ÜGK – COFO – VECOF 2016 results: Technical appendices

Giang Pham, Laura Helbling, Martin Verner,
Franck Petrucci, Domenico Angelone & Alice Ambrosetti

Impressum

Autors	Giang Pham (PHSG), Laura Helbling, Martin Verner (IBE), Franck Petrucci (SRED and CIRSE), Domenico Angelone (ADB), Alice Ambrosetti (CIRSE)
Quotation proposal	Pham, G., Helbling, L., Verner, M., Petrucci, F., Angelone, D. & Ambrosetti, A. (2019). ÜGK – COFO – VeCoF 2016 results: Technical appendices. St.Gallen & Genf: Pädagogische Hochschule St.Gallen (PHSG) & Service de la recherche en éducation (SRED).
Download	www.cofo-suisse.ch/cofo-2016
Layout	Narain Jagasia (SRED)

Table of contents

1 Context variables: social background, home language and migration status

1.1 Social background

The ÜGK social background index (or socioeconomic status - SES) is a composite score. Its calculation is based on three indicators: the highest parental occupational status, the highest parental education level, and the number of books at home. This procedure is in line with the indicators used in the international computer and information literacy study (ICILS, Schulz & Friedman, 2015), the educational standard survey (BIST-Ü) in Austria (Pham et al., 2014), and represents an adaptation of the index of economic, social and cultural status (ESCS) as used in PISA 2012 (OECD, 2014).

1.1.1 Highest parental occupational status

The parental occupations were obtained via student responses (open-response format) to questions MB08 and MB10 in the student questionnaire. The student responses on parental occupations were coded into four-digit codes according to the International standard classification of occupations (ISCO-08) framework (Ganzeboom & Treiman, 2008; Ganzeboom, De Graaf, & Treiman, 1992), then transformed to the international socioeconomic index of occupational status (ISEI-08; Ganzeboom, 2010a, 2010b). These codes are contained in the variables `occupm_isei08` (occupational status of mother – ISEI-08 status) and `occupf_isei08` (occupational status of father – ISEI-08 status).

The highest occupational status of parents (`hisei08`) corresponds to the higher value between `occupm_isei08` and `occupf_isei08`, in case both items were answered. If only one value is available, `hisei08` corresponds to this value. The variable `hisei08` is missing, if both parental occupation items were not answered.

In order to construct the social background index for the national report, all missing values of `occupm_isei08` and `occupf_isei08` were multiply imputed (see chapter 2). Within each imputed dataset, the value of `hisei08` corresponded to the higher value of `occupm_isei08` and `occupf_isei08`.

1.1.2 Highest parental education level

Parental education was assessed by means of questions MB12a and MB13a in the student questionnaire. Based on the following options, students reported on the highest educational attainment of their mother (`meduc_org`) and father (`feduc_org`):

- 1 = never attended school
- 2 = primary level education (4-6 years)
- 3 = compulsory education (primary and lower secondary levels, 7-9 years)
- 4 = upper secondary level VET (including Handels(mittel)schule, Fachmittelschule (formerly Diplommittelschule))
- 5 = Baccalaureate (general or vocational, including former primary teacher training diploma)
- 6 = non-university tertiary level VET (e.g. Eidg. Fachausweis, Meisterdiplom)
- 7 = Tertiary level university (including HTL, HWV, Fachhochschulen [UAS], Pädagogische Hochschulen)
- 8 = Other education or training, that is (open response)
- 19 = I don’t know

In the cleaning process, category 8 (other education or training) was recoded into one of the other seven categories using students' open responses whenever possible. Category 19 was treated as missing. Variables `meduc` and `feduc` contain the cleaned data.

Two new variables `medu` (mother's highest educational attainment) and `fedu` (father's highest educational attainment) were created by recoding `meduc` and `feduc` and reducing them into the following categories:

- 0 = compulsory schooling only
- 1 = upper secondary education
- 2 = tertiary education
- 8 = other

The recoding rules were decided based on the absolute frequency distribution of the seven original categories and the average student achievement in mathematics at two levels: the national level and the linguistic-regional level. In addition, corresponding data of the ÜGK 2017 survey were considered as well, since identical coding rules and calculation of the social background index in both studies were intended.

In the raw dataset, the highest parental educational level (`fmedu`) corresponds to the higher value of `medu` and `fedu` (category 8 was treated as a missing code in recoding process). In case one of these two values is missing, the value of `fmedu` corresponds to the only available value. If both values are missing, `fmedu` has missing value.

The highest parental educational level (`fmedu`) corresponds to the higher value between `medu` and `fedu`, in case both items were answered (category 8 was treated as a missing value during the recoding process). If only one value is available, `fmedu` corresponds to this value. The variable `fmedu` is missing, if both parental education items were not answered.

In order to construct the SES for the national report, all missing values of `medu` and `fedu` were multiply imputed (see chapter 2). Within each imputed dataset, the value of `fmedu` corresponds to the higher value of `medu` and `fedu`.

1.1.3 Number of books at home

The third indicator for the social background index is based on student responses to question F18 in the student questionnaire. Students reported the numbers of books at home by choosing one of the following options (variable `books`):

- 1 = none
- 2 = 1-10 books
- 3 = 11-50 books
- 4 = 51-100 books
- 5 = 101-250 books
- 6 = 251-500 books
- 7 = more than 500 books

On this basis, a new variable `nbooks` was created to construct the index of social background by recoding variable `books` into the following five categories:

- 0 = 0-10 books
- 1 = 11-50 books
- 2 = 51-100 books
- 3 = 101-250 books
- 4 = more than 250 books

The recoding rules were decided based on the frequency distribution of the seven original categories and the average student achievement in mathematics at the national level as well as within each of the three linguistic regions. Corresponding data of the ÜGK 2017 survey were considered as well to enable identical coding rules and calculation of the social background index in both studies.

To construct the social background index for the national report, all missing values of `nbooks` were multiply imputed (see chapter 2).

Notes:

In PISA, one of the three indices incorporated in the ESCS is the index of household possessions, which comprised all items on the family wealth possessions (`wealth`), cultural possessions (`cultpos`), home educational resources (`hedres`) and the number of books at home (OECD, 2014, p. 316, 351). In ÜGK, some items of `wealth`, `cultpos` and `hedres` scales were included in the student questionnaire, however, they were not used to construct the index of social background due to the following reasons:

- High percentages of missing values in ÜGK 2016: Since two student questionnaire versions were used in ÜGK 2016, only about 50% of the survey sample reported on possessions and educational resources (Sacchi & Oesch, 2017).
- Problematic psychometric parameters: The mean scores of several items were very high, e.g. internet connection is available in more than 95% families. Several items correlated not at all or negatively with student achievement in mathematics. Differential item functioning in different linguistic regions was found for one item of the cultural possessions scale (possession of classical literature at home). While the number of books at home was a statistically significant positive predictor of student achievement in mathematics, almost all other items had no predictive power after controlling for the effect of number of books at home, as suggested by multiple regression analyses.
- The number of books at home could be seen as an indicator of both factors representing the wealth and cultural possession indices: Parallel analysis based on a polychoric correlation matrix of all items (number of books at home and all wealth and cultural possession items) suggested that there were two dominant factors underlying all these items. Results of an explorative factor analysis with two factors showed that all wealth items loaded highly positively on one factor and not on the other factor; all cultural possession items loaded highly positively only on the other factor; `nbooks` had high positive loadings on both factors.

In other studies such as the ICILS 2013 (Schulz & Friedman, 2015) or the BIST-Ü in Austria (Pham, Freunberger, & Robitzsch, 2014), wealth, cultural possessions and home educational resources scales were not involved in constructing the index of social background.

