ÜGK – COFO – VECOF 2017 results: Technical appendices

Giang Pham, Laura Helbling,
Martin Verner & Alice Ambrosetti

Impressum

Autors	Giang Pham (PHSG), Laura Helbling, Martin Verner, (IBE), Alice Ambrosetti (CIRSE)
Quotation proposal	Pham, G., Helbling, L., Verner, M. & Ambrosetti, A. (2019). ÜGK – COFO – VeCoF 2017 results: Technical appendices. St.Gallen & Genève: Pädagogische Hochschule St.Gallen (PHSG) & Service de la recherche en éducation (SRED).
Download	www.cofo-suisse.ch/cofo-2017
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 question A04 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 `MISEI` (occupational status of mother – ISEI-08 status) and `FISEI` (occupational status of father – ISEI-08 status).

In the raw dataset (with missing values), the highest occupational status of parents (`HISEI`) corresponds to the higher value between `MISEI` and `FISEI`, in case both values were answered. If at least one value is missing, `HISEI` has a missing value.

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

1.1.2 Highest parental education level

Parental education was assessed by means of question A08 in the student questionnaire. Based on the following options, students reported on the highest education attainment of their mother and father:

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

In the cleaning process, category 7 (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.

Two new variables MEDU (mother's highest educational attainment) and FEDU (father's highest educational attainment) were created by reducing the original data into the following categories:

- 0 = compulsory schooling level or lower
- 1 = upper secondary education
- 2 = tertiary education

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 2016 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 (with missing values), the highest parental educational level (HISCED) corresponds to the higher value between MEDU and FEDU. If at least one of these two values was missing, HISCED has a missing value.

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 HISCED 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 A14 in the student questionnaire. Students reported the number of books at home (variable A14) by choosing one of the following answer options:

- 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 A14 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. In addition, corresponding data of the ÜGK 2016 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). Identical coding rules and calculation of the social background index in both studies were intended.
- Problematic psychometric parameters: The mean scores of several items were very high (relative score > 0.95), e.g. internet connection is available in almost every family. Several items correlated not at all or negatively with student achievement in reading in school language. 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, 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 (HISEI and HISCED) 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 HISEI and HISCED 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:

$$SES_2 = zSES_1,$$

$$SES_1 = \frac{zHISEI + zHISCED + znbooks}{3},$$

zHISEI, zHISCED and znbooks are *z*-scores of the three basic indices (HISEI, HISCED and nbooks). Weighted data (using student weights) were used to standardize variables.

In the raw dataset (with missing values), the social background (variable SES) was calculated based on the raw values of zHISEI, zHISCED and znbooks. The values of SES in this dataset correspond to the values of SES_2 as described above.

For the national report, 20 imputed datasets (see chapter 2) were applied. First, the SES_1 – the weighted mean of zHISEI, zHISCED 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.

1.2 Home language

Questions A12a to A13b in the student questionnaire asked students about their main and second languages spoken at home. Variable A12a contains student responses in regard to the main language spoken at home; variables A13a and A13b contain student responses in regard to the second language spoken at home, if available.

Based on these three variables, three new variables (homelang1, homelang2f and homelang2) were created:

- homelang1: the main language spoken at home is the school language (0 = false, 1 = true)
- homelang2f: another language is spoken at home (0 = false, 1 = true)
- homelang2: the second language spoken at home is the school language (0 = false, 1 = true)

The final variable regarding the language spoken at home or home language (homelang) is coded based on data of three variables homelang1, homelang2f and homelang2. This 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 and Standard German were 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.

The Romansh language was not treated as the school language in the Engadin, since there were no tests in this language.

For the national report, 20 imputed datasets were used. All missing values of the three basic variables `homelang1`, `homelang2f`, and `homelang2` (if exist) were multiply imputed (see chapter 2). Within each imputed dataset, the variable `homelang` was derived from these three basic variables. The reported results derived based on the pooled results over all imputed datasets (see chapter 3).

1.3 Immigration status

The immigration status in ÜGK 2017 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 A10 in the student questionnaire asked students about their country of birth as well as the country of birth of their mother and father.

Based on students' responses, three new variables were coded, which indicate whether the student (`A10cobsaggr`), the mother (`A10cobmaggr`), and the father (`A10cobfaggr`) were born abroad (value = 0) or in Switzerland (value = 1).

For the national report, all missing values of the three basic variables `A10cobsaggr`, `A10cobmaggr`, and `A10cobfaggr` 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 aforementioned 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 of the students were the pooled results over all imputed datasets (see chapter 3).

In the raw dataset, variable `immig_pisa` was derived identically in case there were no missing values of the three basic variables. In case either `A10cobmaggr` or `A10cobfaggr` has value 1 (one parent was born in Switzerland), `immig_pisa` has value 1 regardless if other two variables have missing values or not according to the definition. In case the birth data of the student and of one parent were available, the birth data of the other parent was missing, the available data were used to derive the value of `immig_pisa`. In all other cases, `immig_pisa` has a missing value.

