Sampling techniques and weighting procedures for complex survey designs

–

The school cohorts of the National Educational Panel Study (NEPS)

Dissertation

Presented to the Faculty for Social Sciences, Economics, and Business Administration at the University of Bamberg in Partial Fulfillment of the Requirements for the Degree of

DOCTOR RERUM POLITICARUM

by
Hans Walter Steinhauer, Diplom Volkswirt
born January 26, 1984 in Kaiserslautern, Germany

Date of Submission
June 5, 2014

Principal advisor:	Professor Dr. Susanne Rässler University of Bamberg, Germany
Reviewers:	Professor Mick P. Couper, PhD University of Michigan, United States of America Professor Dr. Mark Trappmann University of Bamberg, Germany
Date of submission:	June 5, 2014
Date of defence:	September 17, 2014

This page is intentionally left blank.

Abstract

The National Educational Panel Study (NEPS) set up a panel cohort of students starting in grade 5 and grade 9. To realize the corresponding samples of students NEPS applied a complex stratified multi-stage cluster sampling approach. To allow for generalizations from the sample to the universe especially aspects of complex sample designs have to be considered and are reflected by design weights.

When applying multi-stage sampling approaches unit nonresponse, that is, units refuse to participate, may occur on each stage where decisions towards participation are made. To correct for potential bias induced by refusals of schools and students the derived design weights need to be adjusted. Since participation decisions differ in many ways, for example by stage (school- or student-level), time (in the forerun of or during the panel) or reasons (school-level: workload or participation in other studies, student-level: not interested in the study, resentment to testing), design weights need to be carefully adjusted to reflect the participation decisions made on each stage properly. Participation decisions on the school level take information from sampling and the school recruitment process into account and are modeled using binary probit models with random intercept considering the federal-state-specific recruitment.

Schools participating are subsampled providing access to students in grade 5 and 9. Subsampling within schools provides a sample of two classes if at least three are present, otherwise all classes are selected. In creating design weights this subsampling needs again to be incorporated in the weights. The students decision process on the next stage has to be accounted for in providing unit nonresponse adjusted weights. These decision processes take clustering at the school level as well as information on the initial sample, that is, respondents and nonrespondents, into account. The resulting net sample forms the panel cohorts of students in grade 5 and 9.

Based on the panel cohorts each student can again decide whether to participate or not for each successive wave. Providing additional information obtained in a parental interview with one parent this multi-informant perspective makes consideration of an additional participation decision necessary. Since participation decisions of a student and a parent are unlikely independent they should be modeled appropriately using bivariate models. To again account for a cluster structure these models are extended with a random intercept on the school level.

All these aspects of complex sample and survey designs as well as the different participation decisions involved need to be considered in weighting adjustments. The results point at typical characteristics influencing partic-

ipation decisions of schools, students and parents. Besides that the results stress the need to account for sample design and the nature of decision processes involved resulting in the actual participation.

Basis for this thesis

Earlier papers

This thesis is in parts based on work published in earlier papers by

- Aßmann et al. (2011),
- Aßmann, Steinhauer, and Rässler (2012),
- Steinhauer, Blossfeld, and Maurice (2012),
- as well as supplements of the data manuals accompanying the Scientific Use Files of the Starting Cohorts 3 (Grade 5 students) and Starting Cohort 4 (Grade 9 students) of the NEPS.

Data used

This thesis uses data from the National Educational Panel Study (NEPS). Results presented in this thesis are based on data mostly available as scientific use files. The corresponding data sets are:

- Starting Cohort 3 – Grade 5 (Paths through Lower Secondary School - Education Pathways of Students in 5th Grade and Higher)
  DOI:10.5157/NEPS:SC3:1.0.0
  DOI:10.5157/NEPS:SC3:2.0.0
- Starting Cohort 4 – Grade 9 (School and Vocational Training - Education Pathways of Students in 9th Grade and Higher)
  DOI:10.5157/NEPS:SC4:1.1.0.

The NEPS data collection is part of the Framework Programme for the Promotion of Empirical Educational Research, funded by the German Federal Ministry of Education and Research and supported by the Federal States.

Statistical software used

All analysis provided are based on The R Project for Statistical Computing (R Core Team, 2014). See Appendix E for further information.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The National Educational Panel Study (NEPS) provides data on various aspects of competence development, educational decisions, learning environments, migrational background and returns to education. The design of the National Educational Panel Study focuses on the life course perspective as introduced by Baltes, Reese, and Lipsitt (1980) and extended by Elder, Johnson, and Crosnoe (2003). Therefore the NEPS adapted a multicohort sequence design (Blossfeld & Maurice, 2011), shown in Figure 1.1.

This stringent commitment to the longitudinal focus is most central for the design of the NEPS. The cohorts are positioned at central stages within the educational system as well as at transitions relevant for educational careers. Since this definition of cohorts differs from others they are referred to as starting cohorts. This design allows to quickly provide data to the scientific community for each of the starting cohorts. Starting Cohorts are samples of a special cohort that will be followed over time. These six Starting Cohorts (SC) cover the entire lifespan and comprise *Early Childhood* (SC1), *Kindergarten children* (SC2), *Grade 5 and grade 9 students* in primary and secondary schools (SC3 and SC4), *First-Year Students* (SC5) as well as *Adults* (SC6).

The starting cohorts positioned at key transitions are kindergarten children and students in the ninth grade of secondary schools. Grade 5 students as well as the freshmen cohort are positioned at the beginning of a new educational stage. Besides that the adults cohort focuses on the educational careers in adulthood and the early childhood cohort studies the infant development (Blossfeld, Maurice, & Schneider, 2011). Each starting cohort is followed up so that target persons can be studied in different stages as well as in different developmental statuses of their individual careers throughout the entire lifespan. The longitudinal design not only permits the analysis of dynamics but also the determinants of individual behaviour. In contrast

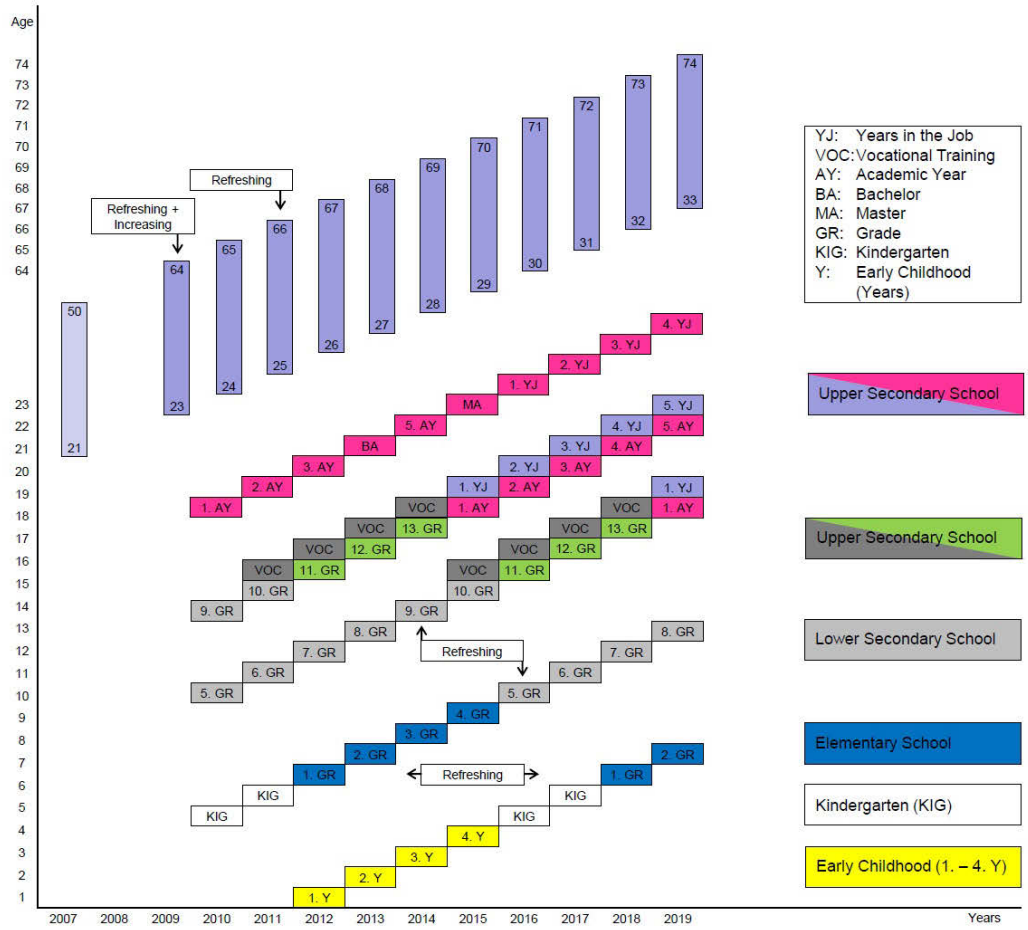


Figure 1.1: The multicohort sequence design of the National Educational Panel Study.

to cross-sectional designs this further allows to study decisions within political, familiar and social contexts (Trivellato, 1999). In order to provide a rich database allowing for analyses of different educational topics, the NEPS uses a multi-informant survey approach. Adapting such a survey design, the NEPS enriches for example data obtained from testing and surveying students with information obtained within a parental telephone interview as well as information provided by teachers and institution heads.

Since this design is complex and to account for particularities of each starting cohort sophisticated sampling designs are applied. Focusing on students surveyed and tested in the school cohorts in grade 5 and grade 9, that is, Starting Cohorts 3 and 4, complex survey designs and their consequences

for sampling designs and weighting strategies will be thoroughly discussed. The emphasis is on complex sampling designs that are applied respecting the stratified and hierarchical school system in Germany as well as deriving design weights and adjusting them to compensate for unit nonresponse.

In the phase of planning the samples of schools for grade 5 and 9 several issues arose concerning stratification, sample sizes, and allocation of sample size within a stratified multistage cluster sampling design. Clusters are sampling units that are grouped together for example in an institution. In an educational context schools or classes within schools form such clusters of students. These clusters (schools as well as classes) are commonly of unequal size. Multistage (cluster) sampling can be applied to hierarchically structured clusters such as students within classes within schools. In this case a school could be sampled on the first stage (on the top level of the hierarchy) and classes could be sampled on the second stage (on a lower level of the hierarchy). From this example it can be seen that in multistage cluster designs the sample size (for example the number of students) on the ultimate stage becomes a random variable (see Kish (1995, pp. 217ff) or Särndal, Swensson, and Wretman (2003, p. 127)). Following Kish (1995) appropriate measures such as stratification by cluster size, splitting or combining clusters or probability proportional to size sampling exist to achieve an approximate control of ultimate stages' sample size. To determine a sufficient sample size of clusters on the first stage appropriate measures were adopted. Furthermore the number of clusters, that is, schools, to sample on the first stage was derived by means of simulation to achieve a desired number of units of the ultimate stage, that is, students.

The Starting Cohorts 3 and 4 focus on students in grade 5 and 9 within secondary schools. Schools were grouped in strata according to their school type to account for the heterogeneity of educational degrees achievable in the different school types. Within each stratum a two-stage cluster sampling approach was adopted. On the first stage schools (as clusters of classes) were sampled providing access to the students clustered in classes within these schools. On the second stage classes (as clusters of students) were sampled and all students therein were asked for participation.

Sampling schools for Starting Cohorts 3 and 4 was done in school year 2009/10 using information on schools from the school year 2008/09. Surveying and testing students took place in school year 2010/11. Within these two years fluctuation (students repeating a class or leaving school), ongoing school reforms or closing down of schools reshape the population of schools. In probability proportional to size (*pps*) sampling (as one measure to achieve an approximate control of sample size) the measure of size is the characteristic to which the probability for sampling is proportional to. When choosing

the number of students or the number of classes per school this characteristic might change over time. Using an appropriate measure of size in *pps* sampling on the first stage, that allows for an inverse *pps* on the second stage results in a sample where each individual has the same design weight, that is, a self weighting sample.[1] Since the characteristic for a measure of size in *pps* sampling based on information from school year 2008/09 and an inverse *pps* based on the same characteristic two years later which might change during that period, an exact self weighting sample can only be realized having a constant characteristic. Due to the changes induced by this gap in time different characteristics for constructing the measure of size were evaluated. The aim was to find a measure of size that is stable over this period to get as close as possible to a self weighting sample. The evaluation uses the actual sampling frame from school year 2008/09 and the one from the previous school year 2007/08. The measure of size for sampling was chosen to yield minimum variation of design weights induced by changes due to the gap in time.

Applying the above described stratified two-stage cluster sampling design schools and classes were sampled and corresponding design weights were derived for schools and students. The sampled schools (also referred to as original schools) were contacted by the survey research institute and asked for participation. Since participation is voluntary and schools could refuse each sampled school was assigned up to four replacement schools and if necessary contacted in a fixed order to counteract a reduction of sample size. Within each participating school two classes were sampled if at least three were present, otherwise all classes were sampled. One teacher within each school responsible for communication with the survey research institute (coordinator) was asked to list all students in the sampled classes on the so called *Schüler-Teilnahme-Liste* (engl.: student participation list, see Table B.2). Each student willing to participate had to return an informed consent signed by a parent if the student was not of legal age. This participation status was recorded on the list and it further contained information on the initial sample such as sex, month and year of birth, school type, etc. One part of this list was returned to the survey research institute and later on to the methods department of the NEPS. The other parts remained in the school.

The derived design weights apply to the initial sample of original schools and their students. They would be applicable if participation in the study is mandatory. Since schools as well as students have the possibility to refuse participation the derived design weights need to be adjusted to compensate for refusal (i.e., unit nonresponse). Due to the two-stage sampling design and

[1] This is true under certain circumstances discussed in more detail in Sections 3.4 and 3.5.

the successive decision processes on each stage unit nonresponse can occur on each stage. Therefore the derived design weights were adjusted on each level, that is, school and student level, successively to correct for unit nonresponse. Adjustments on the school level utilized information from sampling (for example strata, number of classes, etc.) as well as information arising from the recruitment process of schools (for example number of schools recruited per federal state). Adjustments on the student level utilized information provided by the coordinator on the students participation list. The adjustments of the initial sample were done respecting the structure of the school system, two-stage decision processes and the clustering of students within schools. These adjusted design weights apply to students willing to participate in the panel, that is, the panel cohort, and will be referred to as panel entry weight.

At the day of surveying and testing students in schools students were absent due to illness, weather conditions or other reasons although they were willing to participate. The absence of students at 'test days', referred to as temporary drop-out, can occur in each wave of the panel. This makes further adjustments for wave-specific unit nonresponse necessary. These are based on information available for the entire panel cohort. Adjusted weights for the panel cohort are provided for different groups. These groups include wave-specific participants (cross sectional weights), all-time-participants (panel cohort members participating in each wave up to the actual wave) or subgroups of interest (for example students and parents or participants with available tests from each second wave).

Wave-specific adjustments correct for unit nonresponse in the corresponding wave. Therefore participation decisions are modeled using available information which is mostly not varying over time. To account for the clustering of students a random intercept model is adopted using a probit link function.[2] The models will become more sophisticated in the progress of the panel, since more information arises. Information that is missing or not available in the first wave may arise in the second wave so that the model for the first wave can be updated and becomes more accurate.

The group of panel cohort members participating in each wave up to the actual wave, that is, the all-time-participants, is modeled almost analogously to the cross sectional adjustment models. The models are conditioned on the participation status in previous waves and extended by information arising in the progress of the panel.

Finally, models for weighting adjustments in the subgroup of the panel cohort students in grade 5 and their parents have to consider two possibly

[2]The probit specification is used to be consistent with extensions of the model introduced later on.

correlated decisions. Thereby both, the students and the parental survey are subject to nonresponse. To provide unit nonresponse adjusted weights for the relevant group of participating students and parents, a bivariate probit with random intercept allowing for clustering at the school level is used. Yet there is no implementation of this model for the statistical software package R (R Core Team, 2014) available. Thus it is provided in the Appendix D. The model is estimated using a simulated maximum likelihood procedure based on the importance sampler of Geweke, Hajivassiliou and Keane (GHK-simulator) documented in Geweke and Keane (2001).

The empirical results of the adjustment models point at significance of typical explaining factors of unit nonresponse and reveal the importance to consider a clustering structure as well as a correlation parameter regarding the possibly correlated participation processes of parents and students.

This dissertation proceeds along the order of events. Chapter 2 will give a review on sampling and weighting as basis for the description in the following chapters. Chapter 3 gives detailed insights on the sampling design of SC3 and SC4. The weighting procedures applied to SC3 and SC4 are discussed in Chapter 4. The bivariate probit model with random intercepts and its application to weighting adjustments in SC3 is shown in Chapter 5. A summary and an outlook to future research in the field of weighting longitudinal cohorts with complex survey designs will be given in Chapter 6.

The Appendix A contains the lists of abbreviations and symbols used throughout the following chapters. Appendix B includes the tables in order of their appearance within the text. An illustration of the GHK-simulator is given in Appendix C. The syntax of the code used for estimation of the bivariate probit model with random intercept as introduced in Chapter 5 is displayed in Appendix D. Finally Appendix E gives R's session information.

Chapter 2

Reviewing sampling and weighting techniques

Chapter outline:
Reviewing the basic sampling techniques this chapter will serve as theoretical basis for the thorough description of the samples in the subsequent chapter. The review deals with the preparation of the sampling frame, a general description of sampling as well as a formal description of deriving design weights. This is followed by a summary of the sampling techniques applied in SC3 and SC4. It finishes with aspects of weighting adjustments.
This chapter is in parts based on the work published in earlier papers by Aßmann et al. (2011) and Aßmann et al. (2012).