1.1.4 Calculation

The number of books at home `nbooks` was the strongest predictor of student achievement in different domains among three indicators of the social background index (mathematics, ÜGK 2016: *r* = .38, *p* < .001; L1-reading, ÜGK 2017: *r* = .36, *p* < .001). Therefore, this variable should not have lower weight than the other two variables (`hisei08` and `fmedu`) in computing the social background index. This would be the case, if the same statistical approach as in PISA 2012 were applied (component scores for the first principal component, OECD, 2014, p. 352). The two indices `hisei08` and `fmedu` correlated namely stronger with each other (*r* = .43) than with the number of books at home (*r* = .29-.41). In ÜGK 2016 and ÜGK 2017, the normative weights of all three indices were set equal while calculating the social background index. The same approach was applied in the educational standard survey in Austria (Pham, Freunberger, & Robitzsch, 2014).

The calculation of the ÜGK social background index is represented by the following formula:

$$SES_2 = zSES_1,$$

$$SES_1 = \frac{zhisei08 + zfmedu + znbooks}{3},$$

`zhisei08`, `zfmedu` and `znbooks` are the *z*-scores of the three basic indices (`hisei08`, `fmedu` and `nbooks`). Weighted data (using sampling weights) were used to standardize variables.

For the national report, 100 imputed datasets (see chapter 2) were applied. First, the SES_1 – the weighted mean of `zhisei08`, `zfmedu` and `znbooks` – and SES_2 – the *z*-score of SES_1 (using weighted data) – were calculated for each imputed dataset. Then, the final `SES` variable – the social background index – was calculated by transforming SES_2 in each imputed dataset as follows:

$$SES = \frac{SES_2 - \mu_{SES_2}}{\sigma_{SES_2}},$$

μ_{SES_2} represents the overall weighted mean and σ_{SES_2} the overall weighted standard deviation of SES_2 over all imputed datasets (see chapter 3). For this reason, `SES` has an overall weighted mean of zero and an overall weighted standard deviation of one over all imputed datasets.

In order to compute the social background index `SES` based on the raw data, an appropriate approach to deal with missing data should be considered first. Then, the same procedure can be applied to calculate the social background index.

1.2 Home language

Questions MB17a to MB18c in the student questionnaire asked students about their main and second languages spoken at home. Variables `langhome_org`, `langhome_a` contain student responses in regard to the main language spoken at home; variables `langhome2f`, `langhome2_org` and `langhome2_a` contain student responses in regard to the second language spoken at home, if available.

Open responses were considered during data cleaning. Variables `langhome`, `langhome2f` and `langhome2` contain recoded data. Based on these three variables, two new variables (`homelang1` and `homelang2`) were created:

- homelang1: the main language spoken at home is the school language (0 = false, 1 = true)
- homelang2: the second language spoken at home is the school language (0 = false, 1 = true)

The final variable regarding home language (`homelang`) is coded using the same definition as in PISA 2015 (OECD, 2016, p. 243) based on data of three variables `homelang1`, `homelang2f` and `homelang2`. The variable contains three levels:

- `homelang` = 1: only the school language is spoken at home
- `homelang` = 2: the school language and another language are regularly spoken at home
- `homelang` = 3: the school language is not spoken at home

The coding rules were different for different linguistic regions in Switzerland:

- In the German language region Swiss German, Liechtenstein dialect, and Standard German was treated as the school language.
- In the French language region French only (no dialect option in the questionnaire) was treated as the school language.
- In the Italian language region Italian and its dialects were treated as the school language.

For the national report, imputed datasets were used. All missing values of the three basic variables `homelang1`, `homelang2f`, and `homelang2` (if available) were multiply imputed (see chapter 2). Within each imputed dataset, the variable `homelang` was derived from these three basic variables. The reported results were derived based on the pooled results over all imputed datasets (see chapter 3).

1.3 Immigration status

The immigration status in ÜGK 2016 was defined identically as in PISA 2015 (OECD, 2016, p. 243) using three categories:

- *Non-immigrant students* or *'students without an immigrant background'* are those whose mother or father or both was/were born in Switzerland, regardless of the birth place of the student.
- *Immigrant students* or '*students with an immigrant background*' are those whose mother and father were *both* not born in Switzerland. Among them, a distinction is made between students who were born in Switzerland and students who were born abroad:
    - *First-generation immigrant students* are foreign-born students whose parents are both foreign-born.
    - *Second-generation immigrant students* are students who were born in Switzerland and whose parents are both foreign-born.

Question MB14 in the student questionnaire asked students about their country of birth (variable `cobs`) as well as the country of birth of their mother (variable `cobm`) and father (variable `cobf`).

Based on students' responses, three new variables were coded, which indicate whether the student (`cobs_frgn`), the mother (`cobm_frgn`), and the father (`cobf_frgn`) was born abroad (value = 1) or in Switzerland (value = 0).

For the national report, all missing values of the three basic variables `cobs_frgn`, `cobm_frgn`, and `cobf_frgn` were first multiply imputed (see chapter 2). Within each imputed dataset, the variable `immig_pisa` was derived from these three basic variables with three categories corresponding to the above mentioned definition:

- immig_pisa = 1: Non-immigrant student.
- immig_pisa = 2: Second-generation immigrant student.
- immig_pisa = 3: First-generation immigrant student.

The reported results regarding immigration status were the pooled results over all imputed datasets (see chapter 3).

In the raw dataset, an appropriate approach to deal with missing data should be considered first. Then, variable `immig_pisa` can be derived using the same rules based on three variables `cobs_frgn`, `cobm_frgn`, and `cobf_frgn`.

2 Dealing with missing values of context variables

Missing values of context variables could lead to biased estimates. Based on the technique of multiple imputation (MICE, Multiple Imputation by Chained Equations, see van Buuren, 2012; Robitzsch, Pham & Yanagida, 2016), missing values of these variables were imputed multiple times utilizing (correlated) observed student information and taking into account the hierarchical structure of the data (students nested within schools). For this purpose, the R-package miceadds (see Robitzsch, Grund & Henke, 2018) was applied. Separately by canton, each missing value was imputed five times based on one plausible value set (for plausible values see Angelone & Keller, 2019; the imputation approach is called nested multiple imputation by plausible value, cf. Shen, 2000; Rubin, 2003). Given 20 plausible value sets of student achievement, this resulted in 100 (20 x 5) nested multiply imputed datasets, which served as the basis for all reported results and analyses in the national report.

3 Estimation of descriptive results and measurement errors

All results including confidence intervals of the test data – if the context data were not involved – were estimated using standard combining rules based on 20 plausible values (Rubin's rule, Rubin, 1987). All results involving the context data were estimated using the modified combining rules for nested multiply imputed datasets (Rubin, 2003). In addition, due to the complex sampling design (see Verner & Helbling, 2019), there was some disproportionalities in the sample data. All analyses, referring to population measures, were conducted using sampling weights and replicate weights to take this into account (cf. OECD, 2017; Bruneforth et al., 2016; Foy, 2012; Enders, 2010). Analyses for the report were performed using the R-package BIFIEsurvey (BIFIE, 2018). There were exceptions: multilevel analyses (chapter 5.2.3 in the national report) were performed using slightly different data basis, methodological approach and software (see section 4.2).

