5–26. doi:10.1177/0013164412449831

Revelle, W. (2018). psych: Procedures for Psychological, Psychometric, and Personality Research. *R package version 1.8.4*. Evanston, Illinois. Retrieved Jul. 05, 2018, from https://CRAN.R-project.org/package=psych

Revelle, W., & Rocklin, T. (1979). Very simple structure - alternative procedure for estimating the optimal number of interpretable factors. *Multivariate Behavioral Research, 14*(4), pp. 403–414.

Robitzsch, A. (2016). *Essays zu methodischen Herausforderungen im Large-Scale Assessment.* Berlin: Humboldt-Universität zu Berlin. Retrieved Sep. 21, 2020, from https://edoc.hu-berlin.de/handle/18452/18076

Robitzsch, A. (2018). sirt: Supplementary item response theory models. *R package version 2.6-9*. Retrieved Jul. 05, 2018, from https://CRAN.R-project.org/package=sirt

Robitzsch, A., Kiefer, T., & Wu, M. (2018). TAM: Test analysis modules. *R package version 2.12-18*. https://CRAN.R-project.org/package=TAM.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys.* Hoboken, NJ: Wiley.

Sacchi, S., & Oesch, D. (2017). *ÜGK 2016: Assessment of mathematics skills: Documentation of questionnaire-based scales.* TREE / University of Bern: Bern. Retrieved Sep. 21, 2020, from https://boris.unibe.ch/143394/

Samejima, F. (1969). Estimation of latent ability using a response pattern of graded scores. *Psychometrika Monograph No. 17, 34*(4).

Shavelson, R. J., Jubner, J. J., & Stanton, G. C. (1976). Self-Concept: Validation of Construct Interpretations. *Review of Educational Research, 46*(3), pp. 407–441.

Takane, Y., & de Leeuw, J. (1987). On the relationship between item response theory and factor analysis of discretized variables. *Psychometrika, 52*(3), S. 393–408.

TREE. (2016). *Konzepte und Skalen. Befragungswellen 1 bis 9, 2001-2015.* Bern: TREE.

Trendtel, M., Pham, G., & Yanagida, T. (2016). Skalierung und Linking. In C. Schreiner, & S. Breit, *Large-Scale Assessment mit R: Methodische Grundlagen der österreichischen Bildungsstandard-Überprüfung* (pp. 185–224). Vienna: UTB.

Trendtel, M., Schwabe, F., & Fellinger, R. (2016). Differenzielles Itemfunktionieren in Subgruppen. In C. Schreiner, & S. Breit, *Large-Scale Assessment mit R: Methodische Grundlagen der österreichischen Bildungsstandard-Überprüfung* (pp. 111–148). Vienna: UTB.

Van Ginkel, J. R., Van der Ark, A., Sijtsma, K., & Vermunt, J. K. (2007). Two-way imputation: A Bayesian method for estimating missing scores in tests and questionnaires, and an accurate approximation. *Computational Statistics & Data Analysis, 51*, pp. 4013–1027.

Venetz, M., Zurbriggen, C., Eckhart, M., Schwab, S., & Hessels, M. (2015). The Perceptions of Inclusion Questionnaire (PIQ). English Version. Retrieved Sep. 24, 2020, from www.piqinfo.ch

Verner, M., & Helbling, L. (2019). *Sampling ÜGK 2017. Technischer Bericht zu Stichprobendesign, Gewichtung und Varianzschätzung bei der Überprüfung des Erreichens der Grundkompetenzen 2017.* Zurich: Institut für Bildungsevaluation, assoziiertes Institut der Universität Zürich. Retrieved Sep. 21, 2020, from http://uegk-schweiz.ch/wp-content/uploads/2019/05/%C3%9CGK2017_Verner_Helbling_2019_Sampling_%C3%9CGK_2017.pdf

Warm, T. (1989). Weighted Likelihood Estimation of Ability in Item Response Theory. *Psychometrika, 54*, pp. 427–450.

Zinbarg, R. E., Revelle, W., Yovel, I., & Li, W. (2005). Cronbach's Alpha, Revelle's Beta, McDonald's Omega: Their relations with each and two alternative conceptualizations of reliability. *Psychometrika, 70*, pp. 123–133. Retrieved Sep. 21, 2020, from https://personality-project.org/revelle/publications/zinbarg.revelle.pmet.05.pdf

Zumbo, B. D. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-type (ordinal) item scores.* Ottawa ON: Directorate of Human Resources Research and Evaluation, Department of National Defense.