2.1 The sampling frame

The starting point for each sampling design is the definition of the target population in terms of temporal and regional restriction as well as further characteristics describing the population. The basis for sampling is most often provided in form of a complete list of population elements[1] containing available information for each element. This complete listing is referred to as *sampling frame*. Sampling frames usually are obtained from administrative data bases, for example registration offices and their registers or complete lists of schools, universities and communities provided by the States Bureaus of Statistics (*Statistische Landesämter*).

When requesting administrative listings the frames provided (for example for schools or communities) cannot always be up to date since ongoing

[1]The terms element and unit will be used synonymously.

reforms, closing and merging of institutions, deaths and births as well as migration and immigration reshape populations. Therefore any sampling frame available can only be a snapshot of the population at a certain point in time. This in fact has an impact on designing a sampling strategy. Furthermore frames are provided by states in different formats, quality and informational content. So the aim to construct a nationwide frame can become a challenging task. For more details on the provision and harmonization of a frame see Aßmann et al. (2012).

2.2 Sampling techniques

There exist different methods to draw a sample from a target population (also universe) if a sampling frame is available. Let U denote the universe consisting of N units $U = \{u_1, \ldots, u_i, \ldots, u_N\}$. Let further the set of all samples S contain all possible samples s of size n for a given selection scheme so that $s \in S$. The probability p to draw one certain sample $s \in S$ is given by the function $p : S \rightarrow [0; 1]$ with $p(s) > 0$ and $\sum_{s \in S} p(s) = 1$ (see Särndal et al., 2003, pp. 27f). The tuple (S, p) is called sample design. For a given sample design the first order inclusion probability π_i is the probability that the i^{th} element u_i is sampled (see Särndal et al., 2003, p. 31):

$$\pi_i = P(u_i \in S) = \sum_{s \ni u_i} p(s).$$

The probability that the units u_i and u_j $(i \neq j)$ are sampled jointly into the sample is given by the second order inclusion probability π_{ij}

$$\pi_{ij} = P(\{u_i; u_j\} \in S) = \sum_{s \ni \{u_i; u_j\}} p(s).$$

The summation is over all samples s that do contain the element u_i. The design weight d_i for unit u_i is usually given by the inverse of its first order inclusion probability

$$d_i = \frac{1}{\pi_i} \qquad \forall i = 1, \ldots, n.$$

With respect to the design (S, p) for each sampled unit u_i the design weight d_i can be derived. The design weight (or also base weight) can (in some designs) be interpreted as the number of population elements represented by a sampled unit (see Wolter, 2007, p. 18). In simple random sampling without replacement the first order inclusion probability arises as

$$\pi_i = \frac{n}{N} \qquad \forall i = 1, \ldots, n.$$

Hence the design weight is (see Tillé (2006, p. 45) or Särndal et al. (2003, p. 66))

$$d_i = \frac{1}{\pi_i} = \frac{N}{n} \qquad \forall i = 1, \ldots, n$$

being constant for all sampled units. In case of simple random sampling without replacement the design weight d_i gives the number of units represented in the population by the sampled unit. That is, the sum of the design weights results in the population size N, since

$$\sum_{i=1}^{n} d_i = \sum_{i=1}^{n} \left(\frac{n}{N}\right)^{-1} = \sum_{i=1}^{n} \left(\frac{N}{n}\right) = n \cdot \frac{N}{n} = N.$$

This is not necessarily the case for all sampling designs. Such is the usual case in sampling with replacement or under consideration of ordering. In statistical inferences constant design weights (as above) can be ignored. In contrast, nonconstant design weights cannot be ignored, since they arise from complex designs such as stratified, multistage or cluster sampling (Snijders & Bosker, 2012). The design weights are commonly used for estimation of population parameters (for example totals, means, ratios) for some variable of interest y. Applying the Horvitz-Thompson estimator (Horvitz & Thompson, 1952) the estimated population total $\widehat{Y}$ is computed as the sum of weighted observations of y_i for unit $u_i \in s$

$$\widehat{Y}_{HT} = \sum_{i=1}^{n} y_i d_i = \sum_{i=1}^{n} \frac{y_i}{\pi_i}.$$

For a probability distribution p on S and a general estimator τ the tuple (p, τ) is referred to as strategy. One main focus of sample selection theory therefore is to find sample designs well interacting with estimators, which means finding appropriate strategies. To achieve this aim it is necessary *in advance* to be aware of analyses of interest when the sample is realized and information is available.

But designing a sampling scheme also has to take practical and economical aspects into account. Complex survey designs often do not allow for sampling units by simple random sampling. Economic reasons finally drive decisions towards certain sampling designs, even if those result in more complex methods of data analysis. The aim of providing accurate estimates for a population of interest is often achieved by choosing other designs than simple random sampling.

2.2.1 Explicit Stratification

Explicit stratification assigns each population element u_i to distinct non-overlapping strata. After stratifying the population U in $h = 1, \ldots, H$ strata $U_1, \ldots, U_H$ with population sizes $N = \sum_{h=1}^{H} N_h$ samples of size n_h are taken independently from each stratum. Explicit stratification serves several aspects as discussed in Kish (1995, pp. 76-77) and Särndal et al. (2003, p. 100).

- It can be used to gain precision in estimates (i.e. decrease their variance).
- By sampling independently from each stratum different sampling designs can be applied to the strata. This is especially useful when the populations are extremely heterogeneous or their elements differ by nature.[2]
- Stratification by cluster size is one possibility to control sample size in case of clusters with unequal size.[3]
- When samples are drawn from several frames of different quality stratification may become necessary. Characteristics that are relevant for sampling may be measured differently, provided in a different way or even be missing.[4]
- Subpopulations can be of special interest for a study and separate estimates are needed. Dividing the population in strata serves this aspect.[5]
- Administrative reasons may be a further argument for stratification.

Figure 2.1 illustrates explicit stratification of schools. The schools in the population (Subfigure 2.1a) are assigned to the strata $h = 1, \ldots, 5$. Thereby each stratum can contain a different number of schools. The schools within each stratum are in this example identical with respect to the stratification characteristic color of the school building[6] (Subfigure 2.1b) but still different with respect to other characteristics (for example color of the roof). When

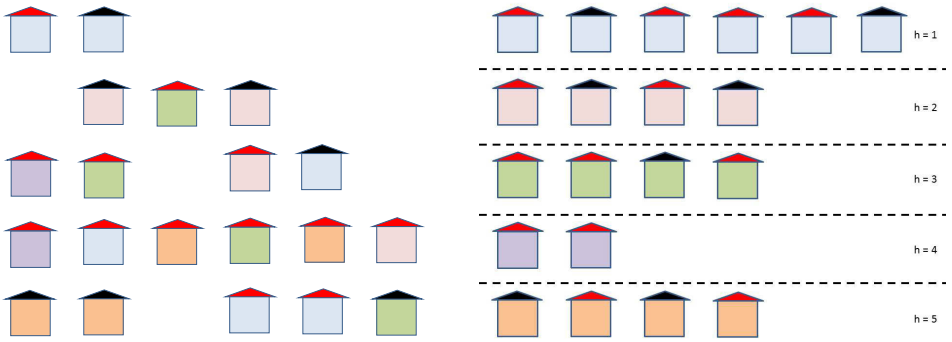
[2]This aspect applies in sampling students. The samples were stratified by school type to account for heterogeneity between different school types; especially between regular and special schools.

[3]See also Section 2.2.2 for further details.

[4]This issue is addressed in Section 3.4 and 3.5 for sampling students in regular (*allgemeinbildende Schulen*) and special (*Förderschulen*) schools.

[5]This is shown in Section 3.4 when oversamplings of students in vocational tracks are considered.

[6]It could have been any other characteristic such as Federal State or school type.



(a) Population of schools. (b) Stratified population of schools.

Figure 2.1: Explicit stratification of schools by color of the school building.

using explicit stratification sample size (for example $n = 10$) needs to be allocated to the strata. The allocation of the total sample size n to the H strata can be achieved in several ways whereas it always has to be ensured that $n = \sum_{h=1}^{H} n_h$ (Cochran, 1977, p. 89). In equal allocation sample size per stratum $n_h = \frac{n}{H}$ is equal across strata $n_1 = n_2 = \ldots = n_H$ (for example sample $n_h = 2$ blue, red, green, purple and orange schools). Allocating the total sample size proportional to the size of the population of stratum h results in unequal sample sizes per stratum $n_h = n \cdot \frac{N_h}{N}$ (for example 3 blue, 2 red, 2 green, 1 purple and 2 orange schools) and equal sampling fractions $f = \frac{n_h}{N_h}$. To gain precision in estimates for small strata the sample size is increased resulting in unequal sampling fractions. This is called oversampling of strata. To ensure a minimum (and maximum) number of sampling units per stratum Gabler, Ganninger, and Münnich (2012) derived an allocation algorithm with respect to bounded design weights while considering optimality in allocation for stratified random sampling.[7] For a numerical solution see Münnich, Sachs, and Wagner (2012)

Explicit stratification results in independent samples. If this is not desired or possible implicit stratification, that is, sorting the sampling frame by characteristics available, together with systematic selection can–to some extend–help to ensure having elements of implicit strata within the sample. The result will be similar to that of a proportionate stratified sample (Kish, 1995, p. 85). Implicit stratification will not be useful for small strata (for example purple schools in Subfigure 2.1b), since they may not be sampled or only in small numbers.

[7] This approach was discussed in sampling the Starting Cohort 1 – Early Childhood of the NEPS.

Besides explicit and implicit stratification it is also possible to stratify the sample after its realization. When information is not available for sampling but is collected during the field period (for example individual characteristics such as age, sex, occupation) these information can be used for post-stratification.

2.2.2 Multistage and multistage cluster sampling

Multistage sampling is used to get access to hierarchical structured or clustered populations. Sometimes it is not possible to sample individuals directly, since no frame is available on the individual level. In this case clusters of individuals can be sampled instead if the individuals are grouped or clustered (and there is an available frame). In cluster sampling a cluster (for example a school) contains a set of units (students) and within a cluster all units can be surveyed (Särndal et al., 2003, p. 124). If further samples are drawn within the cluster, sampling is done on multiple stages and is therefore referred to as multistage sampling. In two-stage sampling units on the first stage are referred to as *primary sampling units* (PSU, e.g. schools) and those on the second stage are called *secondary sampling units* (SSU, e.g. classes). The primary sampling units on the first stage are disjoint sub-populations of grouped secondary sampling units. In multistage designs this hierarchy is extended to the ultimate stage (Särndal et al., 2003, p. 125) and sampling selection processes can differ at each stage so that the variety of designs increases rapidly.

Except for the case of equal sized clusters on each stage the resulting sample size becomes a random variable and is not under control (see Kish, 1995, pp. 217ff or Särndal et al., 2003, p. 127). Controlling sample size is essential to most surveys because there are variable costs increasing the total costs by each sampled unit. Kish (1995, p. 217) points out that "Exact control of sample size is unnecessary and impossible in most situations. [...] We should aim at an *approximate control* that is both feasible and desirable."

To achieve this *approximate control* in the case of unequal cluster size Kish recommends not to use uncontrolled random sampling procedures and to stratify by cluster size. Another way is to split or combine clusters of unequal size to clusters of a more similar size. On the second stage also size stratified sampling can be applied with different sampling fractions or a fix number of elements can be sampled. Finally probability proportional to size sampling of units (no matter on which stage) can help to get less variation in the initial sample size (Kish, 1995, pp. 219f). In probability proportional to size sampling each sampling unit is assigned a measure of size (*mos*) which can be a natural characteristic of that unit (for example number of students

of a school) or any value assigned to it (Kauermann & Küchenhoff, 2011, pp. 104f).

Mehrotra, Srivastava, and Tyagi (1987) show another way for controlling sample size by discarding an excess number of clusters randomly from the sampled clusters. The advantages of the proposed procedure are convergence of planned and realized sample sizes and thereby a reduction of survey costs. But discarding clusters from the sampled clusters therefore results in less efficient estimators. Discarding clusters can be done if information about cluster size is reliable or can be estimated accurately. When sampling and surveying is done at different points in time cluster size can change significantly and though discarding clusters can become a challenging task.[8] Furthermore Aliaga and Ren (2006) determine the optimal number of clusters to sample in a two-stage design for a given linear cost function.

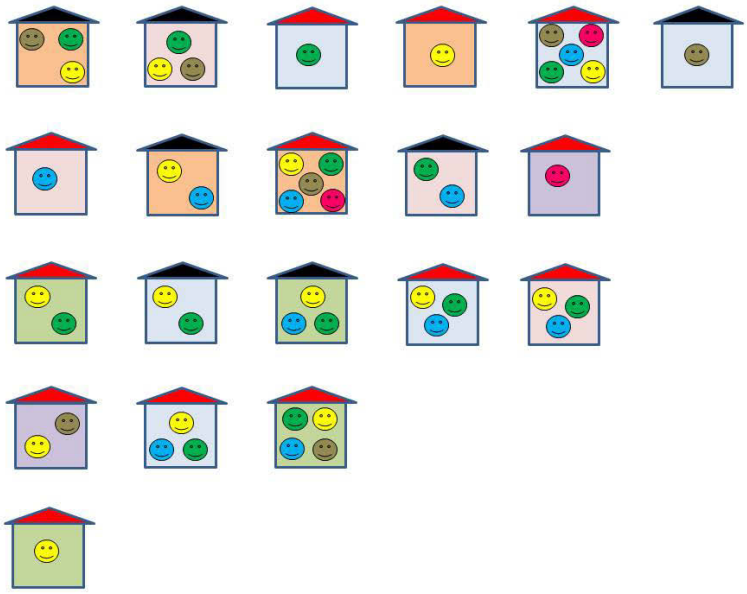


Figure 2.2: Two-stage cluster sampling.

Figure 2.2 illustrates clusters of students (i.e., classes) in schools. The classes indicated by colored smileys are located within schools. In two-stage cluster sampling a number m^I of primary sampling units C^I (e.g. schools) is sampled on the first stage (denoted by the superscript). Within a sam-

[8] In NEPS sampling schools was based on a frame of the school year 2008/09. Sampling was done in 2009 and surveying and testing of students followed in school year 2010/11. See Section 3.4 for further details.

pled first stage cluster C_j^I a number m^{II} of secondary sampling units C^{II} (e.g. classes, colored smileys) is sampled. To demonstrate the problem of unequal cluster sizes let the yellow smileys be classes of 20 students, the green smileys be classes of size 25 and the brown smileys classes of size 30. Having sampled two schools (e.g. the first and the second in the top row) and sampling one class per school the sample sizes can vary. Sampling a yellow and brown smiley will result in the same sample size as sampling two green smileys. But any other combination will either yield smaller sample sizes (green and yellow, yellow and yellow) or larger sample sizes (green and brown, brown and brown). So, depending on the samples drawn, the sample sizes can be $n \in \{40; 45; 50; 55; 60\}$.

2.2.3 Systematic and systematic unequal probability sampling

Systematic sampling is an alternative to random selection of units. In systematic sampling with equal probabilities each unit is assigned an interval of length 1 (for illustration see Figure 2.3). The selection interval length $k = \frac{N}{n}$ is the population size N divided by the sample size n. Starting from a randomly drawn starting point $r \in \{1, \ldots, k\}$ (i.e. within the first interval) every k^{th} unit is selected. The units u_i selected by systematic sampling are then

$$s = \{u_r, u_{r+k}, u_{r+2k}, , u_{r+3k}, \ldots, , u_{r+(n-1)k}\}$$

and the inclusion probability for each unit i is the same $\pi_i = \frac{1}{k}$ (Madow, 1949). The only unit sampled randomly is the first one. Since the rest of the sample is determined by the first unit sampled. Systematic selection can be seen as single stage cluster sampling where only one cluster is selected (Kauermann & Küchenhoff, 2011, p. 172).

One drawback of systematic selection is that some units u_i and u_j do not have a second order inclusion probability π_{ij}. For example let u_i and u_j be neighbours, than there is no chance for them to end up together in a sample. This drawback mainly effects variance estimation, for example for the Horvitz-Thompson estimator. An overview of variance estimation methods that can be applied in this and other cases is given in Münnich (2008). A more details can be found in Wolter (2007).

Figure 2.3 illustrates sampling $n = 12$ units from a population of size $N = 120$ via systematic sampling with a random start value $0 < r \leq 12$. The N units of the population are ordered on the axis, where each tick mark indicates one unit. The interval length between to neighbouring units is

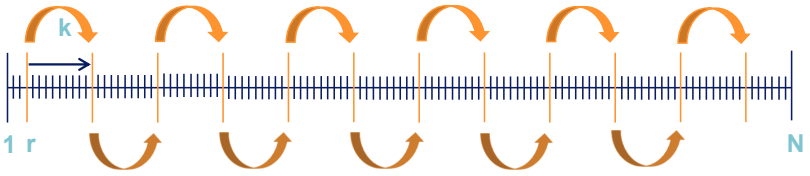


Figure 2.3: Systematic selection of units with random start.

equal to 1. Let $r = 3$ and $k = \frac{120}{12} = 10$ then the sample would consist of the units $s = \{u_3, u_{13}, \ldots, u_{113}\}$ (indicated by the larger tick marks and the arrows). This selection procedure is easily applicable but also can become more sophisticated for example when it comes to selection with probability proportional to size or when k is not an even number. For a brief discussion of systematic sampling procedures see Madow and Madow (1944) and Madow (1949, 1953). For extension to circular methods as solution to non-integer k see Uthayakumaran (1998) or Kish (1995, p. 116)

In systematic sampling (as well as other random sampling procedures) units can be sampled with unequal selection probabilities. When systematic sampling is performed with *pps* each unit u_i is assigned a measure of size mos_i. The total measure of size is

$$MOS = \sum_{i=1}^{N} mos_i \qquad \text{and} \qquad MOS_i = \sum_{j=1}^{i} mos_j$$

is the cumulative measure of size and the selection interval length k changes to $k = \frac{MOS}{n}$ (see Kish (1995, pp. 234ff) or Hájek and Dupač (1981, p. 113)). The inclusion probability for unit i then arises as $\pi_i = \frac{n \cdot mos_i}{MOS}$, see Tillé (1996) or Wolter (2007, pp. 332ff).