3.1 Estimation of point estimates using multiply imputed datasets

3.1.1 Point estimates involving only test data

All reported point estimates involving only test data at different levels (the proportion of students who achieved the minimum standards in mathematics) were pooled estimates using 20 plausible values. This means that each analysis was performed 20 times, each time based on one plausible value. Afterwards, all 20 estimates were pooled to yield the final result. The *pooled point estimate* $\hat{\mu}$ (e.g. mean, effect size) is the arithmetic average over all 20 estimates $\hat{\mu}_i$ (i *= 1, 2… 20*):

$$\hat{\mu} = \frac{\sum_{i=1}^{20} \hat{\mu}_i}{20}$$

3.1.2 Point estimates involving context variables

All reported point estimates involving the context variables at different levels (e.g. the proportion of students without migration status) were pooled estimates using 100 imputed datasets (five imputed datasets per plausible value). This means that each analysis was performed 100 times, each time based on one imputed dataset. Afterwards, all 100 estimates were pooled to yield the final result. The *pooled point estimate* $\hat{\mu}$ (e.g. mean, effect size) is the arithmetic average over all respective 100 estimates $\hat{\mu}_{i,j}$ (i *= 1, 2… 20;* j *= 1, 2… 5*):

$$\hat{\mu} = \frac{1}{20 \cdot 5} \sum_{i=1}^{20} \sum_{j=1}^{5} \hat{\mu}_{i,j}$$

3.2 Estimation of measurement errors and confidence intervals of the point estimates

3.2.1 Measurement errors and confidence intervals of point estimates involving only test data

The estimation variance of a point estimate $\hat{\mu}$ involving only test data was calculated by combining two components: the variance component within each plausible value i $V_{Samp,i}(\hat{\mu})$ (within-imputation variance or sampling variance) and the variance component caused by variation between plausible values $V_{Imp}(\hat{\mu})$ (between-imputation variance, cf. Mislevy et al., 1992).

The between-imputation variance $V_{Imp}(\hat{\mu})$ is the product of the sum of squares of differences between each estimate $\hat{\mu}_i$ and the pooled estimate $\hat{\mu}$ with a constant factor:

$$V_{Imp}(\hat{\mu}) = \left(1 + \frac{1}{20}\right) \cdot \sum_{i=1}^{20} (\hat{\mu}_i - \hat{\mu})^2$$

The within-imputation variance was estimated using Fay's method (Judkins, 1990) as applied in PISA (OECD, 2017). For this purpose, 120 replicate zones were generated (Verner & Helbling, 2019). The point estimate of interest $\hat{\mu}_{r,i}$ was calculated within each replicate zone r (*r = 1, 2… 120*) with corresponding replicate weights. The variance of $\hat{\mu}_{r,i}$ over all 120 replicate zones represents the within-imputation variance per plausible value i and was calculated with a Fay factor of 0.5:

$$V_{Samp,i}(\hat{\mu}_i) = \frac{1}{120 \cdot 0.5^2} \cdot \sum_{r=1}^{120} (\hat{\mu}_{r,i} - \hat{\mu}_i)^2$$

The sampling variance of the pooled estimate $\hat{\mu}$ over all 20 plausible values is:

$$V_{Samp}(\hat{\mu}) = \frac{\sum_{i=1}^{20} V_{Samp,i}(\hat{\mu}_i)}{20}$$

Altogether, the estimation variance of $\hat{\mu}$ is:

$$V_{Total}(\hat{\mu}) = V_{Imp}(\hat{\mu}) + V_{Samp}(\hat{\mu})$$

The measurement error SE of each point estimate $\hat{\mu}$ corresponds to the square root of the estimation variance:

$$SE(\hat{\mu}) = \sqrt{V_{Total}(\hat{\mu})}$$

Finally, the lower and upper bounds of the 95% confidence interval of each reported result were calculated. This statistical interval represents a range of values that might contain (with 95% confidence level) the true value of the result of interest. Unless otherwise indicated, the lower (KI_{low}) and upper (KI_{upp}) bound of this interval were calculated as follows:

$$KI_{low}(\hat{\mu}) = \hat{\mu} - 1.96 \cdot SE(\hat{\mu});\ KI_{upp}(\hat{\mu}) = \hat{\mu} + 1.96 \cdot SE(\hat{\mu})$$

3.2.2 Measurement errors and confidence intervals of point estimates involving context variables

The variance of a point estimate $\hat{\mu}$ involving context variables was calculated slightly differently, since the imputed datasets within a nest (5 imputed datasets were nested within one plausible value set) were correlated.

In order to calculate the between-imputation variance $V_{Imp}(\hat{\mu})$, two components were considered: the between-nest (between-plausible values) variance $V_{Imp,b}(\hat{\mu})$ and the within-nest (within plausible value) variance $V_{Imp,w}(\hat{\mu})$.

Let $\hat{\mu}_{i,j}$ (*j* = 1…5) be the estimate based on the *j*th imputed dataset nested within the *i*th plausible value, $\hat{\mu}_i$ – the average estimate related to the *i*th plausible value – was calculated as follow:

$$\hat{\mu}_i = \frac{\sum_{j=1}^{5} \hat{\mu}_{i,j}}{5}$$

The between-nest variance is:

$$V_{Imp,b}(\hat{\mu}) = \frac{5}{20-1}\sum_{i=1}^{20}(\hat{\mu}_i - \hat{\mu})^2$$

The within-nest variance is:

$$V_{Imp,w}(\hat{\mu}) = \frac{1}{20 \cdot (5-1)}\sum_{i=1}^{20}\sum_{j=1}^{5}(\hat{\mu}_{i,j} - \hat{\mu}_j)^2$$

Now, the between-imputation variance $V_{Imp}(\hat{\mu})$ is:

$$V_{Imp}(\hat{\mu}) = \frac{1}{5}\left(1 + \frac{1}{20}\right)V_{Imp,b}(\hat{\mu}) + (1 - \frac{1}{5})V_{Imp,w}(\hat{\mu})$$

The within-imputation variance was estimated identically as described in section 3.2.1, now based on all 100 imputed datasets. The point estimate $\hat{\mu}_{r,i,j}$ was calculated within each replicate zone r (*r = 1, 2… 120*) with corresponding replicate weights. The variance of $\hat{\mu}_{r,i,j}$ over all 120 replicate zones represents the within-imputation variance per imputed value set *j* per plausible value set *i* and was calculated with the Fay factor equal 0.5 as follows:

$$V_{Samp,ij}(\hat{\mu}_{i,j}) = \frac{1}{120 \cdot 0.5^2} \cdot \sum_{r=1}^{120}(\hat{\mu}_{r,i,j} - \hat{\mu}_{i,j})^2$$

The sampling variance of the pooled estimate $\hat{\mu}$ over all 100 imputed datasets is:

$$V_{Samp}(\hat{\mu}) = \frac{\sum_{1}^{100} V_{Samp,ij}(\hat{\mu}_{i,j})}{100}$$

Altogether, the total estimation variance of $\hat{\mu}$ is:

$$V_{Total}(\hat{\mu}) = V_{Imp}(\hat{\mu}) + V_{Samp}(\hat{\mu})$$

The measurement error SE of each point estimate $\hat{\mu}$ corresponds to the square root of the estimation variance. The lower and upper bounds of the 95% confidence interval of each reported result were calculated identically as described in section 3.2.1 above:

$$SE(\hat{\mu}) = \sqrt{V_{Total}(\hat{\mu})}$$

$$KI_{low}(\hat{\mu}) = \hat{\mu} - 1.96 \cdot SE(\hat{\mu});\ KI_{upp}(\hat{\mu}) = \hat{\mu} + 1.96 \cdot SE(\hat{\mu})$$

3.2.3 Notes

By implementing the aforementioned procedures, an infinite population was assumed during the calculation of sampling variances. Employing this procedure, the cantonal sampling variances were not adjusted for the (unequal) sampling rates in cantons (*no* finite population correction was applied). As a result, for small cantons with comparatively large shares of students participating (e.g., full census cantons), the sampling variance might be large despite full census. With this, we intended to take the possible cohort effect into account. Results of one student cohort might be different from results of another student cohort under the same educational framework and conditions. The cohort effect might be larger in small cantons due to small sample size. If the finite population correction method were applied to calculate the sampling variance, results of small cantons would often differ statistically significantly from the average, even in case the difference were very small. This could sometimes lead to difficulties in interpreting the results.

Therefore, we decided to apply this rather conservative approach in estimating the variance of point estimates, which was applied in PISA (OECD, 2017) as well.