2 Dealing with missing values of context variables

There was a total of 171 student questionnaire items with missing data. The share of missing data of each single item in the student questionnaire ranged between 3% and 26%. For 149 items, the share of missing values exceeded 5%, and by 99 items this proportion exceeded 10% (see Pham, 2019). The share of missing data for derived variables such as the social background `SES`, which was calculated based on the values of other items, was even higher, since they were coded as missing if any of the primary items had no valid response. Variable `SES` had the highest missing rates (40%), mostly due to the high proportion of missing data of variables `MISEI` and `FISEI` (see chapter 1). Generally, ignoring missing data (which is equivalent to the listwise- or pairwise-deletion method of dealing with missing data) would lead to three major problems (Little & Rubin, 2002; Enders, 2010; Lüdtke, Robitzsch, Trautwein, & Köller, 2007; Schafer & Graham, 2002):

- Reduced sample size for analyses in report and publication: The reduced sample size due to missing data does not match the sample procedure. Thus, the recalculation of sample weights for each analysis would be necessary.
- Difficulties in applying standard statistical methods and software which require complete data matrices.
- Risk of having biased estimates due to systematic differences between observed and missing data: students who did not achieve the GK in L1-reading had 25% missing data on average, while students who achieved GK in L1-reading only had 15% missing data of all questionnaire items on average.

In addition, the share of missing data varied between cantons and language regions. Ignoring missing data might lead to biased comparisons between cantons and regions.

Thus, dealing appropriately with missing data in the context of this study was inevitable. Between the two state-of-the-art methods to deal with missing data (Lüdtke & Robitzsch, 2010), the multiple imputation (MI) method (Rubin, 1987) was chosen over the model-based method (full information maximum-likelihood method) for the consistency and reproducibility of the results.

In this study, we adopted the multiple imputation procedures for questionnaire data in large-scale assessments as suggested and described by Robitzsch, Pham, & Yanagida (2016). All missing values were assumed *missing at random* (MAR) (Rubin, 1976). The missing data were imputed by chained equations (MICE approach) (van Buuren, 2012) under the MAR assumption using R package **mice** (van Buuren & Groothuis-Oudshoorn, 2011) with supplement functions from R package **miceadds** (Robitzsch, Grund, & Henke, 2018). The predictor matrix for the imputation model of each variable with missing data involved all available variables in the dataset including questionnaire items[1], the plausible values (PV) of all language tests[2] (Angelone & Keller, 2019), tracking data (Verner & Hebling, 2019), and cantonal-level data. Moreover, school-level aggregated values of all level-1 (individual level) variables were also included in the predictor matrix. Quadratic terms of interactions between important variables[3] and all other variables were included in the prediction matrix to consider

[1] Except B03b (school grades of class repetition), B04b (coachaim-Items) and B05 (school mark) due to problems and limited time in data cleaning process, and other items which had been eliminated based on the results of the pilot study as well as the preliminary check process before data imputation.

[2] In this study, the multiple imputation of all questionnaire data took place after the plausible values of test data had been calculated.

[3] Tracking variables and variables used for the national report including test data.

possible non-linear relationships. The multilevel data structure (students are nested within schools) was taken into account using the random intercept model for the imputation of level 1 (individual level) variables. Since the number of predictors including interaction terms turned out to be large, the partial least squares technique (Abdi, 2010) was applied. That means, a smaller number of uncorrelated factors was stepwise extracted under the criterion of retaining as much as possible of the variation presented in both the dependent variable and the predictor matrix. For this purpose, the R package **pls** (Mevik, Wehrens & Liland, 2016) was used.

The data imputation was conducted iteratively and by multiple times. Within each iteration, missing data of each variable were imputed separately by canton (or linguistic region of one canton) in order to allow for canton specific data structures. Imputed values of one iteration served as starting values for the next iteration. Imputed values after multiple iterations were saved and treated as one imputed dataset. For each imputed dataset, only one set of plausible values (e. g. the first plausible values of all test performances of students) was used as predictors. A total of 20 imputed datasets – corresponding to 20 sets of plausible values of test performances – were generated for reporting result and subsequent analyses. For a more detailed explanation and technical description of the data imputation process see Robitzsch et al. (2016).

3 Estimation of descriptive results and measurement errors

All results including confidence intervals in the national report ÜGK 2017 were estimated using standard combining rules based on 20 plausible values (Rubin's rule, Rubin, 1987). In addition, due to the complex sampling design (see Verner & Helbling, 2019), there were some disproportionalities in the sample data. All analyses, referring to population measures, were conducted using sampling 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).

3.1 Estimation of point estimates using multiply imputed datasets

All reported point estimates (e.g. 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 result 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*):

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

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

The estimation variance of a point estimate $\hat{\mu}$ 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 represented by 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 result of interest. Unless otherwise indicated, the lower (KI_{low}) and upper (KI_{upp}) bound of this interval were calculated as follows:

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

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, the sampling variance might be large. 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 if 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 has been 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 belong to the first group, $\hat{\mu}_2$ and SD_2 are the estimate and corresponding sample standard deviation belong to 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.