Table 2.1: Example for systematic probability proportional to size sampling.

	u_1	u_2	u_3	u_4	u_5	u_6	u_7	u_8	u_9	u_{10}	u_{11}	u_{12}
mos_i	1	3	3	1	1	1	1	2	1	1	2	3
MOS_i	1	4	7	8	9	10	11	13	14	15	17	20

Note: Systematic *pps* sampling is performed using `ppss()` implemented in the package `pps` (Gambino, 2012). Use this function with care, since it can not handle $\pi_i > 1$.

Table 2.1 shows a simple example for a universe consisting of 12 units. From this universe a sample of size $n = 4$ is taken using systematic probability proportional to size sampling. With random start point $r = 1.8812$ and

interval length $k = 5$ the units u_2, u_3, u_8 and u_{11} are sampled. Using systematic unequal probability sampling two neighbouring units can be selected, see Figure 2.4.

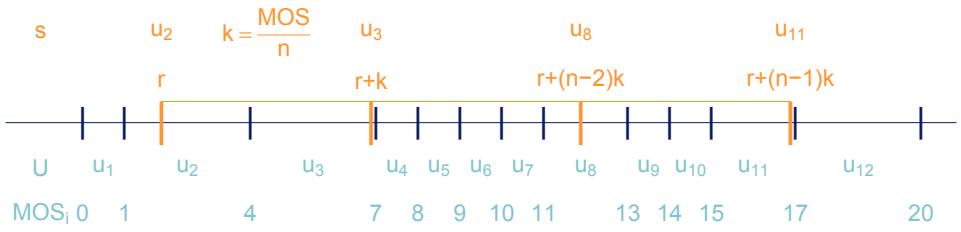


Figure 2.4: Graphical illustration for *pps* sampling according to Kish (1995, pp. 230, 234ff).

Figure 2.4 gives the graphical illustration of the example above. The cumulative measure of size MOS_i is given along the axis. The first unit selected is the unit which includes the random start r in its interval. From this starting point each element is chosen, whose interval includes r plus a multiple of k.

2.3 Design weights and their adjustment

Since non-mandatory surveys are typically affected by nonresponse. Nonresponse may occur for several reasons. Lepkowski and Couper (2002) separate the process leading to participation or nonparticipation into location of units, contacting units (given location) and cooperation of units (given location and contact). Therefore sampled units can end up as nonrespondents in each of the three steps. For example a sampled unit has moved and thus cannot be located. Other units might–for whatever reason–not be contactable. For example they can be in a hospital because of illness or have moved abroad. Thus these persons can not be contacted and asked for participation. Some of the sampled, located and contacted units will refuse to participate in the survey. Reasons for refusal vary between countries, survey topics, etc., see for example Lugtig (21.10.2013) These groups of people, that is those people that could not be located, contacted and those that refuse to participate, form the set of nonrespondents. This so called unit nonresponse might make adjustments of the design weights necessary, depending on the type of the

missing data mechanism. The derivation of final weights is, according to Kalton and Kasprzyk (1986), in general done in three steps:

1. Derivation of design weights
2. Sample weighting adjustments
3. Population weighting adjustments

In the *first step* design weights (also known as base weights) are usually computed as the inverse of the inclusion probability (as shown in Section 2.2). Thus, for most designs they are directly available after sampling. Design weights compensate for unequal probabilities of selection, unequal sampling fractions in stratified samples, that is, oversampling, or for subsampling (Kish, 1990, 1992).

In the *second step* the design weights are adjusted to correct for unit nonresponse. Kalton and Kasprzyk (1986) refer to this step as *sample weighting adjustment*. In multistage sampling procedures this step needs to be considered on each stage where nonresponse occurs. Sample weighting adjustments correcting for unit nonresponse usually result in increasingly varying weights and thereby lower the precision of survey estimates (Kalton & Flores-Cervantes, 2003).

The *third step*, referred to as *population weighting adjustment*, calibrates weights so that estimates conform to known parameters (for example totals or ratios) of the population. This last step corrects for potential bias due to incomplete coverage or non-coverage of the population and sampling error (Brick, 2013).

2.3.1 Sample weighting adjustment

After the sample is realized the sampled units have to be contacted and are asked to participate in the survey. This two stage process gives rise to two reasons why sampled persons might not be surveyed. Survey response depends on contact and cooperation. First, the sampled unit needs to be contacted. Second, given contact the unit decides to cooperate or not. Failing to establish contact as well as noncooperation will result in unit nonresponse, but for different reasons (Groves, 1998). Survey response therefore can become a threefold variable of participation, refusal and noncontact.When modeling unit nonresponse the two components, that is noncontact and refusal, should be modeled to avoid bias (Steele & Durrant, 2011). For both components of unit nonresponse the resulting sample will be biased if the persons not participating form a selective group.

The need for adjustments of the design weights depends on the missing data mechanism. The terminology originates from item nonresponse and multiple imputation, see Rubin (1987) or Little and Rubin (2002). But it is also applicable in the context of weighting adjustments. Therefore we need to come back to the strategy. The aim of every survey is to investigate on a variable of interest (y, for example educational aspirations) and use auxiliary information (x, for example educational degree of parents). Unit nonresponse causes both variables to be missing.

A unit is said to be *missing completely at random* (MCAR) if the probability of responding is depending neither on observed nor on unobserved characteristics. An extreme MCAR case would be if any person in the sample has the same response probability (Valliant, Dever, & Kreuter, 2013). In this case the responding part of the sample is a random subsample of the entire sample that allows for valid inferences. An example would be a student being ill at the day of the survey or a computer crash in computer based assessments.

Unit nonresponse is *missing at random* (MAR) if the probability of response depends on the data but only the auxiliary information available for respondents and nonrespondents. This information can either be marginal distributions from a census or individual-level data available for the entire sample. If this auxiliary information is at hand, a model for response propensities can be estimated. Lohr (2010) describes this as *ignorable nonresponse*. That is if a model can explain the mechanism of nonresponse and that it can be ignored if it is accounted for. This approach does not only allow for re-weighting the initial sample but also for documenting effects significantly influencing participation decisions. Therefore it is used in later re-weighting. Here it is not that nonresponse can be ignored and complete data methods can be applied. In the example MAR would be if the probability of response would depend on the educational degree of the parents which is observed.

When the probability of nonresponse depends on the variable of interest (y, for example educational aspirations) and cannot be accounted for by modeling the response based an the auxiliary information (x) units are *not missing at random* (NMAR). (Valliant et al., 2013, p. 319) also term this nonignorable nonresponse. This type of missing data mechanism is–if at all–hard to detect. One way of finding out about NMAR would make follow-ups necessary.

To correct for potential bias arising through unit nonresponse there are a variety of procedures available. Weighting is one of the most commonly used methods to correct for unit nonresponse in surveys (Little & Vartivarian, 2003). A general overview on weighting methods to correct for unit nonresponse is given by Kalton and Flores-Cervantes (2003). A more technical

overview is given by Holt and Elliot (1991).

One way is to adjust the number of participants to the initial sample size. That is the weight is multiplied by an adjustment factor δ

$$w_i = d_i \cdot \delta, \qquad \text{with } \delta = \frac{n_r + n_n}{n_r}. \tag{2.1}$$

Here n_r denotes the number of participants and n_n the number of nonparticipants. This approach implicitly assumes that unit nonresponse occurs completely at random. It can be extended in two directions. First the adjustment factor can be derived as the fraction of the sum of weights for all units divided by the sum of weights for the respondents. This is more appropriate in probability proportional to size samples. Second the above approach and the first extension can be modified by adjusting the weights within certain cells. These cells are formed by characteristics of the units themselves or of higher level units. For example the cells can be defined by sex, age group and cluster. This approach is referred to as cell weighting and is one of the most commonly used approaches to correct for unit nonresponse in (sample) weighting adjustments (Rässler & Riphahn, 2006; Rässler & Schnell, 2003).

Table 2.2: Example for cell weighting.

	women		men	
	n_w	$\sum_{i=1}^{n_w} d_i$	n_m	$\sum_{i=1}^{n_m} d_i$
sampled	500	1200	500	1100
responding	400	700	250	650
δ	1.250	1.714	2.000	1.692

Assume the realization of a sample of size $n = 1000$ (shown in Table 2.2) with an equal ratio of women and men (i.e. $n_w = n_m$, of course just for convenience), whereas $n_w = 400$ women and $n_m = 250$ men respond to the survey. A naive approach would be adjusting the design weights neglecting information on sex resulting in an adjustment factor $\delta = \frac{1000}{650} \approx 1.538$. Making use of the information on sex results in gender specific participation rates of $p_w = \frac{400}{500} = 0.8$ and $p_m = \frac{250}{500} = 0.5$ and the corresponding adjustment factors (see Equation (2.1)) $\delta_w = p_w^{-1} = \frac{500}{400} = 1.25$ and $\delta_m = p_m^{-1} = \frac{500}{250} = 2.0$. Basing the adjustments on the sum of weights would lead to adjustment factors for women $\delta_w = \frac{1200}{700} \approx 1.714$ and men $\delta_m = \frac{1100}{650} \approx 1.692$.

A more sophisticated approach is to adjust the respondents by the inverse of their estimated response propensity. The basic idea (harking back

to Rosenbaum and Rubin (1983)) is to find a sampled element that is most similar to the refusing. Now this similar element has to "represent" more population elements. To do so the response propensity is most often estimated using logit[9] models for binary data, which need information on participants as well as non-participants. The inverse of the estimated response propensity $\widehat{\lambda}_i$ for element i is multiplied by the design weight and finally the adjusted design weight is (Rendtel & Harms, 2009)

$$w_i = d_i \cdot \widehat{\lambda}_i^{-1}. \tag{2.2}$$

Frameworks that can be applied to estimate these response propensities are more thoroughly discussed in Section 4.2. Note that d_i is a fixed value depending on the sample design (S, p) only. The value of w_i, since multiplied by $\widehat{\lambda}_i^{-1}$ estimated from a model, in contrast is an estimate based on the realized sample. Asymptotic properties of estimators using nonresponse adjusted design weights w_i based on the estimated response propensity are discussed by Holt and Elliot (1991), Kim and Kim (2007) and Henry and Valliant (2012).

2.3.2 Population weighting adjustment

The idea behind population weighting adjustments is to make sample distributions and parameters conform to known distributions and parameters of the population. For population weighting adjustment most of the methods used in sample weighting adjustment can be applied as well (Kalton & Flores-Cervantes, 2003). Unlike sample weighting adjustments population weighting adjustments do not need information for nonrespondents (Brick & Kalton, 1996). For population weighting adjustments distributions or parameters of the population need to be known. Further methods for population weighting adjustments include calibration, general regression estimation (GREG), raking or post stratification.

Post stratification can make use of data collected in the survey that was not available before (for example age or sex). For known totals of subgroups of the population the weights for units are adjusted within subgroups (or classes, poststrata) so that the estimate conforms to the total within this class. This method therefore reduces bias induced by undercoverage. One problem with this approach arises if the characteristics used in forming the poststrata are not measured in the same way for the sample and the population (for example migrational background). Never the less post stratification

[9] Laaksonen (2005) finds the logit link function to be used most often and further discusses the characteristics of probit, log-log and clog-log. In his findings the choice of link functions only differs slightly in estimated propensities.

is, according to Brick and Kalton (1996), one of the most frequently used population weighting adjustments.

For a large number of characteristics available for the sample and the population together with parameters of interest post stratification may suffer from small number of cases within the poststrata. In this case an iterative approach called raking is superior. It iteratively adjusts the weights in a way that marginal distributions of auxiliary information conform to those of the data (Brick & Kalton, 1996). The approach, also referred to as iterative proportional fitting, was suggested by Deming and Stephan (1940)

The calibration approach systematically incorporates auxiliary information into the procedure (Särndal, 2007). Calibration thereby is not only a procedure for population weighting adjustment, but also incorporates the estimation of population parameters. The weighting adjustment computes weights using auxiliary information. These adjustments are–at the same time–restrained to one or more calibration equations, see Särndal (2007).

General regression estimation is another way to incorporate auxiliary information in the estimation step. Deville and Särndal (1992) note that the GREG can be derived also from calibration by focusing on the weights. They show that the weights used in GREG are closely to those derived by calibration according to a given distance measure. A disadvantage of using GREG is that negative weights can occur (Deville & Särndal, 1992).

In the later application we refrain from population weighting adjustments, since there are either no known population parameters (yet) available to adjust to or these are based on non-matching definitions.

Chapter 3

Sampling grade 5 and grade 9 students

Chapter outline:
The samples for students in grade 5 and 9 in secondary schools (Starting Cohorts 3 and 4) will be thoroughly described in this chapter. The populations of students in grade 5 and 9 are described using available information from the sampling frame. On the basis of this information the planning phase with simulations and their corresponding results are presented. The final description of the samples for grade 5 and grade 9 students is then associated by the derivation of design weights. The sampling design of NEPS can be summarized as a stratified multistage cluster sampling design. The selection scheme, that is, the rule of how to select units from the universe, for sampling PSUs is systematic selection with probability proportional to size and the SSUs are sampled using simple random sampling. The focus therefore will be on the particularities of each Starting Cohort.
This chapter is in parts based on the work published in earlier papers by Aßmann et al. (2011) and Aßmann et al. (2012).

3.1 Population

The target population of the NEPS SC3 and SC4 include all students attending primary or secondary schools in grade 5 or grade 9 within the Federal Republic of Germany in the school year 2010/11. Access to the population of students was gained via the corresponding set of schools. This set of schools includes all officially recognized and state-approved educational institutions within the Federal Republic of Germany providing schooling for students in grade 5 and / or grade 9. Excluded from the population were

students attending vocational schools or schools with a predominant foreign teaching language that would hinder the realization of a complete survey procedure with the test instruments available. Further, students attending regular schools being unable to follow normal testing procedures were excluded.[1] Additionally, the NEPS comprises a sample of students attending special schools with main emphasis on special educational needs in the area of learning. Access to this population was gained via special schools with Federal-State-specific provisions explicitly for students with special educational needs in the area of learning. Overall, 80% of students attending special schools have a diagnosed learning disability – constituting the largest group of students in these schools. For more details see also Aßmann et al. (2011).

Table 3.1 shows the population of regular schools by school type and schools having classes in grade 5 and 9. The population of schools consists in total of 29346 schools from which 16273 are of no interest for sampling students in grade 5 or 9 since they neither have any classes in grade 5 nor any in grade 9 (row: Neither grade). Although these schools can have classes in any other grade from 1 to 4 and from 6 to 8. The middle rows give the number of schools having only classes in grade 5 and none in grade 9 (row: Grade 5 only) and vice versa. That is the number of schools having classes only in grade 9 but none in grade 5 (row: Grade 9 only). Schools having only classes in grade 5 are in total 1459 and schools having only classes in grade 9 are 1281. The majority of schools relevant for sampling students in SC3 and SC4 consists of secondary schools having classes in grade 5 and grade 9 are in total 10333 (row: Grade 5 and 9).

Table 3.1: Population of regular schools by school type and schools providing classes in grades 5 and 9 (school year 2008/09).

	School type								
	GS	*HS*	*MB*	*RS*	*IG*	*GY*	*SU*	*FW*	$\sum$
Neither Grade	16109	17	72	14	4	44	4	9	16273
Grade 5 only	913	89	75	25	58	62	222	15	1459
Grade 9 only	0	447	163	346	84	230	0	11	1281
Grade 5 and 9	0	3656	1044	2211	541	2708	0	173	10333
$\sum$	17022	4209	1354	2596	687	3044	226	208	29346

Notes: Abbreviations of school types are *GS*: Grundschule, *HS*: Hauptschule, *MB*: Schule mit mehreren Bildungsgängen, *RS* Realschule, *IG*: Integrierte Gesamtschule, *GY*: Gymnasium, *SU*: Schulartunabhängige Orientierungsstufe and *FW*: Freie Waldorfschule.

[1] Regular schools are all *allgemeinbildende Schulen* according to the definition of Kultusministerkonferenz (2012); special schools (Förderschulen) excluded.

Later on the school types *Integrierte Gesamtschule* (*IG*) and *Freie Waldorfschule* (*FW*) will be joined in one stratum since the degrees achievable are similar. Further school types *Grundschule* (*GS*) and *Schulartunabhängige Orientierungsstufe* (*SU*) will be joined in one stratum because these schools educate students in grade 5. The school type *GS* summarizes primary schools normally educating students in grade 1 to 4. The 913 schools having grade 5 too are schools in Berlin and Brandenburg educating students from grade 1 to grade 6. School type *SU* educates students only in grade 5 and 6 in Hesse and Hamburg.