3.3 Calculation and interpretation of Cohen's *d*

Beside the absolute difference and the statistical significance of differences between any two groups, the effect size Cohen's d (Cohen, 1988) was calculated and reported.

Statistically, an effect size is defined as follows:

$$d = \frac{\delta}{SD}$$

δ is the absolute difference between two groups, SD is the pooled sample standard deviation:

$$\delta = \hat{\mu}_1 - \hat{\mu}_2$$

$$SD = \sqrt{({SD_1}^2 + {SD_2}^2)/2}$$

$\hat{\mu}_1$ and $SD_1$are the estimate and corresponding sample standard deviation in the first group, $\hat{\mu}_2$ and SD_2 are the estimate and corresponding sample standard deviation in the second group. The reported d values were calculated based on all imputed datasets as described in section 3.1.

All reported Cohen's *d* effect sizes were derived as mentioned above, except for comparisons between cantonal and national levels (shown in part 2 of the report). To calculate the effect size regarding the difference between a population (e.g. Switzerland) and one of its sub-sample (i.e. canton), the population standard deviation was used instead of the pooled standard deviation.

Cohen (1988) suggested that $d \geq 0.2$ can be interpreted as a small, $d \geq 0.5$ a medium, and $d \geq 0.8$ a large effect size. Hattie (2009, p. 9) suggested $d \geq 0.2$ for small, $d \geq 0.4$ for medium, and $d \geq 0.6$ for large effect size when judging educational outcomes. In this report, we used the suggestions of Hattie to interpret the effect sizes.

3.4 A note on the results of the subscales

The standard setting method used in PISA was adopted to determine the cut-off value between two levels – minimum standard attained and not attained – based on the whole item pool and the total test scores in mathematics (Angelone & Keller, 2019). By applying this approach, the same cut-off value was assumed for all subscales in mathematics. Therefore, the proportions of students who achieved the minimum standard in all subscales (*absolute* results) at the national level were all identical to the result in mathematics as a whole. Only if this assumption holds, the absolute results of the subscales can be interpreted.

Of each subscale, a comparison between the cantonal and the national result can still be made. Over all subscales, these differences can be considered together to judge whether a canton has special strengths/weaknesses in comparison to other cantons. Nevertheless, this does not tell if the absolute result of one subscale is better/worse than the absolute result of another subscale at the cantonal level.

The confidence intervals regarding the results of the subscales were not illustrated in the report due to two reasons. First, the absolute values at the national level were not measured in an exact manner. Second, the number of items of each subscale per booklet – which have the item difficulty lower than the cut-off value – was small. This number ranged between 0 and 8 depending on booklet and subscale. Therefore, both, the absolute values and the corresponding confidence intervals regarding the results of the subscales did not possess the same accurateness as other reported results.

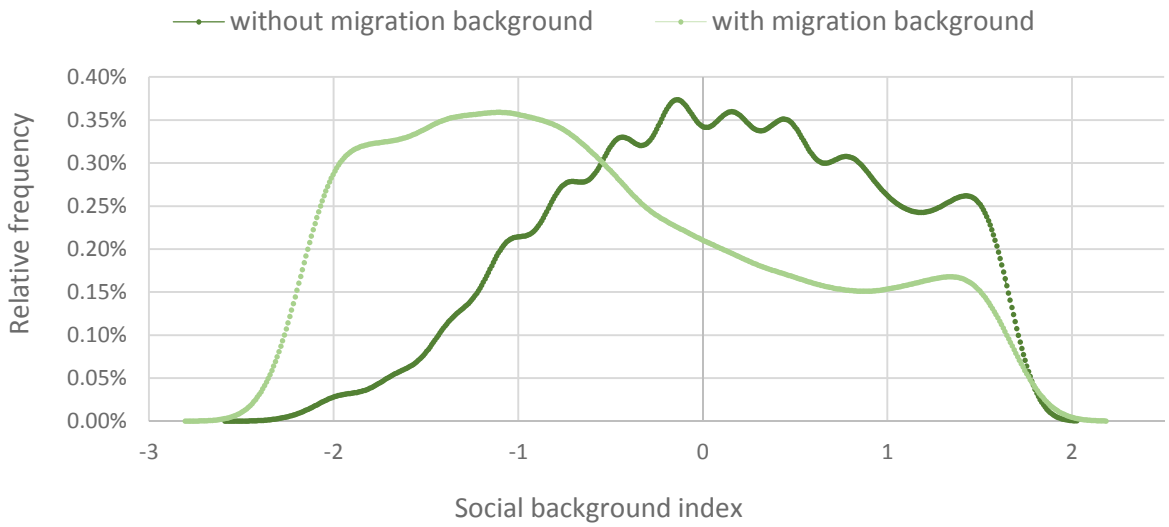
In summary, the results of the subscales should be interpreted with caution.

4 Special analyses

4.1 Differences between students with and without an immigrant background after controlling for social background

The achievement differences between students with and without immigrant background after controlling for the effect of social background were reported in chapter 5.2.2 of the national report. For this purpose, the potential outcome approach (POA) was applied. This is one of the most established approaches to study causal relationship between variables (Gangl, 2010; Lüdtke et al., 2010; Imbens & Wooldridge, 2009; Morgan & Winship, 2007; Winship & Morgan, 1999). It considers and explicitly deals with the different distributions of the index of social background (see *Figure 1*) between two student groups and does not assume the same effect of interest over all groups of comparison (see *Table 1*). This approach has been introduced to the educational research field (Lüdtke et al., 2010) and was applied in the educational standard survey (BIST-Ü) in Austria (Freunberger et al., 2014; Pham et al., 2014).

Figure 1: Distribution of social background index of students with and without migration status



While a large proportion of students with an immigrant background has an index of social background lower than 0, more than 50% students without migration status has an index of social background higher than 0. Due to this difference, it was suspected that the effect of social background on the attainment of minimum standards in mathematics might vary between the two groups of students. In fact, the results of two logistic regressions with social background index as predictor and attainment of minimum standards (0 = not attained, 1 = attained) as dependent variable confirmed this assumption. The social background effect differed significantly between the two groups as shown in *Table 1*:

Table 1: Effect of social background on the attainment of minimum standards

	Students without immigrant background	Students with an immigrant background
Intercept $\boldsymbol{\beta_0}$	0.71 (*SE* = .03)	0.19 (*SE* = .04)
Regression coefficient $\boldsymbol{\beta_1}$	0.80 (*SE* = .03)	0.66 (*SE* = .04)

Notes: results in log odds. SE = standard error

Using the terminology of experimental studies, this means that students were not randomly assigned to these two groups considering their social background. Thus, the mean difference in student outcomes (attainment of minimum standards) without adjustment might be biased and does not match the true difference with exclusive reference to the different migration statuses.

The reported difference between the two groups of students (with and without immigrant background) after controlling for the effect of social background was the *Average Treatment Effect* (ATE) as called in the POA. It can be interpreted as the mean difference in the outcome variable between two groups of students, if they had the same social background. For all students of each group, a *potential outcome* was calculated under the assumption that they belonged to the other group. Thus, for every student, a real outcome and a potential outcome were available. The ATE reflects the mean difference in student outcomes between students without and with an immigrant background considering both the real and the potential outcomes:

$$ATE = E[\delta] = E[Y \mid SES = s, M = 0] - E[Y \mid SES = s, M = 1],$$

δ is the individual difference in outcomes (Y) of each student (with $SES = s$) between two statuses: having no immigrant background ($M = 0$) and having an immigrant background ($M = 1$); $E[]$ denotes the average or mean of the value in brackets.

The (potential or real) outcome of student i without an immigration background $M = 0$ is denoted by y_{i0} and the outcome of students with an immigration background $M = 1$ is denoted by y_{i1}. The individual difference in outcomes between two statuses is:

$$\delta_i = y_{i0} - y_{i1}.$$

The potential outcomes of every student *with* an immigration background were estimated using their own social background index and the group-specific SES effect of students *without* immigrant background (Table 1, column 2). In this case, y_{i1} represents the real outcome while y_{i0} stands for the potential outcome.