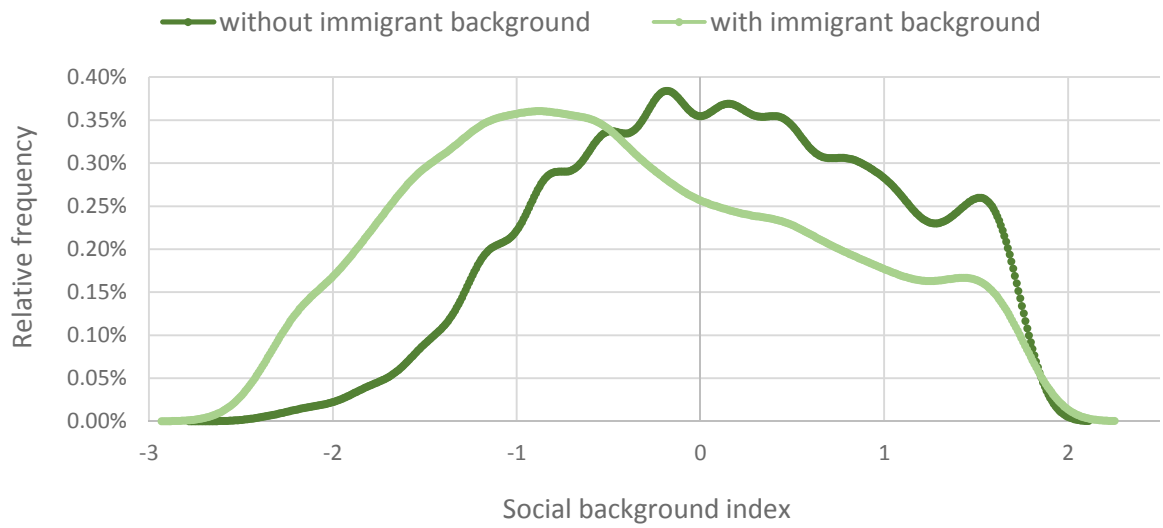
Cohen (1988) suggested that $d \geq 0.2$ can be interpreted as a small, $d \geq 0.5$ a medium, and $d \geq 0.8$ a large effect size. Hattie (2009, p. 9) suggested $d \geq 0.2$ for small, $d \geq 0.4$ for medium, and $d \geq 0.6$ for large effect size when judging educational outcomes. In this report,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

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 5.1.6 of the 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 the 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 two groups of students. In fact, the results of two logistic regressions with social background index as the predictor and attainment of minimum standards (0 = not attained, 1 = attained) as the dependent variable confirmed this assumption. The social background effect differed 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 $\boldsymbol{\beta_0}$	2.42 (*SE* = .05)	1.69 (*SE* = .06)
Regression coefficient $\boldsymbol{\beta_1}$	0.62 (*SE* = .06)	0.49 (*SE* = .06)

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 result 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 the two groups of students, if they had the same social background.

The general idea is as follows. For each student 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 was 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 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 minimal standards. 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 minimal standards 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 minimal standards. On the basis of these Swiss national associations one can calculate the expected probability to achieve the minimal standards for all combinations of background characteristics. As an example, the expected (Swiss national) probability of achieving the minimal standards 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 minimal standards 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

Abdi, H. (2010). Partial least squares regression and projection on latent structure regression (PLS regression). *Wiley Interdisciplinary Reviews: Computational Statistics, 2*(1), S. 97-106.

Angelone, D. & Keller, F. (2019). *ÜGK 2017 Schulsprache und erste Fremdsprache. 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.

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.

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

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.

Little, R., & Rubin, D. (2002). *Statistical analysis with missing data* (2 ed.). New York: Wiley.

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

Lüdtke, O., & Robitzsch, A. (2010). Umgang mit fehlenden Daten in der empirischen Bildungsforschung. In S. Maschke, & L. Stecher, *Enzyklopädie Erziehungswissenschaft Online. Fachgebiet Methoden der empirischen erziehungswissenschaftlichen Forschung, Quantitative Forschungsmethoden* (S. 723-729). Weinheim: Juventa.

Lüdtke, O., Robitzsch, A., Trautwein, U., & Köller, O. (2007). Umgang mit fehlenden Werten in der psychologischen Forschung. *Psychologische Rundschau, 58*(2), S. 103-117.

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.

Mevik, B.-H., Wehrens, R., & Liland, K. H. (2016). pls: Partial Least Squares and Principal Component Regression. *R package version 2.6-0*. https://CRAN.R-project.org/package=pls.

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. (2019). *ÜGK 2017 – Technical report: Student questionnaire data*. St. Gallen: Pädagogische Hochschule St. Gallen.

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.

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, D. B. (1976). Inference and missing data. *Biometrika, 63*, S. 581-592.

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

Schafer, J. L., & Graham, J. (2002). Missing data: Our view of the state of the art. *Psychological Methods, 7*(2), S. 147-177.

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.

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

van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software, 45*(3), S. 1-67. Von http://www.jstatsoft.org/v45/i03/ abgerufen

Verner, M., & Helbling, L. (2019). *Sampling ÜGK 2017. Technischer Bericht zu Stichprobendesign, Gewichtung und Varianzschätzung bei der Überprüfung des Erreichens der Grundkompetenzen 2017*. 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.