For the population of schools displayed in Table 3.1 the number of students within these schools are given in Table 3.2. The table is twofold since the number of students in grade 5 and 9 needs to be reported separately.That is because grade 5 students can be in schools providing either classes in grade 5 only or they can be in schools providing classes in grade 5 and 9. For example there are 2708 Gymnasia (*GY*) providing access to students in grade 5 and students in grade 9. These 2708 schools educate 294624 students in grade 5 (Table 3.2 upper half) and 253929 (Table 3.2 lower half) students in grade 9. Further there are 62 Gymnasia providing access to 3093 students in grade 5 only and 230 Gymnasia educating 14724 in grade 9 only.[2] In total there are 794317 students in schools providing at least one class in grade 5. The corresponding number of students in schools providing at least one class in grade 9 is 806964.

The zeros in the table are due to the fact that the group of schools providing access to classes in grade 5 and 9 do not have students in grade 5 for school types *GS* and *SU* (upper half of the table). These school types educate students only in grades from one to four (*GS*) and five and six (*SU*) respectively. So there are no classes in grade 9. In the lower half of the table the zeros arise from school types *GS* and *SU* not providing any classes in grade 9.

3.2 Summarizing sampling for school cohorts

The variety of Federal-State-specific school systems is challenging for sampling grade 5 and grade 9 students. Several school types related to different transitions between elementary and secondary school institutions form

[2]Suppose a Gymnasium has one class in grade 5 having 17 students and one class in grade 9 having 33 students. This school is reported together with the 2708 Gymnasia having classes in grade 5 as well as in grade 9 in Table 3.1. In Table 3.2 the 17 students in grade 5 are reported among the 294624 fifth grade students and the 33 ninth grade students are reported among the 253929 students in grade 9.

Table 3.2: Population of Students in grade 5 and 9 by school type and schools providing classes in grades 5 and 9 (school year 2008/09).

	Population of grade 5 students School type								
School provides	*GS*	*HS*	*MB*	*RS*	*IG*	*GY*	*SU*	*FW*	$\sum$
Grade 5 only	43374	2079	2791	653	1824	3093	13919	301	68034
Grade 5 and 9	0	120816	50303	186621	67301	294624	0	6619	726284
$\sum$	43374	122895	53094	187274	69125	297717	13919	6920	794317
	Population of grade 9 students School type								
School provides	*GS*	*HS*	*MB*	*RS*	*IG*	*GY*	*SU*	*FW*	$\sum$
Grade 9 only	0	15967	8251	19928	9672	14724	0	319	68861
Grade 5 and 9	0	162171	48470	198815	68576	253929	0	6142	738103
$\sum$	0	178138	56721	218743	78248	268653	0	6461	806964

Notes: Abbreviations of school types are *GS*: Grundschule, *HS*: Hauptschule, *MB*: Schule mit mehreren Bildungsgängen, *RS* Realschule, *IG*: Integrierte Gesamtschule, *GY*: Gymnasium, *SU*: Schulartunabhängige Orientierungsstufe and *FW*: Freie Waldorfschule.

the set of schools providing access to the target population of grade 5 and grade 9 students. To reflect this variety, seven explicit strata have been defined to sample schools. The first stratum comprises all Gymnasien (stratum *GY*: Gymnasien), the second stratum consists of all Hauptschulen (stratum *HS*: Hauptschulen), the third stratum refers to all Realschulen (stratum *RS*: Realschulen), the fourth to comprehensive schools (stratum *IG*: Integrierte Gesamtschulen, Freie Waldorfschulen), the fifth includes schools with several courses of education (stratum *MB*: Schulen mit mehreren Bildungsgängen). The sixth explicit stratum comprises schools offering schooling to students with special educational needs in the area of learning (stratum *FS*: Förderschule). The seventh explicit stratum comprises all schools providing schooling to grade 5 students, but not to grade 9 students (stratum *N*5). The definition of these seven explicit strata allows fulfilling two important aspects.

1. A requisite of NEPS is to establish a sample of grade 9 students as the starting point of a longitudinal survey of young adults entering vocational education over the coming years. In order to ensure sufficient sample sizes for statistical analyses within this heterogeneous population, who, to a large extent, come from Hauptschulen, Gesamtschulen, and Schulen mit mehreren Bildungsgängen, NEPS comprises an oversampling of grade 9 students attending these school types.

2. Most secondary schools offer schooling to grade 5 and grade 9 students, so that they can be reached via the same set of schools and, thus,

reducing administrative survey costs.

In addition to the explicit stratification according to school types, an implicit stratification, that is, sorting the frame by certain characteristics, based on Federal States, regional classification, and sponsorship was used. Given the first-stage sample of regular schools, on the second stage two school classes within each school were sampled randomly, if at least three classes were present, otherwise all classes were surveyed. In special schools a census for all students was held.

3.3 Planning samples for school cohorts

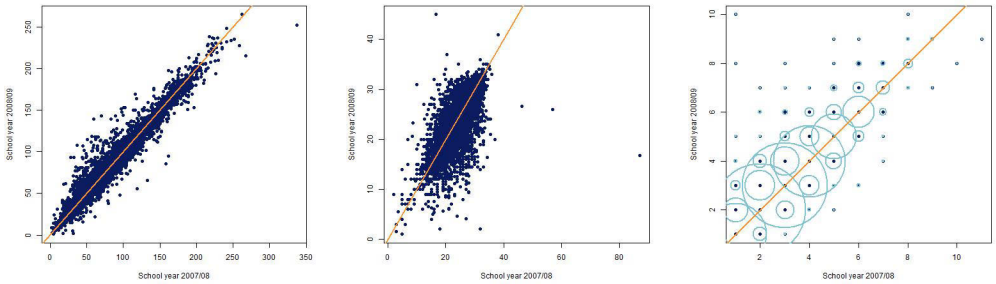
3.3.1 Sample design

For sampling of students in Germany a sampling frame containing a complete listing of all students is not available. In contrast a complete listing of schools providing access to students is available through the Statistical Offices. Access to the target population is gained via the corresponding institution, so that cluster sampling is appropriate. Furthermore the entire age group or only a part of it can be surveyed. When subsampling of age group within a school is performed this is referred to as two-stage (in more general multistage) sampling. One more aspect to consider is stratification since the landscape of school systems in Germany is heterogenous. The population of schools and students was stratified by the school type–more precisely the degree a student can achieve at the school. Lastly there are several selection schemes (i.e., the rules of how to select units from the universe) for sampling units in stratified or multistage designs including simple random sampling (with or without replacement), systematic selection and unequal probability designs. Because the context of the learning environment, that is classes, should be reflected in the data later on cluster sampling of a certain number classes was applied in regular schools. In special schools a census was preferred.

3.3.2 Determining the measure of size

As discussed in Subsection 2.2.2 sampling clusters of unequal size, for example schools or classes, leads to a random sample size on the level of students. One way of achieving an approximate control of sample size on the student level is probability proportional to size sampling. For *pps* sampling a measure of size is assigned to each unit. Using such a sampling design larger schools can be preferred over smaller by assigning them a larger measure of size.

This allows for reducing survey costs by sampling fewer schools to achieve an equal sample size compared to simple random sampling. On the school level characteristics such as total number of students per school, number of students per grade or number of classes per grade were directly available from the frame. The information provided by the frame from school year 2008/09 were used in sampling SC3 and SC4 in 2009. Since reporting information on schools and the preparation of these data by the Statistical Offices is time consuming this was the most actual sampling frame available. So the current situation in schools is not mirrored to the full extent by this frame. Students were surveyed and tested in school year 2010/11. To evaluate how much uncertainty in sample sizes and resulting weights is induced by the lag of time (i.e., two years) between sampling and surveying a simulation was set up. Therefore another frame from the year 2007/08 was made available so that it was possible to simulate several scenarios.



(a) Number of students. (b) Average class size. (c) Number of classes.

Figure 3.1: Changes from school year 2007/08 to 2008/09 for certain characteristics.

Figure 3.1 shows the changes for a one year difference for selected characteristics. The angle bisector indicates no changes from one year to another. The points above indicate an increase, whereas points below the angle bisector indicate a decrease in the corresponding characteristic. The number of students (Subfigure 3.1a) is varying from school year 2007/08 (x-axis) to school year 2008/09 (y-axis) with a covariance of $\varrho = 1725.3211$ and a correlation of $\rho = 0.9796$. The average class size (Subfigure 3.1b) is varying from school year 2007/08 (x-axis) to school year 2008/09 (y-axis) with a covariance of $\varrho = 19.0445$ and a correlation of $\rho = 0.7808$. The smallest covariance can be found for the number of classes changing only slightly from school year 2007/08 (x-axis) to school year 2008/09 (y-axis) with a covariance of $\varrho = 1.8597$ and a correlation of $\rho = 0.9332$. The combinations along the

angle bisector show those schools with no changes in the number of classes from one year to another. This case is most common. The radius of the circles around the combinations of x and y values in Subfigure 3.1c is proportional to the number of this specific combination of x and y values, that is, the larger the radius around (x,y) pairs the more often this (x,y) pair exists. From the subfigure it can be seen that the number of classes does not change in the majority of schools. It also can be seen that small changes from one school year to another are also common. These are the (x,y) pairs near the angle bisector. A large change in the number of classes can occur if schools are merged or if a school has several locations and is once reported with one location an in the next year with several locations. These schools are found on the top right part of the subfigure. Both have only one class in school year 2007/08 but have eight and ten in school year 2008/09 respectively.

For the following simulation the intersection of schools in the frames of the school years 2007/08 and 2008/09 was used, that is, the set of schools which is contained in the frames of both school years. Due to ongoing reforms, closing or merging of schools a small number of schools had to be discarded. For the remaining schools the primary sampling units were selected based on the frame of the school year 2007/08. Sampling of secondary sampling units was based on the frame of the school year 2008/09. Thus this simulation covers the variation induced by a time lag of one year.

For sampling schools a systematic probability proportional to size sampling was applied. Therefore a measure of size needed to be assigned to each unit. The objective of the simulation was to find a measure of size for which the resulting inclusion probabilities (and thus design weights) yield the least variation. For selection of schools the measures of size evaluated in the different scenarios were (among others):

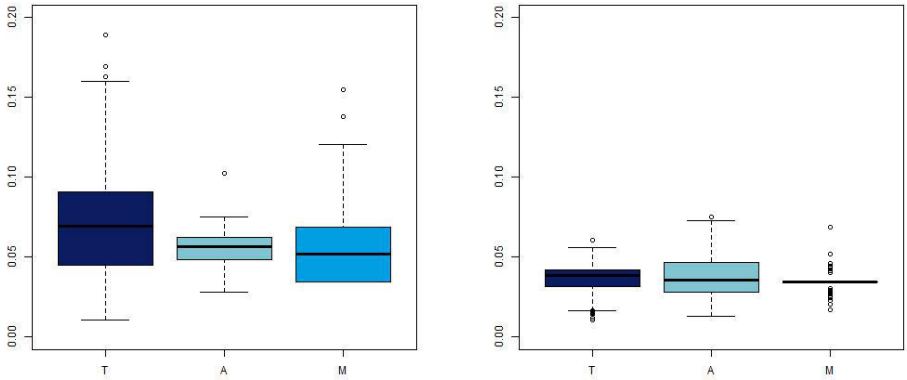
T the number of students in grade 9

A the average number of students per class in grade 9 (i.e. T/C)

M the minimum of the number of classes (C) in grade 9 and 2 classes (i.e., $\min\{C; 2\}$)

Scenario T uses the number of students in grade 9 as a measure of size which is highly correlated as can be seen from Subfigure 3.1a. The second scenario A uses the average class size as a measure of size. The average class size is also positively correlated but not as strong. Further the variation is higher. The last scenario M uses $\min\{C; 2\}$, that is the minimum of the existing number of classes and two classes. As shown in Subfigure 3.1c the number of classes does not vary strongly and is also highly correlated. Figure 3.2a

shows boxplots for the inclusion probabilities on the school level by scenario. Sampling proportional to the average number of students (A) yields least variation in design weights on the school level. In contrast scenario M yields least variation in design weights on the level of students (see Figure 3.2b). This result is due to the fact that within each school two classes out of at least three were sampled (otherwise all were chosen).



(a) School level. (b) Student level.

Figure 3.2: Inclusion probabilities for different scenarios.

When sampling two classes (if available) within each school, scenario M results in a self weighting sample, that is a sample where each unit is assigned the same design weight. This result holds if and only if the number of classes does not change within the time lag between sampling and surveying. That the number of classes do change between two school years can be seen from Figure 3.2b. The Outliers are schools in which the number of classes changed between school year 2007/08 and 2008/09.

The comparison of the two frames (see Figure 3.1) and the following simulation (see Figure 3.2) showed that the variation of the number of classes was much lower from one year to another than the variation of any student number related characteristic. Due to fewer changes in this characteristic a measure of size based on the number of classes induces less variation in design weights over time. Due to this result any sampling design based on the number of classes should be favoured in sampling proportional to size when sampling PSU and SSU fall apart in time over at least one school year.

3.3.3 Determining the first stage sample size

Since sampling is done on two stages containing unequal sized clusters in each cohort the resulting ultimate stages' sample size becomes a random variable (see Kish, 1995, p. 183). This makes sample sizes (on the student level) less predictable (see also the corresponding paragraph in Subsection 2.2.2 for a more detailed discussion) in two-stage cluster sampling with unequal cluster sizes. Besides this other topics should be kept in mind:

- Unit nonresponse, that is, sampled units are not willing to participate. This problem can possibly occur on several stages within multistage designs.
- A minimum net sample size may be desired and this is closely related to the topic before.
- A budget for a survey giving the financial restriction or in other words a maximum net sample size.

When thinking about these topics it is straightforward that there may be trade-offs between sample size and the budget. To overcome the trade-off between a random sample size and a fixed budget another simulation was set up.

To determine the number m^I of primary sampling units (PSU) that need to be sampled to realize a sample of an approximate size n the idea is as follows: Sample m^I PSUs on the first stage and then sample m^{II} secondary sampling units (SSU) on the second stage.[3] Given that the sizes of the PSUs and SSUs are either known from a frame (see Section 2.1) or can be estimated by an average size (for example an average class size) the initial sample size can be calculated. The number of targets participating in the survey can be calculated for a given value or range of a participation rate p. The resulting net sample size $n_{net} = n \cdot p$ can be translated into a number of test groups (t) or an interviewer field (i.e., the financial restriction or a budget). For a specific sample the net sample size (n_{net}, that is the number of targets participating) should be least equal to the desired minimum net sample size (n_0). At the same time the number of resulting test groups (t) should not exceed the number of financed test groups (t_0).

To calculate the net sample size and to allow for refusals different participation rates p are assumed. For a given value of the participation rate p the number of participating students n_{net} in school j is

$$n_{net,j} = \lfloor n_j \cdot p \rfloor.$$

[3]The corresponding stage is denoted by the superscripts.

Here $\lfloor \cdot \rfloor$ denotes the floor function that rounds downwards, that is, $\lfloor 5.9 \rfloor = 5$. Rounding the number downwards implicates a conservative measure of the number of students participating. Surveying and testing students is done in groups of size

$$\tau = \begin{cases} 30 & \text{if } h = GY \\ 25 & \text{else} \end{cases}$$

and so the number of test groups needed per school arises as

$$t_j = \lceil \frac{p \cdot n_j}{\tau} \rceil.$$

Here $\lceil \cdot \rceil$ denotes the ceiling function that rounds values upwards, that is, $\lceil 1.2 \rceil = 2$. Again the rounding (in this case upwards) indicates a conservative measure for the number of test groups. For a drawn sample with given m^I and p the net sample size is

$$n_{net} = \sum_{j=1}^{m^I} n_{net,j} = \sum_{j=1}^{m^I} \lfloor n_j \cdot p \rfloor$$

and the total number of test groups is

$$t = \sum_{j=1}^{m^I} t_j = \sum_{j=1}^{m^I} \lceil \frac{p \cdot n_j}{\tau} \rceil.$$

For given values of m^I (see Table 3.3 for their value and the allocation to different strata) and p (e.g., $p \in [0.4; 0.8]$) these steps are repeated $R = 1000$ times realizing a distribution for the net sample size and the number of test groups. The distributions obtained from the replications (see Table B.1 in Appendix B) can be evaluated for the different combinations of m^I and p with respect to the desired minimum net sample size and the budget in terms of a maximum number of test groups.

The combinations of the number of primary sampling units m^I and participation rate p, simultaneously fulfilling the conditions that (over R replications) the average number of students participating ($\mu_{n_{net}}$) is least equal to the desired minimum net sample size ($n_0 = 12500$) and – at the same time – the average number of resulting test groups (μ_t) does not exceed the maximum number of financed test groups ($t_0 = 840$), are referred to as *favorable samples* and shown in Table 3.4. This table therefore is just a simplified representation of Table B.1 in Appendix B.

Table 3.3: Allocation of first stage's sample sizes m^I.

m^I	m_h^I				
	HS	MB	RS	IG	GY
450	120	39	102	45	144
460	123	40	104	46	148
470	125	41	106	47	151
480	128	42	108	48	154
490	131	43	111	49	157
500	133	44	113	50	160

For example, sampling $m^I = 480$ schools on the first stage and two classes (if available) on the second stage yields $\mu_t = 782$ test groups (on average) for a participation rate of $p = 0.6$. At the same time this combination of m^I and p yields on average a number of participants $\mu_{n_{net}} = 12986$. So the simulation gives a basis for decision making when having a trade-off between a desired minimum sample size and a fixed budget. In our application we used the participation rates from pilot studies as reasonable estimates.