The potential outcome of every student *without* an immigration background was estimated using their own social background index and the group-specific SES effect of students *with* immigrant background (Table 1, column 3). In this case, y_{i1} represents the potential outcome while y_{i0} stands for the real outcome.

As described above, the ATE was calculated as the mean value of δ over all students at the level of interest (national or cantonal level).

4.2 Technical notes on multilevel regression analysis

4.2.1 Method

The multilevel model was developed in the mid-80s to study the influence of context on individuals (Aitkin & Longford, 1986; Goldstein, 1986; Mason, Wong & Entwisle, 1983; Raudenbush & Bryk, 1986). In this model, the context is conceptualized as a hierarchical configuration composed of different levels nested within each other (micro-units should only belong to one higher level unit). In this case, considering that a student's performance (micro-unit) in the UGK 2016 tests depends on their own characteristics but also on their environment characteristics (in particular the canton where they are attending school (macro-unit)) leads to integrate the hierarchical structure of the

data into the analytical process. The multilevel analysis described in chapter 5.2.3 of the report was performed using a multilevel logistic regression[1]. A two-level regression analysis was carried out, with students serving as level 1 and cantons as level 2.The model coefficients and statistics were estimated using a restricted maximum likelihood procedure[2]. Non-response adjusted student base weights were used at level 1. Twenty binary plausible values (PVs) for the students' attainment of minimum standards served as the outcome variable. Results of the final model are the average of the twenty estimates obtained with each of the PVs.

4.2.2 Data and recoding variables

The data file used for the multilevel analysis included 22'423 students from 29 cantons (half-cantons and parts of cantons (in the case of multilingual cantons) are here also referred to as "cantons").

The explanatory variables used in the analysis are briefly described in Table 2 and correspond to the means of the 100 multiply imputed datasets**.** For further information about these variables see chapter 1.

Table 2: Explanatory variables used in multilevel analysis

Variable Name	Variable type	Categories	Categories labels/description
Gender *(Gender)*	Categorical	0 1	Male Female
Social background *(SES)*	Continuous		Z-standardized indice
Migration status *(Immig)*	Categorical	1 2 3	Native Migrant 2nd generation Migrant 1st generation
Language spoken at home *(Tlh3)*	Categorical	1 2 3	Only school language is spoken at home Different languages are spoken at home among these the school language The school language is not spoken at home

Categorical variables were recoded into a set of dummy variables. The number of dummy variables created from a categorical variable is smaller than the number of categories of the variable since one category is always used as a reference group. For each category, a dummy variable was created with the value of 1 if the student belongs to the respective category and 0 otherwise.

4.2.3 Modelling student performance

This section outlines the modelling strategy used in the multilevel analysis. For building the multilevel model, a step-by-step approach was adopted, starting from the student level upwards to the cantonal level. Readers interested in learning more about multilevel logistic regression can refer to Snijders & Bosker (1999), Bressoux (2010), Heck, Thomas & Tabata (2012) or Sommet & Morselli (2017).

[1] The commercial software HLM 7 (developed by Raudenbush, Bryk, Cheong, Congdon & du Toit) was used.

[2] As mentioned in Bressoux (2010), there are two maximum likelihood estimation methods (one full and one restricted) and there is no total agreement among statisticians that one method is superior to the other. Snijders and Bosker (1999) point out that the full maximum likelihood can lead to significant biases when the number of groups is small which is the case with the UGK 2016 dataset (the practical rule they deliver is that "small" means lower than 30).

Step 1. «Empty» model and calculation of the intraclass correlation coefficient (ICC)

The first information we look for when analysing hierarchical data is to estimate how the variance of the phenomenon is distributed over the different levels that are supposed to structure the data. To do this, a so-called «empty» model which does not include any explanatory variables is constructed. Its specification is defined as follows for a multilevel logistic regression:

Level 1 (students)

$$\log\left(\frac{\mathrm{P_{ij}}}{1-\mathrm{P_{ij}}}\right) = \beta_{0\mathrm{j}} \quad \textbf{(1)}$$

Level 2 (cantons)

$$\beta_{0\mathrm{j}} = \gamma_{00} + \mathrm{u_{0j}} \quad \text{with} \quad u_{0j} \sim N\left(0, \sigma^2_{u_0}\right) \quad \textbf{(2)}$$

Where *i* = 1,..., 22423 students, *j* = 1,…, 29 cantons and P_{ij} is the probability to achieve the core competencies in mathematics for student *i* within canton *j*.

Substituting **(2)** into **(1)** leads to:

$$\log\left(\frac{\mathrm{P_{ij}}}{1-\mathrm{P_{ij}}}\right) = \gamma_{00} + \mathrm{u_{0j}} \quad \textbf{(3)}$$

$$\leftrightarrow$$

$$P_{ij} = \frac{1}{1+\exp(-(\gamma_{00}+u_{0j}))} \quad \textbf{(4)}$$

This model is a decomposition of the variance of the dependent variable into the different levels. It provides the basic partition of the variability in the data between the different levels. In other words, the total variance of the UGK 2016 results can be decomposed as the sum of the 2 different levels variances:

- a within-canton variance (level 1): the variance within the cantons about their true means,
- a between-canton variance (level 2): the variance between the cantons' true means.

The main goal of this step is to identify differences in results between the cantons that are not due to randomness. If such differences did not exist, it would be pointless to develop more complex multilevel models aimed precisely at identifying and explaining these differences. The «empty» model allows to assess the statistical significance and the size of the between-canton variance i.e. the existence and the size of the cantonal effect on the probability to achieve the core competencies. As mentioned in Heck, Thomas & Tabata (2012), «little variability between the Level-2 units would suggest little need to conduct a multilevel analysis» (p.19).

Fitting the empty model yields the parameter estimates presented in Table 3.

Table 3: Estimates for empty model

Fixed effect	Coefficient	S.E.	p-value
γ_{00} = Intercept	0.5149	0.0755	<0.001
Random effect			
Level-two variance:			
$\sigma^2_{u_0}$ =	0.1464		<0.001

The between-canton variance ($\sigma^2_{u_0}$) is equal to 0.1464 and is statistically significant. It is therefore possible to calculate the intraclass correlation coefficient (ICC) which is the fraction of total variability that is due to the cantonal level. In a multilevel logistic regression the ICC is defined as follows:

$$\rho = \frac{\sigma^2_{u0}}{\sigma^2_{u0}+{\pi^2}/{3}} \leftrightarrow \rho = \frac{\sigma^2_{u0}}{\sigma^2_{u0}+3.29} \leftrightarrow \rho = \frac{0.1464}{0.1464+3.29} = 0.0426 \quad \textbf{(5)}$$

The fraction of total variability that is due to the cantonal level is around 4% in mathematics on the national level. From this empty model it is also possible to estimate the probability to achieve the core competencies in a «typical» canton (which is a canton where $u_{0j} = 0$) using **(4)**.

$$P_{ij} = \frac{1}{1+\exp(-(\gamma_{00}+u_{0j}))} = \frac{1}{1+\exp(-(0.5149+0))} = 0.625 \quad \textbf{(6)}$$

This probability is equal to 62.5%, extremely close to what we can estimate for the national level (62.2%). The small difference between these two values is due to the non-linear relationship between the logit (whose distribution is symmetrical) and the probability (whose distribution is asymmetrical). For more details on this point, see Bressoux (2010) or Raudenbush & Bryk (2002).