When taking institutional refusals into account predefined replacement schools were asked to participate instead of the originally sampled schools. When they decide to take part in the survey, estimates in the simulation change due to different numbers of classes and students. This topic can be taken into account within this simulation setting as well. Reasonable participation rates on the school level were not available at that time. These were first available after the first wave and were incorporated in later simulations serving as a basis for decision making in subsequent samples.

3.3.4 Replacing nonparticipating schools

Preventing institutional refusals is an important topic since the participation for schools as well as for students is not mandatory in the NEPS. In order to prevent from reduction of sample size and introduction of potential bias a replacement strategy was designed. This is supposed to compensate for nonparticipating institutions and the corresponding reduction of sample size on the student level. Thereby a sampled and nonparticipating institution is replaced by an institution that is most similar in structure with respect to explicit, implicit stratification (i.e., sorting the frame by certain characteristics) and sampling characteristics (i.e., school type and Federal State, sponsorship regional classification, measure of size). A similar approach is used

Table 3.4: Favorable samples.

m^I	p	μ_t	$\mu_{n_{net}}$
470	0.60	766	12724
480	0.60	782	12986
490	0.60	799	13284
500	0.60	815	13532
450	0.65	766	13194
460	0.65	784	13521
470	0.65	799	13784
480	0.65	816	14068
490	0.65	835	14391
450	0.70	788	14209
460	0.70	808	14562
470	0.70	823	14845
480	0.70	840	15151
450	0.75	803	15224
460	0.75	823	15602
470	0.75	838	15905
450	0.80	821	16239

in the Programme for International Student Assessment (PISA) as well as in Progress in International Reading Literacy Study (PIRLS) and in Trends in International Mathematics and Science Study (TIMSS), see OECD (2012), Martin, Mullis, and Kennedy (2007) and Olson, Martin, and Mullis (2008). This is especially important for schools, since there is a large variety of studies (mandatory and non mandatory) surveying students in institutional contexts which put additional workload on the schools' staff. The replacement strategy cannot rule out potential bias and therefore needs to be evaluated after the realization of the institutional sample. Few schools cannot be replaced because they were shut down or there were no institutions similar in structure available.

Baker et al. (2013) give a review on the topic non-probability sampling which provides some frameworks, including the adopted matching, that might serve as a basis to develop a more sophisticated framework for replacing nonparticipating schools.

3.4 Sampling for grade 9

The complete list, that is, the frame, of schools in the school year 2008/09 was used in sampling schools. It was by that time the most actual one available and via the sampled schools access was provided to grade 9 students in the school year of 2010/11. The population of regular schools was stratified into $H = 5$ strata. Thereby the stratum GY consists of *Gymnasien*, stratum HS of *Hauptschulen*, stratum IG includes *(Integrierte) Gesamtschulen* and *Freie Waldorfschulen*[4], stratum MB consists of *Schulen mit mehreren Bildungsgängen*, and stratum RS includes all *Realschulen*. The sample of regular schools was selected via systematic probability proportional to size sampling. The measure of size was chosen proportionally to the number of classes of seventh grade in 2008/09 as the best available proxy for the number of classes in the ninth grade two years later (see Subsection 3.3.2 or Aßmann et al. (2012)). Simulation studies as described earlier, found 629 out of a population of 11570 regular schools with seventh classes to be sufficient for providing intended sample sizes. Another stratum FS contained 1488 special schools with a focus on learning disabilities. For this stratum a separate frame was available containing information on schools also from the school year 2008/09, These were selected using systematic *pps* with measure of size proportional to the squared number of students reported in grade 9.[5] This measure of size allows handling the trade-off between sample size in this stratum and the number of special schools that need to be sampled by sampling large schools with higher probabilities. Let M_h^9 denote the total number of schools in stratum h and m_h^9 the number of schools sampled in stratum h, where $h \in \{GY, HS, IG, MB, RS, FS\}$. The measure of size for sampling a school j in stratum h is then defined as

$$mos_{jh}^9 = \begin{cases} \frac{C_{jh}^7}{\min\{C_{jh}^7;2\}}, & \text{if } h \in \{GY, HS, IG, MB, RS\}, \\ (S_j^9)^2, & \text{if } h \in \{FS\}. \end{cases} \tag{3.1}$$

C_{jh}^7 denotes the number of classes in grade 7 in school j in stratum h in the school year of 2008/09. S_j^9 denotes the approximated number of students

[4]These two school types were put together in one stratum because the achievable degree is similar.

[5]In most special schools students in grade 7, 8, and 9 attend the same courses. That is, in the majority of cases no number of students in grade 9 can be reported. Instead the total number of students in grades 7 to 9 is reported. Therefore, the number of grade 9 students is approximated by one third of the reported number of students in grades 7 to 9.

attending grade 9 in special schools. The strata-specific total measure of size is

$$MOS_h^9 = \sum_{jh}^{M_h^9} mos_{jh}^9, \text{ for } h \in \{GY, HS, IG, MB, RS, FS\}. \tag{3.2}$$

For each stratum the considered values of M_h^9, m_h^9, mos_h^9, and MOS_h^9 are given in Table 3.5 in the following section. In special schools all students were asked to participate, that is a census took place. In regular schools a subsample of classes was drawn. That is, if in a school j in stratum h the number of classes in grade 9 C_{jh}^9 in school year 2010/11 was larger than two, a random sample of two classes in grade 9 C_{jh}^9 was drawn. If the number of classes in grade 9 was two or less all available classes were selected. This selection scheme results in the following inclusion probabilities π_{ijh}^9 for a student i in secondary school j in stratum h

$$\pi_{ijh}^9 = \begin{cases} m_h^9 \cdot \frac{mos_{jh}^9}{MOS_h^9} \cdot \frac{\min\{C_{jh}^9;2\}}{C_{jh}^9}, & \text{if } h \in \{GY, HS, IG, MB, RS\}, \\ m_h^9 \cdot \frac{(S_j^9)^2}{MOS_h^9}, & \text{if } h \in \{FS\}. \end{cases} \tag{3.3}$$

When the number of classes in grade 7 used in constructing the measure of size mos_{jh}^9 is the same number of classes in grade 9 two years later this would lead to a self weighting sample within the strata of regular schools that were not oversampled. Self weighting sample means that each sampled unit is assigned the same design weight. The inclusion probability (and thus the design weight) would only depend on the number of schools sampled and the total measure of size MOS_h^9.

3.5 Sampling for grade 5

In Germany there is a Federal-State-specific timing of transition from primary to secondary education. Most Federal States offer four years of primary schooling followed by secondary schooling. Other States offer six years of primary education and start secondary schools afterwards. Lastly few states have a four year primary school, followed by two years of transition system and start secondary education afterwards (see also Section 3.1). The sample of grade 5 students accounts for this Federal-State-specific timing via explicit stratification. It is based on seven explicit strata. Here the strata are the same as for the grade 9 sample of schools, namely GY, HS, IG, MB, RS, and FS. The stratum $N5$ comprises schools that are mainly elementary schools

educating students in grades 1 to 6 in Berlin and Brandenburg as well as schools in Hesse and Hamburg educating students in grades 5 and 6 only (so called *Schulartunabhängige Orientierungsstufen*). To ensure little variation in design weights across the considered strata, the number of schools that were sampled per stratum was, again, calibrated via simulation studies. To reach the intended sample sizes for grade 5 students 240 regular schools and 65 special schools were found to be sufficient. The sample of grade 5 schools for the strata GY, HS, RS, IG, MB, and FS was established as a subsample of the realized grade 9 school sample to reduce administrative survey costs. In stratum FS, subsampling of schools was performed via simple random sampling (srs). A srs design was found to be sufficient because the number of students in grade 9 was positively correlated with the number of grade 5 students.

Table 3.5: Population sizes (M_h^5 and M_h^9), sample sizes (m_h^5 and m_h^9), and total measures of size (MOS_h^5 and MOS_h^9) for schools with classes in grade 5 and 9 by strata ($h \in \{N5, GY, HS, IG, MB, RS, FS\}$).

	SC4			SC3		
Stratum	m_h^9	M_h^9	MOS_h^9	m_h^5	M_h^5	MOS_h^5
N5	-	-	0.000	26	1383	1914.500
GY	154	2970	5589.500			
HS	233	3990	4641.500			
IG	70	822	1727.500	214	480*	559.060
MB	64	1288	1524.000			
RS	108	2500	3936.500			
FS	110	1488	489128.264	65	110	-

Note: * For $h \in \{GY, HS, IG, MB, RS\}$ the population of 480 schools with fifth classes is a subsample of the realized grade 9 school sample.

Schools in stratum $N5$ contained in the frame and referring to school year 2008/09 were selected randomly using systematic *pps* sampling. Let M_h^5 denote the stratum-specific total number of schools with classes in grade 5 considered for sampling and m_h^5 denotes the stratum-specific number of schools in the grade 5 sample. The measure of size mos_{jh}^5 in stratum $N5$, for sampling a school j is computed analogously to the measure of size for a comparable regular school in the grade 9 sample. It is

$$mos_{jh}^5 = \frac{C_{jh}^5}{\min\{C_{jh}^5; 2\}}, \quad \text{if } h \in \{N5\}$$

where C^5_{jh} is the number of classes in grade 5 hosted by school j in stratum $N5$ in school year 2008/09. For a regular school j in stratum h the measure of size is

$$mos^5_{jh} = \frac{\frac{C^5_{jh}}{\min\{C^5_{jh};2\}}}{\frac{C^7_{jh}}{\min\{C^7_{jh};2\}}}, \quad \text{if } h \in \{GY, HS, IG, MB, RS\}$$

where C^5_{jh} is the number of classes in grade 5 and C^7_{jh} the number of classes in grade 7 in school j in stratum h in school year 2008/09. The corresponding stratum-specific total measure of size is

$$MOS^5_h = \sum_{jh}^{M^5_h} mos^5_{jh}, \quad \text{for } h \in \{GY, HS, IG, MB, RS, N5\}.$$

For the grade 5 sample, the strata-specific values of M^5_h, m^5_h, mos^5_h, and MOS^5_h are also shown in Table 3.5. Sampling grade 5 students in strata $h \in \{GY, HS, IG, MB, RS, N5\}$ was performed similarly to sampling students in regular schools for the grade 9 sample: If at least three classes in grade 5 were available in a sampled school in the school year of 2010/11, a random sample of two classes in grade 5 was drawn; otherwise all classes were selected. Finally, the inclusion probability π^5_{ijh} for a grade 5 student i in school j in stratum h is

$$\pi^5_{ijh} = \begin{cases} m^5_h \cdot \frac{mos^5_{jh}}{MOS^5_h} \cdot \frac{\min\{C^5_{jh};2\}}{C^5_{jh}}, & \text{if } h \in \{N5\} \\ \pi^9_{jh} \cdot \frac{m^5_h}{m^9_h}, & \text{if } h \in \{FS\} \\ \pi^9_{jh} \cdot m^5_h \cdot \frac{mos^5_{jh}}{MOS^5_h} \cdot \frac{\min\{C^5_{jh};2\}}{C^5_{jh}}, & \text{if } h \in \{GY, HS, IG, MB, RS\}. \end{cases} \tag{3.4}$$

Here π^9_{jh} denotes the probability that a school j in stratum h is part of the grade 9 sample, $h \in \{GY, HS, IG, MB, RS, FS\}$. Again, if the number of classes in school year 2008/09 used for the measure of size is the same as the actual number of classes in grade 5 in school year 2010/11 this design would yield a self weighting sample for strata $h \in \{N5, GY, HS, IG, MB, RS\}$. The corresponding design weight of a grade 5 student i in school j in stratum h is then the inverse of the inclusion probability: $d_{ijh} = 1/\pi^5_{ijh}$.

The sample of grade 5 students was enriched by a supplement of 214 students with a Turkish migration background or a migration background

related to the former Soviet Union. As no frame information was available that allowed for direct identification of the relevant groups, a two-stage selection was applied. On the first stage, 500 schools from 10 groups of Federal States (50 schools per group) were sampled systematically proportional to the number of students with foreign citizenship attending grade 5. Foreign citizenship was available for all schools in the actual frame. For Hamburg a frame was available that further provided information on students with a migration background. In finding a good predictor for schools with a large number of students with a migration background the information on foreign citizenship outperformed the others.

For this sample of 500 schools, the educational ministries of the Federal States were asked to quantify the number of students related to the two migration backgrounds of interest. This quantification took different forms based on the amount of information available in the Federal-State-specific school statistics. Some Federal States reported the number of students with migration background directly, other Federal States provided a ranking of schools attended by the highest number of students with the two migration backgrounds. With this informational content, a five categorical ordinal scale for the measure of size was defined for a probability proportional to size sampling of schools with a high number of Turkish migrants as well as those schools with students with a migration background from the former Soviet Union. Within the sampled schools all students with the Turkish migration background or a migration background related to the former Soviet Union were asked to participate in the NEPS.

Chapter 4

Weighting adjustments

Chapter outline:
The previous chapter derived the 'pure' design weights for grade 5 and grade 9 students. Section 2.3 already pointed out the need for further adjustments. In multistage samples decisions to participate take place on multiple stages. Therefore weighting adjustments have to be considered on each stage. In case of the NEPS this is due to nonparticipation on the institutional as well as on the individual level. Thus multiple adjustments of the design weights are necessary. This chapter considers the weighting adjustments on the institutional level as well as on the individual level. The adjustments on the individual level include adjusting the initial sample to the panel cohort and afterwards additional adjustment to wave specific participation patterns.

4.1 Decision processes involved

Because participation is neither mandatory for schools nor for students, both can refuse to participate. The actions and decisions that are relevant to form a sample out of the population are stylized in Figure 4.1. Starting from a defined target population to the corresponding panel cohorts, that is, the part of the initial sample participating in the panel, there are three steps (or four stages).

The first step is drawing a sample of schools from the defined target population using the sample design described in the previous chapter. The recruitment of sampled schools started in April 2010 and ended in October the same year. During this time each originally sampled school was invited to participate in the NEPS. For those schools not willing to participate, the first replacement school was invited. If the first replacement school was also unwilling to participate the second replacement school was asked to partici-

pate and so on. In case of nonparticipation of the originally sampled school and all available replacement schools there was no further recruitment possible. For later adjustments of design weights on the school level information was available from the sampling frame as well as information arising from the recruitment process. This information was used to model the schools decision to participate. So the decision to participate on the school level is the first to be kept in mind for later adjustments. Step two is sampling two

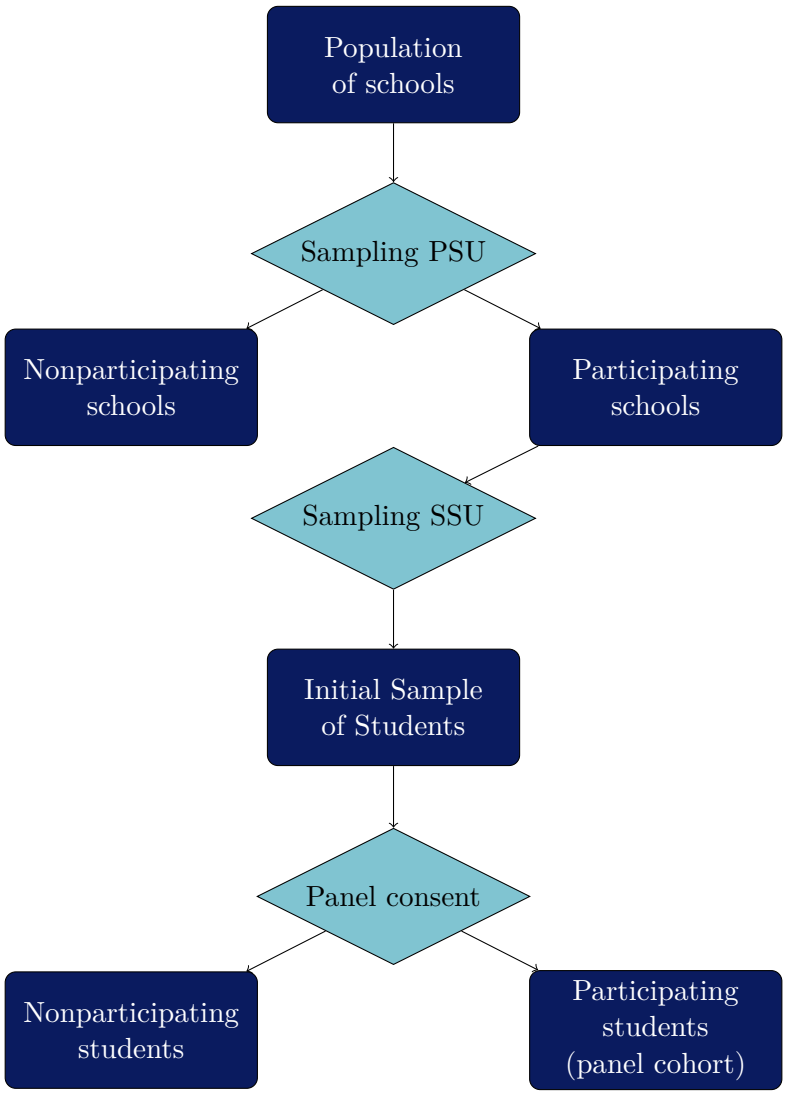


Figure 4.1: Flowchart of decision processes ranging from the population to the panel cohort.

classes (if available) within each participating school. In each school participating sampling of two classes (if available) was performed without refusal of classes. This sample of classes than yields the initial sample of students. The teacher responsible for correspondence with NEPS and the survey research institute was provided a students participation list with three parts (left, middle, right, see Table B.2 in Appendix B) to fill in information on the students in the sampled classes. The information was provided regardless of the students' participation decision. That is, information is provided on respondents as well as on nonrespondents. On the left part (remaining in the school) there was information on the school name and identifier, the sampled classes name and identifier as well as the students names and identifier. The middle part (returned to the survey research institute and later on to NEPS) contained information on the students, such as nationality (German, Non-German), language spoken at home (German, other), grades in Maths and German, special educational needs, dyslexia, gender, month and year of birth and participation status. Further information was provided on school type, sampled age group and Federal State. The right part (handed out to the interviewer on the test day) contains information on the participation decision of the student (if necessary on the parental consent), gender, month and year of birth, grades in Maths and German, information relevant for testing and the identifier of the student, the school, the class and the test group.[1]

In the last step all sampled students again have to choose to participate or not. This is the second decision process that has to be regarded for sample weighting adjustments. The information provided on the middle part of the students participation list on the initial sample, that is, responding and nonresponding students, form the set of information available for modeling the students' decisions to participate. Those students finally deciding to participate in the panel form the so called panel cohort.