Step 2. Modelling within-group variability: construction of a model for level one

In this second step of the multilevel model building process we select relevant available level-one variables (i.e. students' characteristics) to explain differences in the achievement of core competencies. As mentioned in chapter 5 of the national report, social background, migration status, language spoken at home and gender are selected because of their strong correlation with student performance and because they are not influenced by the educational offer. Here it is necessary to specify to what extent the effects of each of these variables should be modelled as fixed or random effects. Indeed, it is possible that the effect of some explanatory variables differ from canton to canton, that is, some variables have random slopes. Therefore, the significance of the random slope variance was tested for the whole set of explanatory variables. Social background was the only one to have such a significant variance. The model specification is defined as follows:

Level 1 (students)

$$log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \beta_{0j} + \beta_1 Gender_{ij} + \beta_{2j} SES_{ij} + \beta_3 Immig2_{ij} + \beta_4 Immig3_{ij} + \beta_5 Tlh3\_2_{ij} + \beta_6 Tlh3\_3_{ij} \quad \textbf{(7)}$$

Level 2 (cantons)

$\beta_{0j} = \gamma_{00} + u_{oj}$ with $u_{0j} \sim N(0, \sigma^2_{u0})$ **(8)**

$\beta_1 = \gamma_{01}$ **(9)**

$\beta_{2j} = \gamma_{02} + u_{2j}$ with $u_{2j} \sim N(0, \sigma^2_{u2})$ **(10)**

$\beta_3 = \gamma_{03}$ **(11)**

$\beta_4 = \gamma_{04}$ **(12)**

$\beta_5 = \gamma_{05}$ **(13)**

$\beta_6 = \gamma_{06}$ **(14)**

Substituting **(8)** to **(14)** into **(7)** leads to:

$$log\left(\frac{P_{ij}}{1-P_{ij}}\right) = \gamma_{00} + \gamma_{01}Gender_{ij} + \gamma_{02}SES_{ij} + \gamma_{03}Immig2_{ij} + \gamma_{04}Immig3_{ij} + \gamma_{05}Tlh3\_2_{ij} + \gamma_{06}Tlh3\_3_{ij} + \left(u_{oj} + u_{2j}SES_{ij}\right) \quad \textbf{(15)}$$

Expressing **(15)** as a probability leads to:

$$P_{ij} = \frac{1}{1 + exp\left[-\left(\gamma_{00} + \gamma_{01}Gender_{ij} + \gamma_{02}SES_{ij} + \gamma_{03}Immig2_{ij} + \gamma_{04}Immig3_{ij} + \gamma_{05}Tlh3\_2_{ij} + \gamma_{06}Tlh3\_3_{ij} + \left(u_{oj} + u_{2j}SES_{ij}\right)\right)\right]} \quad \textbf{(16)}$$

Fitting the model yields the parameter estimates presented in *Table 4*.

Table 4: Estimates for random slope logistic model

Fixed effects	Coefficient	S.E.	Sign.
Intercept	0.8982	0.0732	***
Gender *(ref. = male)*	-0.1516	0.0265	***
Social background (SES)[(e)]	0.7479	0.0275	***
Migrant 2nd generation *(ref. = native)*	-0.1671	0.0620	*
Migrant 1st generation *(ref. = native)*	-0.2759	0.0542	**
Different languages are spoken at home, among these the school language *(ref. only language spoken at home = school language)*	-0.3355	0.0673	***
The school language is not spoken at home *(ref. only language spoken at home = school language)*	-0.5626	0.0813	***
Random effects			
Level-two variances:			
σ^2_{u0} = Intercept variance	0.1838		***
σ^2_{u2} = Social background Slope variance	0.0091		*

[(e)] Random slope
*** Significant of p<0.001; ** Significant of p<0.01; * Significant of p<0.05; n.s = not significant.

In theory, the third and last step of multilevel analysis should be to introduce some relevant level-two variables (i.e. cantons' characteristics) into the model in order to explain the intercept and slope variances that appear in equations (8) and (10). Unfortunately, it has not been possible to identify any significant cantonal explanatory variable in the UGK 2016 dataset. Therefore, the final model is the one presented in Table 4.

Step 3. Probability to achieve the core competencies for a specific student's profile

The level-two variances are still statistically significant after controlling for individual characteristics (Table 3). Differences in results between cantons remain even when we control some usual sociodemographic characteristics (social background, migration status, language spoken at home and gender). In other words, this means that a student with the same sociodemographic characteristics

does not have the same chances of achieving the core competencies depending on the canton in which they are attending school. One way to illustrate that result is to calculate the estimated probabilities of achieving the core competencies (and their confidence interval) in all the cantons for some given students' profiles. To estimate one probability, simply replace the coefficients associated with each of the explanatory variables with their estimates in equation **(16)** and the explanatory variables with their value for the selected student profile. The detailed equation to be used is presented in **(17)**:

$$P_{ij} = \frac{1}{1 + exp\left[-\left(0.898 - 0.151Gender_{ij} + 0.747SES_{ij} - 0.167Immig2_{ij} - 0.275Immig3_{ij} - 0.335Tlh3\_2_{ij} - 0.562Tlh3\_3_{ij} + \left(u_{oj} + u_{2j}SES_{ij}\right)\right)\right]} \quad \textbf{(17)}$$

To complete the calculation, it is also necessary to estimate the u_{oj} and u_{2j} parameters. These random group effects, also known as *posterior means*, are estimated using the empirical Bayes estimation method. An estimate of their values is available in HLM level-two residual files (Raudenbush et al., 2011). The average of the 20 estimates of these parameters for each canton is presented in Table 5.

Table 5: Estimates for posterior means

		Posterior means	
		$\hat{u}_{0j}$	$\hat{u}_{2j}$
AG	Argovie	-0.2666	0.0442
BE_d	Berne (germanophone)	-0.6642	0.1432
LU	Lucerne	-0.2844	0.0319
SG	Saint-Gall	0.3649	-0.0392
ZH	Zurich	-0.3164	0.1010
VD	Vaud	0.4628	-0.1138
BL	Bâle Campagne	-0.6035	0.1176
BS	Bâle Ville	-0.6897	0.1196
FR_d	Fribourg (germanophone)	-0.0238	0.0157
GR	Grisons	0.1611	-0.0239
SO	Soleure	-0.3469	0.0723
TG	Thurgovie	0.0637	-0.0364
SZ	Schwytz	0.2896	-0.0248
FR_f	Fribourg (francophone)	1.1727	-0.2227
GE	Genève	-0.0095	0.0140
NE	Neuchâtel	-0.2979	0.0220
VS_f	Valais (francophone)	1.0452	-0.2121
TI	Tessin	-0.0825	-0.0112
AI	Appenzell Rhodes Intérieures	0.4150	-0.0739
AR	Appenzell Rhodes Extérieures	-0.1531	0.0214
GL	Glaris	0.1831	-0.0495
NW	Nidwald	-0.1272	0.0163
OW	Obwald	-0.1468	0.0255
SH	Schaffhouse	0.1563	-0.0164
UR	Uri	-0.0470	-0.0025
VS_d	Valais (germanophone)	0.2437	-0.0317
ZG	Zoug	-0.0047	0.0377
BE_f	Berne (francophone)	-0.0655	0.0071
JU	Jura	0.0825	-0.0436

By analogy to what is described in Hosmer & Lemeshow (2000) we provided 95% interval estimates for the fitted values (i.e. the predicted probabilities). Basically, the calculation of this confidence interval (CI) associated with the probability of achieving the core competencies for a given student's profile can be summarized in the 5 steps listed below:

- Calculation of the student's logit
- Estimation of the variance and standard error of the student's logit
- Calculation of the logit CI (lower and upper limits)
- Estimation of the probability to achieve the core competencies
- Calculation of the probability CI (lower and upper limits)

The general expression for the student's logit is given in **(15)** and the estimator of this logit, as described in **(18)**, simply corresponds to the same equation in which the coefficients have been replaced by their estimated value as shown in Table 4 (thus, the value of the logit depends on the student's characteristics and on the canton in which they are attending school).

$$log\left(\frac{P_{ij}}{1-P_{ij}}\right) = 0.898 - 0.151Gender_{ij} + 0.747SES_{ij} - 0.167Immig2_{ij} - 0.275Immig3_{ij} - 0.335Tlh3\_2_{ij} - 0.562Tlh3\_3_{ij} + \left(\hat{u}_{oj} + \hat{u}_{2j}SES_{ij}\right) \quad \textbf{(18)}$$

$$\leftrightarrow$$

$$log\left(\frac{P_{ij}}{1-P_{ij}}\right) = (0.898 + \hat{u}_{oj}) - 0.151Gender_{ij} + (0.747 + \hat{u}_{2j}) \times SES_{ij} - 0.167Immig2_{ij} - 0.275Immig3_{ij} - 0.335Tlh3\_2_{ij} - 0.562Tlh3\_3_{ij} \quad \textbf{(19)}$$

An alternative way to express the estimator of the logit in **(18)** is to use of vector notation as $\hat{g}(X) = X'\hat{\beta}$

where

- the vector $X' = \left(1, Gender_{ij}, SES_{ij}, Immig2_{ij}, Immig3_{ij}, Tlh3\_2_{ij}, Tlh3\_3_{ij}\right)$ represent the intercept and a set of values of the 6 explanatory variables and
- the vector $\hat{\beta}' = \left(\hat{\beta}_{0j}, \hat{\beta}_{01}, \hat{\beta}_{2j}, \hat{\beta}_{03}, \hat{\beta}_{04}, \hat{\beta}_{05}, \hat{\beta}_{06}\right)$ denotes the estimator of intercepts and slopes. For the 5 variables that were modelled as fixed effects, terms $\hat{\beta}_{01}, \hat{\beta}_{03}, \hat{\beta}_{04}, \hat{\beta}_{05}$ and $\hat{\beta}_{06}$ are the 5 coefficients estimates ($\hat{\gamma}_{01}, \hat{\gamma}_{03}, \hat{\gamma}_{04}, \hat{\gamma}_{05}$ and $\hat{\gamma}_{06}$). Terms $\hat{\beta}_{0j} = \hat{\gamma}_{00+\hat{u}_{0j}}$ and $\hat{\beta}_{2j} = \hat{\gamma}_{02+\hat{u}_{2j}}$ correspond to the estimates of the model's random intercepts **(8)** and slopes **(10)** that vary from one canton to another.

To estimate the variance and the standard error of the logit above, the estimated covariance matrix of the estimated coefficients is also needed. The latter (noted as $\hat{V}$) is provided by HLM software and the average of the 20 estimates is presented in *Table 6*.

Table 6: Estimated Covariance Matrix of the Estimated Coefficients in Table 3

$\hat{V} =$

	Intercept	Gender	SES	Immig2	Immig3	Tlh3_2	Tlh3_3
Intercept	0.005378	-0.000559	-0.001348	-0.001253	0.001118	0.002023	0.002682
Gender	-0.000559	0.000728	0.000185	0.000415	-0.000257	-0.000663	-0.000682
SES	-0.001348	0.000185	0.000764	0.000490	-0.000323	-0.000813	-0.000823
Immig2	-0.001253	0.000415	0.000490	0.003895	0.000645	-0.002687	-0.003196
Immig3	0.001118	-0.000257	-0.000323	0.000645	0.003006	0.000664	-0.000224
Tlh3_2	0.002023	-0.000663	-0.000813	-0.002687	0.000664	0.004579	0.004302
Tlh3_3	0.002682	-0.000682	-0.000823	-0.003196	-0.000224	0.004302	0.006652

Once again to use of vector notation is the easiest and most concise way to express the estimator of the variance of $\hat{g}(X)$. The expression for the estimator of the variance is

$$\hat{Var}[(\hat{g}(X))] = X'\hat{V}X \quad \textbf{(20)}$$

and the estimator of the standard error is

$$\widehat{SE} = \sqrt{\hat{Var}[(\hat{g}(X))]} = \sqrt{X'\hat{V}X} \quad \textbf{(21)}$$

From the equations above, the 95% student's logit confidence interval can be derived

$$Student's\ logit\ CI = \hat{g}(X) \pm 1.96 \times \widehat{SE} \quad \textbf{(22)}$$

and the estimator of the probability to achieve the core competencies for a given students' profile in a given canton and its 95% CI

$$Student's\ probability = P_{ij} = \frac{1}{1+e^{-\hat{g}(X)}} \quad \textbf{(23)}$$

$$Student's\ probability\ CI = \left[\frac{1}{1+e^{-[\hat{g}(X)-1.96\times\widehat{SE}]}}; \frac{1}{1+e^{-[\hat{g}(X)+1.96\times\widehat{SE}]}}\right] \quad \textbf{(24)}$$

4.3 Approaches for the adjustment of cantonal estimates

In *approach 1*- separate logistic regression analyses on the basis of multiply imputed and weighted data per canton were estimated (see Long, 1997) using the R-package BIFIEsurvey (BIFIE, 2018). The regression coefficients mirror the cantonal associations between student background covariates and the probability to achieve the basic competences. The covariates included in the model are: gender, the language spoken at home, the immigrant status and the social background (SES). Based on these canton-specific regression coefficients and the matrix of the student population that corresponds to the Swiss population (on the included covariates) we estimated the hypothetical (potential) basic competence shares achieved by canton. These hypothetical shares show what shares of students within cantons potentially achieved the basic competences if the cantonal student distribution on the select covariates corresponded to the Swiss national distribution while the associations between background characteristics and achievement remained as they were within cantons (counterfactual approach). The main findings remained the same, when robustness checks were conducted by including different models and specifying interaction terms between covariates. The main disadvantage of approach 1 is, that it bases on a strong and potentially untenable model assumption. Namely, it is assumed that the cantonal associations between student background characteristics and achievement remained the same even if the composition was different. Hence, in essence, the absence of compositional effects was assumed.

In *approach 2*- logistic regression analyses on the basis of multiply imputed and weighted Swiss national data were conducted (see Long, 1997) using the R-package BIFIEsurvey (BIFIE, 2018). In parallel to student-level covariates, aggregate covariates at the cantonal level were included in order to account for the varying cantonal compositions of students (due to differences in school systems between cantons, aggregate variables on school level were not included). The covariates included in

the model were: gender, the language spoken at home, the migrant status and the socio-economic status (SES). Due to a curvilinear relationship with the outcome, the aggregate SES was also included as quadratic term. Moreover, interactions between the SES and the language spoken at home were included. Again, different models for robustness checks were specified. The regression coefficients mirror the Swiss national associations between student background covariates, cantonal student compositions and the probability to achieve the basic competences. On the basis of these Swiss national associations one can calculate the expected probability to achieve the basic competences for all combinations of background characteristics. As an example, the expected (Swiss national) probability of achieving the basic competences for a male student with second generation migrant status who does not speak the test language at home and who attends school in a (cantonal) setting of above average shares of migrants and below average SES can be calculated. These expected probabilities by covariate combination can then be used in a next step to compute the adjusted shares of students achieving the basic competences for the student characteristic distributions in each canton. These adjusted shares represent the expected competences for each canton, when the different student population compositions are taken into account. An advantage of approach 2 is that it explicitly takes into account student composition effects. A disadvantage is that the expectations are modelled based on a comparison of similarities across cantons and it could be that some combinations are rare (at the cantonal level). This would then result in the computation of expectations, which are close to the (unadjusted) observed achievement levels for the cantons affected (on the problem of overfitting, see e.g., Pham, Robitzsch, George & Freunberger, 2016, p. 317).

5 References

Aitkin, M. & Longford, N.T. (1986). Statistical modelling in school effectiveness studies. *Journal of the Royal Statistical Society, Series A, 149*, 1–43.

Angelone, D. & Keller, F. (2019). *ÜGK 2016 Mathematik. Technische Dokumentation zu Testentwicklung und Skalierung*. Aarau: Geschäftsstelle der Aufgabendatenbank EDK (ADB).