The process of school recruitment and subsequently the student recruitment process is consecutive by nature and thus reflected by sequential modeling. That is, first school nonresponse is modeled. Second, the panel cohort sample is established on the basis of the active consent to participate in the panel. This consent is provided by parents, since a student is possibly not of legal age. Such is the case for a grade 5 student. Otherwise the students consent is sufficient (for example students in grade 9 that are of legal age). After correcting for unit nonresponse on the school and student level each student of the panel cohort is assigned an adjusted design weight; the panel

[1]Test groups were of size 25 students each (30 in Gymnasien). If one (or two classes) were larger than this they were split up into two (or more) test groups.

entry weight. On the institutional level nonresponse adjustments take sampling information as well as information from the recruitment process into account. On the individual level we take clustering on school level into account by specification of random intercepts to account for correlation within schools.

Third, given the panel consent provided, actual participation in each successive wave of the panel survey needs again to be analyzed. The decision processes leading to actual participation within the following waves are hence modelled subsequently on the basis of the panel cohort. In contrast to the adjustments for the initial sample we have further information available arising from surveying and testing students in wave 1.

Section 4.3 will give further insights on the adjustments performed on the school level and on the student level of the initial sample. The students participating in the panel (further referred to as panel cohort or panel cohort sample) then can decide to participate in each wave again. These wave-specific adjustments are based on the available information on the panel cohort, so that the user can reproduce and modify the weighting adjustment provided with the scientific use files (or other versions of the data).

4.2 Frameworks for decision modeling

The literature on modeling participation decisions stems to a large extend from household surveys. Some of the methods applied there are directly applicable to educational surveys, for example modeling decisions using statistical frameworks, such as (binary or multinomial) logit or probit regressions. Also respecting cluster structures such as a household can directly be transferred to educational surveys where students are clustered in schools and classes. In educational surveys focusing on schools some of the methods need to be adjusted.

Lepkowski and Couper (2002) separate the process ending in participation of sampled units into three subprocesses. Locating of sampled units is the first field operation. Without having located sampled units there is no way to contact them. Establishing contact with sampled units (given location) is the second step considered. Lastly units decide to cooperate or not given contact and location. This three step process applies to the different stages in educational surveys, too. Locating students by gaining access via the corresponding schools makes location easier, since in Germany schools are registered at the Ministries of Education. This also eases the second step since the Ministries provide contact information on the schools. What remains are the decisions to participate (i.e., cooperation) on the school and

student level. The decision processes involved from a sample of schools to the final realized panel cohort involve aspects not covered by household survey methods. Also the sequential processes of location, contact and cooperation as described by Lepkowski and Couper (2002) and applied in Iannacchione (2003) can either be directly applied or transferred to educational surveys. In a household survey one person living in this household might decide whether or not the entire (or parts of the) household participates in the survey. In contrast, the schools' decision to participate in the survey is made by the schools head (and probably other teaching staff of the school) so that the decision on the higher level is not made by the target persons of the survey (i.e., the students). Further, adopting a multi-informant perspective[2] in educational surveys to enrich data on the students' environments (class, school, home, etc.) opens a wide field of interesting research questions focusing on subgroups (for example students and parents).

Re-approaching panel members in the second wave of a survey is different from the first contact. The differences arise from panel members moving out of household. Getting into contact with them again might make locating them first necessary. Factors influencing the mobility in household surveys are mostly related to household or individual life stage and cohort effect (Lepkowski & Couper, 2002). In case of students this is more important to grade 9 students than to grade 5 students, because they leave school to enter the vocational track or change school to achieve a university entrance certificate. In grade 5 the necessity to re-locate students will become more important in the case of schools closing or age groups expiring within a school.[3]

In educational studies such TIMSS and PIRLS conducted by the International Association for the Evaluation of Educational Achievement (IEA, see Olson et al. (2008) and Martin et al. (2007)) as well as in PISA done by the OECD (see Rust, Krawchuk, & Monseur, 2013, pp. 87f) the decision processes in multistage sample designs are reflected by sequential modeling.

In weighting adjustments for household surveys addressing the differences between noncontact and refusal is stressed by various authors, such as Groves (1998), Durrant and Steele (2009), and Steele and Durrant (2011). Distinguishing between non-contactability and refusal allows for consideration of differences in characteristics determining the two components of nonresponse. To analyse the determinants of these response processes several statistical

[2]Different informants such as the target person, parents, teachers, etc. are asked to provide information on the same topic (for example occupational status of the students parents).

[3]An age group expires for example if too few students are within an age group and are thus handed over to a school nearby.

frameworks are at hand.

Recent work analyzes the determinants of this three categorical response process using sequential univariate models, bivariate sample selection models, multinomial models and their extensions to multilevel models (Durrant and Steele (2009), O'Muircheartaigh and Campanelli (1999) or Steele and Durrant (2011)).

O'Muircheartaigh and Campanelli (1999) apply multilevel logistic regression and multilevel multinomial regression to investigate the influence of the interviewer over those of a geographic region on household nonresponse in the British Household Panel Study (BHPS). Their findings indicate that good interviewers reduce refusal as well as noncontacts, since variance is induced by differences between interviewers rather than between geographic regions.

Durrant and Steele (2009) use the 2001 UK Census Link Study incorporating response outcomes for six household surveys.[4] They apply multilevel multinomial models to explore the effects of household characteristics on noncontact and refusal, as well. The multilevel structure allows for correlation in response probabilities for households allocated to the same interviewer. Their findings, according to the interviewer effects, are in line with those of O'Muircheartaigh and Campanelli (1999). Further results indicate that noncontact is related to household and lifestyle characteristics, that is, variables related to the propensity of being at home. In contrast, refusal is found to be a "complex social phenomenon that is explained by individual characteristics[...]" (Durrant & Steele, 2009, p. 378).

Steele and Durrant (2011) focus on alternatives in modeling noncontact, refusal and cooperation. They review sequential models, sample selection models and their extensions with a random effect and multinomial models, too. The authors find the sequential model (modeling contact first and refusal second) to be the most commonly used although sometimes only one of the two is estimated. The sequential modeling approach is also appealing since it separates the process of contact and participation, assuming independence of noncontact and nonparticipation. Besides that coefficients are more easy to interpret than in the multinomial model; although Steele and Durrant (2011) find the coefficients to be very similar. Furthermore they apply sample selection models allowing for residual correlation between the equations for noncontact and refusal.[5] Throughout their paper Steele and Durrant choose a probit link function in their analysis. For easing computational burden

[4]These are: Expenditure and Food Survey (EFS), the Family Resources Survey (FRS), the General Household Survey (GHS), the Omnibus Survey, the National Travel Survey (NTS) and the Labour Force Survey (LFS).

[5]With zero correlation the sample selection modeling approach would decompose into the sequential modeling approach.

a logit link is used in their simulation study. They find little difference in estimates using the multinomial and the sequential model. This fact is due to different sets of variables significantly effecting noncontact and refusal.

Also Nicoletti and Peracchi (2005) use a bivariate probit model to account for possible correlations between the ease of contact and the willingness to participate. After controlling for personal and household characteristics on the one hand and data collection characteristics on the other hand they find no residual correlation.

Skinner and D'Arrigo (2011) point out that nonresponse is commonly correlated within clusters since the access to the sampled targets is depending on authorities at the cluster level (as is in educational surveys within schools). This is mostly the case when multistage sampling designs are applied. Yuan and Little (2007) state that using random effects models in adjusting for unit nonresponse can yield biased estimates if the cluster-specific response rates vary across clusters. Skinner and D'Arrigo (2011) further show via simulation that for small cluster sizes a little negative relative bias is induced on the inverse probability weighted estimator.

Besides individual characteristics, household composition, social environment or survey design features most of these models use para data (Couper, 1998; Groves & Heeringa, 2006) in modeling response processes. For example these models incorporate interviewer characteristics (O'Muircheartaigh & Campanelli, 1999), level of effort measures (Biemer, Chen, & Wang, 2013) or the number of (failed) contact attempts (Wood, White, & Hotopf, 2006).

In the later applications probit models are used to model participation decisions. We decide for the probit link function to be consistent with later extensions of the probit model. This is because one extension to a bivariate binary probit model with random intercept will allow for estimation of a correlation parameter, which is not possible within a logit framework. The univariate probit framework describing the participation decisions for individuals within clusters can be described as follows. Let $j = 1, \ldots, m$ denote the cluster indicator and $i = 1, \ldots, n_j$ the indicator for individuals within clusters. The dichotomous participation decision y_{ij} can then be modelled using the probit framework given as

$$y_{ij} = \begin{cases} 1 & \text{if} \quad \widetilde{y}_{ij} > 0, \\ 0 & \text{else} \end{cases} \qquad \text{with} \quad \widetilde{y}_{ij} = X_{ij}\beta + \alpha_j + \epsilon_{ij}, \tag{4.1}$$

where $\widetilde{y}_{ij}$ denotes the latent variable, X_{ij} the regressors, β the coefficients, $\epsilon_{ij} \sim N(0, 1)$ denotes the disturbance, and $\alpha_j \sim N(0, \omega)$ denotes the random intercept. Note that the standard probit framework occurs if α_j is not taken into account. Both frameworks are implemented in standard statistical soft-

ware packages.[6] We base all adjustments on the individual level on response propensity reweighting, harking back to Rosenbaum and Rubin (1983). In unit nonresponse adjustments using auxiliary information the set of variables is often small, since information on nonrespondents is sparse. When modeling nonresponse the available variables should be good predictors for nonresponse to adjust the weights so that the nonresponse bias of the estimate is reduced. In a recent comment Little (2013) states that "attempting to address unit nonresponse without modeling the outcome is [...] like trying to tie a shoelace with one hand behind one's back." (Little, 2013, p. 363). This statement refers to the fact that weighting adjustments become most effective, that is, reduce nonresponse bias without increasing variance, when the variables used in adjusting the weights are also predictive for the variable of interest, as demonstrated by Little and Vartivarian (2003, 2005).

As noted by Kreuter et al. (2010) and Kreuter and Olson (2011) the selection of variables in nonresponse adjustments faces on one hand the problem of sparse information on nonrespondents and on the other hand the only few (if any) variables that are related to the response propensity and the key outcome variables. For example Rust et al. (2013) report that for student level adjustments in PISA information is limited to school, gender, month of birth and grade.[7] In TIMSS adjustment factors are mostly derived by dividing the number of sampled units by the number of responding units (Joncas, 2008). That is, unconditionally re-weighting by the inverse of the participation rate. Hawkes and Plewis (2006) use age, birth weight, sex, a dummy for being born in Wales and a dummy for mother stayed at school for modeling unit-nonresponse in a panel setting of the National Child Development Study (NCDS). Within the National Assessment of Educational Progress (NAEP) cell weighting is applied. The cells are formed by PSU, age, grade (modal or higher vs. lower than modal), see Rust and Johnson (1992). The NEPS has information on the students, the classes and the institutions provided by teachers that help to organize the test procedures in school. Besides that there is para data (Couper, 1998; Groves & Heeringa, 2006) available arising from test and telephone interview protocols during field work.

[6]The probit model implementation in `R` is given by the function `glm(formula, family = binomial(link = 'probit'), ...)` included in the `stats` package (R Core Team, 2014). The extension to the multilevel probit model by `glmer(formula, data, family = binomial(link = 'probit'), ...)` provided by the `lme4` package (Bates, Maechler, & Bolker, 2012).

[7]Since PISA focuses on students at the age of 15 years there is no need for information on the year of birth.

4.3 Adjusting design weights for nonresponse

Lugtig (21.10.2013) states that "[...] in survey methods, nonresponse is one of the phenomena that is contextual. Nonresponse always occurs, but the predictors of nonresponse differ across countries, survey topics, time, survey mode, and subpopulations. In other words, that is what makes building a theory about nonresponse so difficult." Due to this fact nonresponse adjustments have to account for particularities of the various contexts. For this reason the adjustments performed on the design weights are separated by the stages of the sampling design, that is schools as clusters of students and the students themselves. This is in line with methods commonly used in educational studies such as TIMSS (Olson et al., 2008), PIRLS (Martin et al., 2007), PISA (OECD, 2012; Rust et al., 2013) or the National Survey of Student Engagement (NSSE, Pike (2008)). These surveys also apply an unconditional cell weighting on the student level and adjust their weights by the inverse of the participation rate within each cell. In the context of educational surveys Porter and Whitcomb (2005) use an ordered logistic regression model because in the Cooperative Institutional Research Program (CIRP) first-year students had up to four opportunities to participate in the survey. We model school- and student-level participation processes separately. On both levels we take clustering into account. Schools are clustered by federal states because in Germany they are in charge of providing access to the schools. Further, the clustering by federal states accounts for different levels of effort in recruitment of the schools as well as for support by the different federal states. In contrast to the studies mentioned above we account for clustering on the school level via random intercepts when modeling individual participation decisions.

4.3.1 Adjusting for nonparticipation on the institutional level

When analyzing school participation, the replacement rule (see Subsection 3.3.4) implemented to address school nonresponse in advance has to be considered, see also Aßmann et al. (2011). Most often, schools refused participation in order to avoid additional workload that arises from participation in other studies. To counteract the resulting sample size reduction on the level of students as strongly as possible, replacement schools were defined in advance. However, the implemented replacement strategy was unable to prevent the nonparticipation of schools completely.

Table 4.1 shows the sampled and realized schools per stratum. The rates of schools refusing participation by strata varies from 1.81% in stratum FS

Table 4.1: Sampled vs. realized regular and special schools after replacement.

Sampled	Realized								
	Refused	FS	GY	HS	IG	MB	$N5$	RS	$\sum$
FS	2	108	0	0	0	0	0	0	110
GY	5	0	149	0	0	0	0	0	154
HS	52	0	0	181	0	0	0	0	233
IG	15	0	0	0	55	0	0	0	70
MB	8	0	0	0	0	56	0	0	64
$N5$	5	0	0	0	0	0	21	0	26
RS	4	0	0	0	0	0	0	104	108
$\sum$	91	108	149	181	55	56	21	104	765

Notes: Table is based on process information from the school recruitment.

to 22.75% in stratum HS. Participation rates in the other strata are 15.38% in stratum $N5$, 3.25% for Gymnasien, 21.43% in stratum IG which is almost as high as in HS. Stratum MB has a refusal rate of 12.5% and for stratum RS it is 3.7%. The overall refusal rate for schools is 11.9%. In summary the replacement strategy seemed not to be able to compensate refusals of schools to the full extend.

We checked via probit regressions (see Equation (4.1)) whether the available variables influenced school participation. To model school participation of all the contacted schools, their participation status was regressed on explaining factors reflecting the response burden for schools, the efforts involved in recruiting schools, and all variables defining the considered explicit stratum and implicit strata. The effort of recruiting schools is measured by the number of schools contacted per Federal State. A corresponding dummy variable separates the efforts by the median. As the legal basis for school participation differs across Federal States, a Federal-State-specific random intercept was considered. For schools not related to the migrants' supplement, the estimated random intercept models are shown in Table B.3 in Appendix B. The results indicate the significance of variables defining the explicit and implicit strata.[8] The results of the regression model describing the participation propensity of schools contacted during the supplement of migrants are given in Table B.4 in Appendix B.

For the model estimating the participation of special schools the dummy

[8]Note that some variables are stratum-specific and can only be considered in certain models, for example number of classes in grade 7 cannot be considered in stratum $N5$. Besides, collinearity within the $N5$ stratum does not allow for the consideration of the variable effort made in recruitment in the associated stratum-specific model.

for the effort measurement is significant. For the schools sampled in the supplement of students with migrational background this dummy could not be formed. Furthermore, here the state-specific random intercept is not significant. For each of the strata models for the regular schools (except for strata HS and RS) as well as for special schools the effort made in school recruitment (if in the model) is significant. In the case of Hauptschulen ($h = HS$), the dummy indicating public funding is significant. To adjust the design weights on the school level cell weighting is applied. The cells are formed by strata, Federal States and sponsorship. The weight of school j in weighting cell f was adjusted by the factor

$$\delta_j = \frac{m_f^r + m_f^n}{m_f^r}, \quad \text{for } j \in f$$

where m_f^r denotes the number of participating schools and m_f^n the number of nonparticipating schools in weighting cell f. The nonresponse adjusted design weight for school j in adjustment cell f then finally arises as

$$w_j = d_j \cdot \delta_j = d_j \cdot \frac{m_f^r + m_f^n}{m_f^r} \quad \text{for } j \in f.$$

The adjustment on the school level had to be done only once at the start of the panel because the participation of schools is the necessary condition for getting access to the students on the second stage.

The panel entry weight on the school level resulting from this adjustments can be found in the scientific use files of SC3 and SC4 under `w_i`. This denotes the weight for the institutions, which is corrected for institutional nonresponse.

4.3.2 Adjusting for nonparticipation on the individual level

As we gain access to the sampled targets via the participating schools, the decision to participate in the survey is made at the student level. At the time of the survey most students were not of legal age, thus, they needed the permission of their parents to participate in the panel survey. As panel consent from students and their parents was obtained before the actual survey, nonparticipation is a two-phase process. First, students and (if underage) also their parents, agree to participate in the panel. Second, nonresponse may occur in the actual wave due to temporary drop-out. Within NEPS a panel cohort member not participating is considered as a temporary drop-out. After two years without any information about the target she or he

is considered as final drop-out. Considering the decision to participate in the panel cohort Table 4.2 shows the percentages of students willing to participate per test group.[9] The table shows that in each starting cohort and

Table 4.2: Distribution of participation rates per test group by starting cohort.

		p_{Min}	$p_{0.25}$	$p_{0.5}$	μ_p	$p_{0.75}$	p_{Max}
SC3	REG	0	0.410	0.542	0.543	0.688	1
	FOE	0	0.423	0.636	0.601	0.800	1
	MIG	0	0.142	0.229	0.300	0.431	1
SC4	REG	0	0.463	0.667	0.627	0.826	1
	FOE	0	0.392	0.571	0.567	0.750	1

Note: REG: regular schools, FOE: special schools, MIG: migrants supplement.

each subgroup there were test groups in which no student was willing to participate (i.e., the minimum $p_{Min} = 0$ in all groups). On the other hand there were also test groups throughout all groups in which all students were willing to participate (i.e., the maximum $p_{Max} = 1$ in all groups). On average (μ_p) the participation rate per test group is highest in regular schools of SC4 (66.7%), followed by special schools in SC3 (63.6%). The lowest average participation rate per test group is found within the supplement of migrants. The participation rate of 30% on average shows the challenging task of recruiting hard to reach respondents and minorities. These findings are in line with the literature showing that active parental consent lowers participation rates towards a range of 40% to 60%, see Esbensen, Miller, Taylor, He, and Freng (1999).

More detailed pictures are given by the Tables 4.3 and 4.4. They show the participation status of the initial samples by strata. In both tables the minority of the initial sample has a undetermined participation status, that is at the day the students participation list was returned to the survey research institute the students did not hand back a informed consent to participate. Differences to the number of students in the panel cohort are due to this group. Students handing back a consent form after the students participation list was sent back to the survey research institute the student was allowed to participate.[10] Also students willing to participate revised their decision.

[9]To bring it to the readers mind again: A test group is a group of students tested together. These groups are mostly neither equal to classes nor to schools. The only case in which a test group is identical to school and class is when the school has only one class with less students then the maximum size of the test group.

[10]Unfortunately the participation status could not be updated since the returned part

For some returned students participation lists the school type needed to be anonymized (column NA), so that re-identification is not possible. This is due to the fact that information on nonrespondents come under the law of informational self-determination, derived from Article 2 (1) together with Article 1 of the Basic Constitutional Law of the Federal Republic of Germany (GG: Grundgesetz). It further appears in German Federal Data Protection Act (BDSG: Bundesdatenschutzgesetz) as well as in Data Protection Acts in each Federal State. For more detailed information on data protection issues in the NEPS see Meixner, Schiller, Maurice, and Engelhardt-Wölfler (2011). In SC3 the participation rates by strata range from 29.50% (stratum MIG) to 58.32% (stratum GY). In most strata ($h \in \{N5; HS; IG; MB\}$) the participation rate is around 50%. For the other strata ($h \in \{NA; RS; FS\}$) it is above 50%.

Table 4.3: Participation status of the initial sample by strata for SC3.

	$N5$	GY	HS	IG	MB	NA	RS	FS	MIG
participant	278	2187	666	270	292	416	1082	584	290
nonparticipant	219	1463	611	260	238	285	814	378	546
undetermined	77	100	35	8	38	14	105	74	147
$\sum$	574	3750	1312	538	568	715	2001	1036	983

Notes: Abbreviations of strata are $N5$: Grundschule & schulartunabhängige Orientierungsstufen, GY: Gymnasium, HS: Hauptschule, IG: Integrierte Gesamtschule & Freie Waldorfschule, MB: Schule mit mehreren Bildungsgängen, NA: not available, RS: Realschule, FS: Förderschule and MIG: schools of the migrants supplement.

In SC4 participation rates vary from 54.22% (stratum FS) to 68.12% (stratum GY) across strata. In stratum NA the participation rate is about 56% whereas the participation rates in the other strata are above 60%.

These figures reveal a more detailed difference in participation rates that are lower throughout the strata in SC3 than in SC4.

For some students the participation status could not be determined, because the did not return their consent form by the time the lists were sent to the survey research institute. Having no information on wether or not these students are willing to participate in the panel they were excluded in estimating the propensities to participate. The propensity of students to participate in the panel cohort is modeled via probit models (see Equation (4.1)) regressing the participation status (participant, nonparticipant) on characteristics available for the initial sample of students (that is, information from the students participation list, see Table B.2 in Appendix B).

did not allow for re-identification.

Table 4.4: Participation status of the initial sample by strata for SC4.

	GY	HS	IG	MB	NA	RS	FS
participant	4974	3753	1394	1032	1000	3174	1286
nonparticipant	2125	2070	738	580	669	1818	860
undetermined	203	211	33	51	119	285	226
$\sum$	7302	6034	2165	1663	1788	5277	2372

Notes: Abbreviations of strata are GY: Gymnasium, HS: Hauptschule, IG: Integrierte Gesamtschule & Freie Waldorfschule, MB: Schule mit mehreren Bildungsgängen, NA: not available, RS: Realschule and FS: Förderschule.

Starting-cohort-specific models were estimated with a further distinction between students attending regular and special schools. In this way the different sets of information available for the different starting cohorts were taken into account. For example, grades from the previous year were provided by teachers for students in grade 9, but they were not necessarily available for students in grade 5. For SC3 a separate model for the supplement of migrants was estimated. In addition to the information available on student characteristics, random intercepts were considered to reflect the cluster structure of the initial samples of students. The complete models by starting cohorts and samples are given in Table B.5 in Appendix B.

For SC3, models were estimated for the sample of students in regular and special schools as well as for the sample of students with migration background. For all three samples the propensity to participate is significantly negatively influenced by a language other than German spoken at home and also by missing values in characteristics related to competencies, that is, competencies in maths and reading, dyslexia, special educational needs, and attention deficit hyperactivity disorder. For students in regular schools the participation propensity is positively influenced by good competencies in maths and reading and negatively influenced by missing values in personal attributes, that is, gender, year of birth, and migration characteristics, that is, language spoken at home, nationality.

For all five subgroups of the starting cohorts considered (students in grade 5 and 9 in regular schools and special schools, as well as migrants' supplement of fifth graders) some of the findings can be generalized. In each subgroup the participation propensity of the students is effected in the same way: Male students and students speaking a language other than German at home have a lower propensity to participate. Furthermore, missing values in personal, migrational, and competence characteristics are strong predictors negatively influencing participation. A positive effect was found for good grades in maths and, except for the students with migration background,

younger students are more willing to participate.

For students in special schools, mathematical competence has a significant positive effect on the participation propensity, and for students with a migration background the dummy indicating a Turkish migration background is highly significant and negatively influences response propensity. The size of the test groups yields different effects throughout the three samples. Whereas it is insignificant in regular schools, it is significantly negative for special schools and positive for students with a migration background, see also Table B.5 in Appendix B.

In SC4 the effects of gender, grades in math, and the missing values in competence characteristics are of similar kind. Male students and students with at least half of the competence characteristics missing are less willing to participate, whereas students with good grades in math are more willing. Of the students in regular schools the younger students are significantly more willing to participate; also, good grades in German have a positive effect on participation propensity. Negative effects are found for students with special educational needs and for those with missing values in personal characteristics.

The inverse of the participation propensities derived by these models constitute adjustment factors for the computation of nonresponse adjusted design weights of students.[11] In more detail, the nonresponse adjusted weight w_{ijh} of a student i in school j in stratum h entering the panel is the product of the design weight d_{ijh}, the institution participation adjustment factor δ_{jf}, and the estimated participation propensity $\widehat{\lambda}_{ij}$ of a student estimated from the models displayed in Table B.5 in Appendix B), that is

$$w_{ijh} = d_{ijh} \cdot \delta_j \cdot \widehat{\lambda}_{ij}^{-1}.$$

The panel entry weight on the student level w_{ijh} is the basis for all further adjustments of the panel cohorts and can be found in the scientific use files of SC3 and SC4 (also SC2) under `w_t`. This denotes the 'design weight' for the target persons, that is, the students, which is corrected for individual nonresponse within the participating schools.

4.4 Adjustments of the panel cohort for successive waves

The adjustments described in the previous sections were necessary to adjust the initial sample of schools and students to the final panel cohort. For each

[11] To simplify notation, some of the super- and subscripts are omitted.

member of the panel cohort there are subsequent decisions for participating in each wave of the panel. When panel members participate in different waves of the survey this will result in several participation patterns.

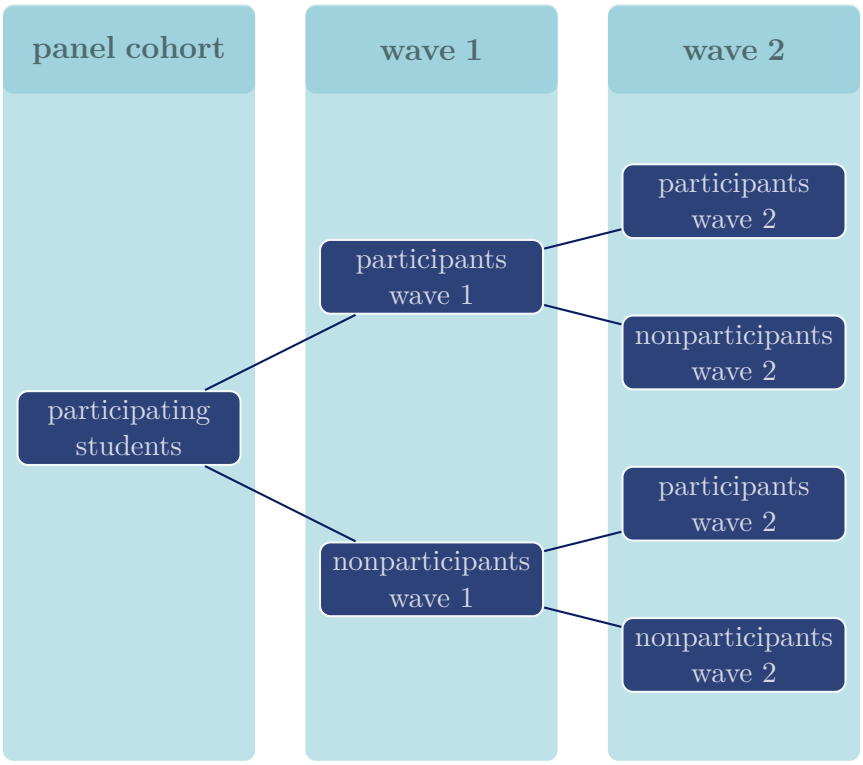


Figure 4.2: Participation patterns for panel cohort members.

Figure 4.2 somewhat simplifies these different patterns for two waves. Since panel cohort members are those that decide to participate in the panel, there are no nonparticipants. This is due to the fact that there was a small lag in time between the panel consent and the test days for wave 1. In the first wave there are again students that participate in surveying and testing and others that do not (for example because of illness or bad weather conditions that did not make it possible for them to show up). The same is true for wave two. Here each member of the panel cohort can participate (again) or not participate (again). Figure 4.2 simplifies also according to the declaration of nonparticipants. Within the NEPS any target person not participating in a wave is defined as temporary drop-out, whereas final drop-outs are defined as target persons for which over a time of two years no information was given. So nonparticipants in the figure summarize temporary and final drop-outs. The decision to participate in the second wave might be independent of the first

wave participation status. Students not participating a wave are considered as temporary drop-outs. Students that do not provide any information in surveying or testing within a two year period are considered as final drop-outs, that is, panel attrition. The figure shows on the top track of the pattern the 'all-time-participants' and on the lower track what might become panel attrition in wave three. All other tracks might be considered as irregular participation patterns. Others categorize the all-time-participants simply as 'respondents' and the irregular participants as 'non-attriters', see Kalton (1986) and Lepkowski (1989). The number of participation patterns increases with each wave, whereas at wave T there can be 2^T different participation patterns. For a survey running for three waves this makes $2^3 = 8$ different patterns, including the permanent nonrespondents pattern. To comply with the situation of drop-outs, the panel entry weights of the students need to be adjusted again for unit nonresponse among the panel cohort. As this would yield within only a few waves an enormous number of available weights, these published weights are selected on the basis of groups. These groups are either defined by a participation pattern (for example 'all-time-participants' or participant in the actual wave) or subgroups of the panel cohort that are of special interest in analyses. The latter is discussed in the succeeding chapter.

Table 4.5: Participation status for starting cohorts by wave.

	SC3 $n = 6112$		SC4 $n = 16425$	
	Wave 1	Wave 2	Wave 1	Wave 2
participant	5774	5790	15629	16017
temporary drop-out	338	308	796	408
final drop-out	0	14	0	0

Table 4.5 shows the distribution for participation status of the panel cohorts of grade 5 and 9 students by waves. In SC3, $n^r = 5778$ students (of the $n = 6112$ students in the panel) participated in the first wave yielding a participation rate of 94.54%. In the second wave of SC3 the participation rate is slightly higher with 94.73% and $n^r = 5790$ students participating. Of $n = 16425$ students in the SC4 panel, 15629 student took part in the first wave and 16017 students in the second wave, yielding participation rates of 95.15% and 97.52%. For the participation in the first wave of SC3 and SC4 and for the participation in the second wave of SC4 response propensity

models were estimated. In doing so, the participation propensity in the second wave of SC4 was modelled conditioned on a student's participation in the first wave. A cross tabulation of the participation status of SC4 students by wave is given in Table 4.6. The relevant models and their outcomes can be found in Table B.6 in Appendix B.

Table 4.6: Participation status by Starting Cohort and wave.

	Wave 2		
		drop-out	
Wave 1	participant	temporary	final
	Starting Cohort 3		
participant	5473	287	14
temporary drop-out	317	21	0
final drop-out	0	0	0
	Starting Cohort 4		
participant	15308	321	0
temporary drop-out	709	87	0
final drop-out	0	0	0

As shown in Table 4.6 the majority of 15308 (93.20%) students in SC4 participated in both waves. Whereas respectively 709 (4.32%) students participated in wave 2 and 321 (1.95%) in wave 1 only. The smallest fraction of 0.53% (87 students) participated neither in wave 1 nor in wave 2. The participation rate for both waves in SC3 is with 85.55% lower than in SC4. Thus the fractions of students participating in SC3 in one of the two waves only is higher with 5.19% (wave 2) and 4.70% (wave 1) respectively. Furthermore in the second wave of SC3 14 students refused further participation in the panel cohort, that is result in panel attrition.

Again the participation propensity for students is modelled using random intercept models with a probit link function (see Equation (4.1)). Models for wave 1 are based on information collected in the first wave. The information is provided by different informants. Mainly information arises from the students survey. But also information on the school or on classes are made available by the institution heads and teachers. In contrast, models for wave 2 can be based on more information since some of the students not participating in wave 1 provide information in wave 2 (see Table 4.6). In this case the

estimated models for wave 1 need to be updated as soon as new information arises. In SC4 information from wave 1 and wave 2 became available in the same time so that missing information in wave 1 could directly be edited if made available in wave 2. In SC3 new information from wave 2 was made available some time after wave 1. Here the new information was used in the second release for SC3.

Modeling students' participation decisions in SC3 is based on information on the first and second wave. In contrast to SC4 the time between wave 1 and wave 2 was about a year in SC3. In SC4 the time between the waves was about half a year. Since there are only 14 final drop-outs in wave 2 in SC4 the model is estimated based on 6098 remaining observations. On the one hand these 14 final drop-outs will for sure attrite for different reasons than those student temporarily dropping out. On the other hand extending models for these 14 cases cannot use much information. Therefore these cases are considered separately by unconditional modeling, see Table B.6 in Appendix B columns Wave 1 and Wave 2 for Starting Cohort 3.

In SC3 the decision to participate is negatively influenced by missing values in personal characteristics, that is, sex as well as month and year of birth. The same is the case for wave 1 in SC4, whereas the effect losses strength and significance in wave 2. Furthermore, students educated in schools of stratum IG and in schools in stratum MB have lower propensities to participate in wave 1 of SC3. This effect is not significant for wave 2 anymore. Instead the effect of being educated in special schools becomes negative and highly significant in wave 2. The decision to participate in wave 2 of SC3 is mostly negative influenced by not being in the context of a school anymore, that is being in the field for individual re-tracking. In this field the students are sent the survey and test questionnaires to their home instead of filling them out in their school. This is because 17 schools are not willing or able to cooperate, since the workload of the study organization is to high, the age group is expired, there are to few students willing to participate or the school was closed (reasons in descending order with respect to frequencies of occurrence).