BIFIE (2018). BIFIEsurvey: Tools for survey statistics in educational assessment. R package version 2.5-44. https://CRAN.R-project.org/package=BIFIEsurvey. Retrieved: 28.12.2018.

Bressoux, P. (2010). *Modélisation statistique appliquée aux sciences sociales.* Bruxelles, Editions De Boeck Université.

Bruneforth, M., Oberwimmer, K., & Robitzsch, A. (2016). Reporting und Analysen. In S. Breit & C. Schreiner (Eds.). *Large-Scale Assessment mit R: Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (pp. 333–362). Wien: facultas.

Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Hillsdale, NJ: Erlbaum.

Enders, C. K. (2010). *Applied Missing Data Analysis*. New York: The Guilford Press.

Foy, P. (2012). Estimating standard errors for the TIMSS and PIRLS 2011 achievement scales. In M. O. Martin & I. V. S. Mullis (Hrsg.), *Methods and procedures in TIMSS and PIRLS 2011*. Chestnut Hill, MA: TIMSS/PIRLS International Study Center, Boston College.

Freunberger, R., Robitzsch, A. & Pham, G. (2014). Hintergrundvariablen und spezielle Analysen. Technische Dokumentation – BIST-Ü Mathematik, 4. Schulstufe, 2013. Salzburg: BIFIE. https://www.bifie.at/node/2765

Gangl, M. (2010). Causal inference in sociological research. *Annual Review of Sociology, 36 (1)*, 21–47.

Ganzeboom, H. B. (2010a). *International standard classification of occupations ISCO-08 with ISEI-08 scores*. http://www.harryganzeboom.nl/isco08/isco08_with_isei.pdf. Retrieved Jul. 12, 2018.

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. Annual Conference of International Social Survey, Lisbon.

Ganzeboom, H. B., & Treiman, D. J. (2008). *International Stratification and Mobility File: Conversion Tools*. http://www.harryganzeboom.nl/ismf/index.htm. Retrieved Jul. 12, 2018

Ganzeboom, H. B., De Graaf, P., & Treiman, D. (1992). A standard international socio-economic. *Social Science Research, 2(1)*, pp. 1-56.

Goldstein, H. (1986). Multilevel mixed linear model analysis using iterative generalized least squares. *Biometrika, 73*, 43–56.

Hattie, J. (2009). Visible learning: A synthesis of over 800 meta-analyses relating to achievement. London: Routledge.

Heck, R.H., Thomas, S.L & Tabata, L.N. (2012). *Multilevel modeling of categorical outcomes using IBM SPSS*. New York: Routledge.

Hosmer, D.W. and Lemeshow, S. (2000). *Applied logistic regression* (2nd ed.). New York: John Wiley & Sons, Inc.

Imbens, G. W. & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature, 47 (1)*, 5–86.

Judkins, D.R. (1990), Fay's Method for Variance Estimation, *Journal of Statistics, Vol. 6*, 223–229.

Long, S. J. (1997). Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: Sage Publications.

Lüdtke, O., Robitzsch, A., Köller, O. & Winkelmann, H. (2010). Kausale Effekte in der Empirischen Bildungsforschung. Ein Vergleich verschiedener Ansätze zur Schätzung des Effekts des Einschulungsalters. In W. Bos, E. Klieme & O. Köller (Hrsg.), *Schulische Lernangelegenheiten und Kompetenzentwicklung. Festschrift für Jürgen Baumert* (pp. 257–284). Münster: Waxmann.

Mason, W.M., Wong, G.M., & Entwisle, B. (1983). Contextual analysis through the multilevel linear model. In S. Leinhardt (Ed.), *Sociological methodology* (pp.72-103). San Francisco: Jossey-Bass.

Mislevy, R. J., Beaton, A. E., Kaplan, B. & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. *Journal of Educational Measurement, 29*, 133–161.

Morgan, S. L. & Winship, C. (2007). *Counterfactuals and causal inference*. Cambridge: Cambridge University Press.

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

OECD (2016). *PISA 2015 Results (Volume I): Excellence and Equity in Education*. Paris: OECD Publishing. dx.doi.org/10.1787/9789264266490-en.

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

Pham, G., Freunberger, R. & Robitzsch, A. (2014). Hintergrundvariablen und spezielle Analysen. Technische Dokumentation – BIST-Ü Englisch, 8. Schulstufe, 2013. Salzburg: BIFIE. https://www.bifie.at/node/2849

Pham, G., Robitzsch, A., George, A. C., Freunberger, R. (2016). Fairer Vergleich in der Rückmeldung. In S. Breit, & C. Schreiner (Hrsg.), *Large-Scale Assessment mit R. Methodische Grundlagen der österreichischen Bildungsstandardüberprüfung* (pp. 295-332). Wien: Facultas.

Raudenbush, S. W., Bryk, A. S., Cheong, A. S., Fai, Y. F., Congdon, R. T., & du Toit, M. (2011). *HLM 7: Hierarchical linear and nonlinear modeling*. Lincolnwood, IL: Scientific Software International.

Raudenbush, S.W., & Bryk, A.S. (1986). A hierarchical model for studying school effects. *Sociology of Education, 59*, 1–17.

Raudenbush, S.W., & Bryk, A.S. (2002). Hierarchical linear models. *Applications and data analysis methods* (2e éd.). Thousand Oaks, London, New Dehli: Sage.

Robitzsch, A., Pham, G. & Yanagida, T. (2016). Fehlende Daten und Plausible Values. In Breit, S. & Schreiner, C. (Eds.), *Large-Scale Assessment mit R: Methodische Grundlagen der österreichischen Bildungsstandard-Überprüfung*. Wien: facultas, S. 259–294.

Robitzsch, A., Grund, S. & Henke, T. (2018). *Miceadds: Some additional multiple imputation functions, especially for 'mice'. R package version 2.15-22*. https://cran.r-project.org/web/packages/miceadds/miceadds.pdf.

Rubin, DB. (1987). Multiple Imputation for Nonresponse in Surveys. New York: Wiley.

Rubin, DB. (2003). Nested multiple imputation of NMES via partially incompatible MCMC. *Statistica Neerlandica, 57(1)*, 3–18.

Sacchi, S. & Oesch, D. (2017). ÜGK 2016: Documentation of questionnaire-based scales. Bern: TREE.

Schulz, W. and Friedman, T. (2015). Chapter 12: Scaling procedures for ICILS questionnaire items. In J. Fraillon, W. Schulz, T. Friedman, J. Ainley & E. Gebhardt (Eds.), *ICILS 2013 technical report* (pp. 177-220). Amsterdam: IEA.

Shen, Z. J. (2000). Nested multiple imputation. *Ph.D. Thesis*, Department of Statistics, Harvard University, Cambridge, MA.

Sommet, N. & Morselli, D. (2017). Keep Calm and Learn Multilevel Logistic Modeling: A Simplified Three-Step Procedure Using Stata, R, Mplus, and SPSS. *International Review of Social Psychology, 30(1)*, 203–218. doi. org/10.5334/irsp.90

Snijders, T. & Bosker, R. (1999). Multilevel Analysis: an introduction to basic and advanced multilevel modeling. London: Sage.

van Buuren, S. (2012). *Flexible imputation of missing data*. Boca Raton, FL: CRC press.

Verner, M., & Helbling, L. (2019). Sampling ÜGK 2016. Technischer Bericht zu Stichprobendesign, Gewichtung und Varianzschätzung bei der Überprüfung des Erreichens der Grundkompetenzen 2016. Zürich: Institut für Bildungsevaluation, assoziiertes Institut der Universität Zürich.

Winship, C. & Morgan, S. L. (1999). The estimation of causal effects from observational data. *Annual Review of Sociology, 25 (1)*, 659–706.