A common characteristic positively influencing the participation decision across both starting cohorts and in each wave is speaking German as native language. Furthermore being schooled in special schools not only lowers participation propensities in wave 2 of SC3 but also in both waves of SC4. Visiting a special school has a significantly negative effect in both waves, though; in the second wave, students in special schools show an even lower participation propensity than in the first wave.

Students' propensity to participate in the first and second wave of SC4 is negatively influenced by having missing values in migration characteristics,

that is, native language and Nationality. The effects of being in a school of stratum HS as well as being German both negatively influence the students' propensity of participation but do not remain stable over the two waves. The influence of belonging to the younger half of the age cohort stays positive while growing stronger, and it becomes significant in the second wave. Surely, participating in the first wave is a strong predictor for participation in the second wave, exerting a highly significant positive effect.

In both starting cohorts the variance estimate for the random intercept at the school level increases from wave 1 to wave 2.

Having the information that arose in wave 2 of SC3 and SC3 the model for wave 1 participation was re-estimated making use of new information. Comparing the models for wave 1 based on the different sets of information they yield similar estimates.

The weight w_{ijh} for individual i in school j in stratum h is now adjusted again by a factor $\widehat{\lambda}_i^{-1}$ that comes from the appropriate model specification relying on the starting cohort and on the specific situation.[12] Hence any weight for individual i in the first wave of the panel will be computed as $w_i^I = w_i \frac{1}{\widehat{\lambda}_i^I}$ and for wave 2 as $w_i^{II} = w_i \frac{1}{\widehat{\lambda}_i^{II}}$, respectively.

The weight resulting from these adjustments can be found in the scientific use files of SC3 and SC4 under `w_t1`. This denotes the weight for the target persons (i.e., the students) in wave 1, which is corrected for individual nonresponse within the first wave.

[12] Again, we will skip further indices to simplify notation.

Chapter 5

Weighting multi-informant surveys in institutional contexts

Chapter outline:
After showing how we deal with wave-specific unit nonresponse this chapter focuses on weighting adjustments for special subgroups of interest. The NEPS adapts a multi-informant (also: multi-actor) perspective and enriches the data for students arising in the tests and surveys by an additional telephone interview with one parent. Thus one interesting subgroup is students and parents participating jointly, because the parents provide background information on the students environment. This widens the range for interesting analysis. Therefore we provide additional weighting adjustments accounting for nonresponse within this group. Because the participation decisions for students and parents are unlikely to be independent these decisions are modeled jointly resulting in a bivariate probit model allowing for the estimation a correlation parameter. Finding a correlation in a bivariate binary probit model motivated an extension for random intercepts. As in the previous chapter the random intercept respects the cluster structure of students within schools. Thus we adapt a bivariate probit model with random intercepts allowing for clustering at the school level to model jointly the (possibly) correlated participation processes of students and parents. The model results in a complex likelihood function with evaluations of the bivariate normal distribution. Therefore the approach developed by Geweke (1991), Hajivassiliou (1990) and Keane (1994) (the GHK-simulator) is suitable to estimate the model by means of simulation.

5.1 Students and parents participation decisions

Applying a multi-informant or multi-actor design can validate information given by the 'key-informant', that is, students, enrich information on their formal learning environment (for example information given by teachers on the context of the class or by the institutions head on the school) and their social background (for example parents providing further information on living conditions, socioeconomic status, etc.), see Wagner, Rau, and Lindemann (2010) for more details.[1] Maaz, Kreuter, and Watermann (2006) discuss the problems that can occur in multi-informant designs. For the PISA 2000 and the PISA-E (extension study surveying the student's parents, German: PISA-Erweiterung; see Baumert et al. (2002)) the authors show that students can provide reliable information on their parents educational school-leaving qualification and their occupation (Maaz et al., 2006).

Problems resulting in unit-nonresponse in multi-informant surveys occur for example when one or more informants refuse to participate. When enriching students data by information provided by their parents this may be the result of nonparticipation of the student or the parent (or both). The clear advantage of this multi-informant perspective is that in case of one informant not providing information on the topic of interest the information of the other informant can be used (Maaz et al., 2006).

The data collected in SC3 (and SC4 as well) is enriched by information on the school, the classes (especially for German and Maths) and the educators to allow for consideration of learning environments of students. Furthermore a computer assisted telephone interview (CATI) with one parent (if willing to participate in the survey as well) provides information on the social background of the student.

For example analysis of social disparity in educational participation or competence acquisition make use of characteristics on the social background provided multiple informants (Maaz et al., 2006). This is why students and parents participating together (stemming from the multi-informant design) form an interesting population for analysis. Therefore nonresponse adjusted weights need to be provided for this subgroup as well.

The participation processes resulting in the final panel cohort are embedded in the sampling and recruitment process as follows. The panel cohort sample has been established using a stratified two-stage cluster sampling ap-

[1] Multi-informant and multi-actor design describe the same design, whereas multi-informant is more prominently used in economic studies while multi-actor is predominantly used in family research.

proach. Stratification reflects the different school systems in Germany via seven explicit strata, see Aßmann et al. (2011) for details. Access to the target cohort, that is, students in grade 5, is gained via schools ensuring thus the contactability of students in sampled classes. In NEPS, participation is not mandatory and thus unit nonresponse can occur on each level, that is, schools, students, and parents. The process of school recruitment and subsequently the student recruitment process is consecutive by nature and thus reflected by sequential modeling (see previous chapter for details).

That is, first the participation decisions of schools are modeled as shown in Subsection 4.3.1. Second, the panel cohort is established by parents allowing their children to participate via an active consent, since a fifth grade student is not of legal age. After correcting for unit nonresponse on the school and student level each student of the panel cohort is assigned a panel entry weight, as shown in Subsection 4.3.2. These steps are, due to the nature of the decision process of schools and students, done using sequential adjustments applying appropriate model specifications. Since the students need their parents permission (granted by a parents' signature) to participate in the NEPS, a first contact with the parents is already established in the forerun of the survey. The provided consent to participate in the panel survey establishes the panel cohort for SC3. Third, given the panel consent provided by parents for their children, actual participation in each wave including testing of students in schools and the telephone interview of parents needs to be analyzed. The decision processes leading to actual participation within wave 1 are hence modeled subsequently.

However, data availability on students of the cohort depends on actual first wave participation. Further availability of data provided by parents on the student does depend on the participation decision of parents. This is to a larger extend embedded in the threefold process (location, contact and cooperation) leading to cooperation described by Lepkowski and Couper (2002), since not all parents provided sufficient contact information. Given the decoupled participation decisions, that is parents may grant their children to participate but refuse participation for themselves, the participation decision of parents realizes either when they provide consent for their children or during the contact procedure of the telephone interview.

The decision processes described above result in the joint participation statuses shown in Table 5.1. The table gives the participation statuses for students and parents by wave. The panel cohort consists of 6112 students from which 5774 participated in wave 1 (participation rate: 94.47%) and 338 were classified as temporary drop-outs due to illness, bad weather conditions, etc. Students participation rates in SC3 by institution range from 30.77% up to 100% (with median of 96.67%). The parents of the students were

Table 5.1: Participation statuses for students in SC3 and their parents by wave.

	parents		
		drop-out	
students	participant	temporary	final
	Wave 1		
participant	3974	462	1338
temporary drop-out	177	28	133
final drop-out	0	0	0
	Wave 2		
participant	3727	636	1427
temporary drop-out	92	104	112
final drop-out	1	2	11

less likely to participate in the CATI. Altogether 4151 parents participated in wave 1. The other 1961 parents did not participate in the first wave due to temporary drop-out or refusal. For the subgroup of 3974 of wave 1 participants an additional interview with one parent is available.

In wave 2 there are fewer students and parents participating together. The subgroup consists of 3727 students and parents in wave 2. For one student of SC3 who finally dropped out, an interview with one parent is available. For the other 13 students the parents could not be contacted or refused to participate in wave 2.

5.2 Model specifications for decision modeling

The following model specifications focus on appropriately modeling the two distinct participation decisions of students and parents. Therefore a bivariate binary probit model is set up and extend by random intercepts. A random intercept on the institutional level is chosen in modeling the students and parents participation decision accounting for clustering. This is in line with the model framework introduced in Equation (4.1) and used for analyzing participation decisions as discussed in the previous chapter.

According to Laaksonen (2005) binary regression models using a logit link

function seem to be dominant in modeling participation decisions in social research (and also in other fields). We decide for the probit link function because it allows for the estimation of a correlation parameter, which is not straightforward within a logit framework.

The bivariate model setting allows to consider possible correlations in the decision processes of students and parents and is therefore not a sample selection model in the sense of Heckman (1979) as in the literature discussed in Section 4.2. Steele and Durrant (2011) use a multilevel extension of the sample selection model to model non-contact and refusal.

5.2.1 Univariate probit model

The univariate probit model for $i = 1, \ldots, n$ individuals (for example students or parents) with dichotomous participation decisions y_i is given by

$$y_i = \begin{cases} 1 & \text{if} \quad \widetilde{y}_i > 0, \\ 0 & \text{else} \end{cases} \qquad \text{with} \quad \widetilde{y}_i = X_i\beta + \varepsilon_i, \tag{5.1}$$

where $\widetilde{y}_i$ denotes a latent variable, X_i the regressors, β the coefficients and $\varepsilon_i \sim N(0, \sigma)$ denotes the disturbance (with $\sigma = 1$). The univariate probit model estimates the probability for $y_i = 1$ by

$$\begin{aligned} \mathrm{P}(y_i = 1) &= \mathrm{P}(\widetilde{y}_i > 0) \\ &= \mathrm{P}(X_i\beta + \varepsilon_i > 0) \\ &= \mathrm{P}(\varepsilon_i > -X_i\beta) \\ &= \int_{-\infty}^{X_i\beta} \phi(\varepsilon_i)d\varepsilon_i = 1 - \int_{-X_i\beta}^{+\infty} \phi(\varepsilon_i)d\varepsilon_i \\ &= \Phi(X_i\beta) = 1 - \Phi(-X_i\beta), \end{aligned} \tag{5.2}$$

with ϕ denoting the density function of the normal distribution and Φ the distribution function of the normal distribution, see Greene (2012, p. 728).[2] Thus the contribution of an individual i to the likelihood $\mathcal{L}$ is

$$\mathcal{L}_i(\beta) = \mathrm{P}(Y_i = y_i | X_i, \beta) = \int_{D_{i\mathcal{L}}}^{D_{i\mathcal{U}}} \phi(\varepsilon_i)d\varepsilon_i = \Phi(X_i\beta), \tag{5.3}$$

where $D_{i\mathcal{L}} = (-\infty^{1-y_i}, -X_i\beta^{y_i})$ denotes the lower and $D_{i\mathcal{U}} = (-X_i\beta^{1-y_i}, +\infty^{y_i})$ the upper integration limits corresponding to $y_i = (0, 1)$.[3]

[2] Because of the symmetry of the normal distribution $1 - \Phi(X_i\beta) = \Phi(-X_i\beta)$.

[3] Because $\mathrm{P}(\widetilde{y}_i > 0|X) = \mathrm{P}(\varepsilon_i > -X_i\beta|X)$ and the symmetry of the normal distribution it is $\mathrm{P}(\widetilde{y}_i > 0|X) = \mathrm{P}(\varepsilon_i < X_i\beta|X) = \Phi(X_i\beta)$, see Greene (2012, p. 726).

Finally the likelihood for the model results in

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} \mathcal{L}_i(\beta). \tag{5.4}$$

This model framework can be extended with a random intercept. More precisely: to a random intercept probit model. This extension allows to take clustering on a higher level (for example school level) into account. Individuals (denoted by i) are clustered in groups $j = 1, \ldots, m$ of size n_j. Therefore the error term from Equation (5.1) is decomposed into

$$\varepsilon_{ij} = \alpha_j + \epsilon_{ij}. \tag{5.5}$$

Inserting Equation (5.5) in Equation (5.1) the extended model can be rewritten as

$$y_{ij} = \begin{cases} 1 & \text{if} \quad \widetilde{y}_{ij} > 0, \\ 0 & \text{else} \end{cases} \qquad \text{with} \quad \widetilde{y}_{ij} = X_{ij}\beta + \underbrace{\alpha_j + \epsilon_{ij}}_{\varepsilon_{ij}}, \tag{5.6}$$

where $\alpha_j \sim N(0, \omega^2)$ denotes the random intercept and $\epsilon_{ij} \sim N(0, \sigma)$ is the disturbance (again with $\sigma = 1$).[4] This model is equal to the model already given in Equation (4.1). Summarizing model parameters as $\theta = (\beta, \omega^2)$ the contribution to the likelihood by cluster j is the joint probability for all n_j individuals within the cluster, that is

$$\begin{aligned} \mathcal{L}_j(\theta) &= \mathrm{P}(Y_{\cdot j} = y_{\cdot j} | X_{\cdot j}, \theta) \\ &= \mathrm{P}(Y_{1j} = y_{1j}, \ldots, Y_{n_j j} = y_{n_j j} | X_{1j}, \ldots, X_{n_j j}, \theta) \\ &= \underbrace{\int_{D_{1j\mathcal{L}}}^{D_{1j\mathcal{U}}} \cdots \int_{D_{n_j j\mathcal{L}}}^{D_{n_j j\mathcal{U}}}}_{n_j} \phi_{n_j}(\varepsilon_{1j}, \varepsilon_{2j}, \ldots, \varepsilon_{n_j j}) d\varepsilon_{1j} d\varepsilon_{2j} \ldots d\varepsilon_{n_j j} \end{aligned} \tag{5.7}$$

We get the joint density of ϵ_{ij} integrating α_j out of the joint density of $(\varepsilon_{1j}, \varepsilon_{2j}, \ldots, \varepsilon_{n_j j})$, that is

$$\phi_{n_j}(\varepsilon_{1j}, \varepsilon_{2j}, \ldots, \varepsilon_{n_j j}) = \underbrace{\phi_{n_j}(\epsilon_{1j}, \epsilon_{2j}, \ldots, \epsilon_{n_j j} | \alpha_j)}_{\prod_{i=1}^{n_j} \phi(\epsilon_{ij} | \alpha_j)} \phi(\alpha_j). \tag{5.8}$$

[4] For further simplification in notation of the likelihood functions it is $\epsilon_{ij} = \widetilde{y}_{ij} - (X_{ij}\beta + \alpha_j)$.

Because the errors are uncorrelated given the cluster we can rewrite the multivariate normal distribution as the product of univariate normal distributions. Inserting this in Equation (5.7) yields

$$\begin{aligned}\mathcal{L}_j(\theta) &= \mathrm{P}(Y_{.j} = y_{.j}|X_{.j}, \theta) \\ &= \int_{-\infty}^{+\infty} \left(\int_{D_{1j\mathcal{L}}}^{D_{1j\mathcal{U}}} \cdots \int_{D_{n_j j\mathcal{L}}}^{D_{n_j j\mathcal{U}}} \prod_{i=1}^{n_j} \phi(\epsilon_{ij}|\alpha_j) d\epsilon_{ij} \right) \phi(\alpha_j) d\alpha_j \end{aligned} \tag{5.9}$$

In the above equation the order of the integrals can be rearranged because the ranges of the integration are independent. Further ϵ conditioned on α_j is independent so we can rewrite the above equation to

$$\begin{aligned}\mathcal{L}_j(\theta) &= \mathrm{P}(Y_{.j} = y_{.j}|X_{.j}, \theta) \\ &= \int_{-\infty}^{+\infty} \left(\prod_{i=1}^{n_j} \left[\int_{D_{ij\mathcal{L}}}^{D_{ij\mathcal{U}}} \phi(\epsilon_{ij}|\alpha_j) d\epsilon_{ij} \right] \right) \phi(\alpha_j) d\alpha_j. \end{aligned} \tag{5.10}$$

Butler and Moffitt (1982) derived this simplification for a one-factor multinomial probit model. A general derivation is given in Greene (2012, pp. 758 ff.). In this case the likelihood arises the product of the likelihood contributions from each of the clusters j, that is

$$\mathcal{L}(\theta) = \prod_{j=1}^{m} \mathcal{L}_j(\theta). \tag{5.11}$$

5.2.2 Bivariate probit model

The probit model given in Equation (5.1) can further be extended to allow for modeling two (possibly) correlated decisions. Let $c \in \{s, p\}$ denote a couple of a student s and a parent p. Further suppose the two decisions are somehow correlated.[5] This is measured via the correlation parameter ρ. Extending the univariate to the bivariate binary probit model Equation (5.1) changes to

$$\begin{aligned} y_i^s &= \begin{cases} 1 & \text{if} \quad \widetilde{y}_i^s > 0, \\ 0 & \text{else} \end{cases} \qquad \text{with} \quad \widetilde{y}_i^s = X_i^s \beta^s + \varepsilon_i^s \\ y_i^p &= \begin{cases} 1 & \text{if} \quad \widetilde{y}_i^p > 0, \\ 0 & \text{else} \end{cases} \qquad \text{with} \quad \widetilde{y}_i^p = X_i^p \beta^p + \varepsilon_i^p \quad . \end{aligned} \tag{5.12}$$

[5]Slightly abusing the notational conventions some symbols are defined different than in previous chapters. The substitutes will become clear at appropriate places in the text.

The correlation parameter ρ enters via the correlation matrix Σ of the residuals so that

$$\varepsilon_i^c = \begin{pmatrix} \varepsilon_i^s \\ \varepsilon_i^p \end{pmatrix} \sim N(0, \Sigma) \quad \text{with} \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}. \tag{5.13}$$

The bivariate probit model estimates the regression coefficients $\beta^c = (\beta^s, \beta^p)$ and the correlation coefficient ρ using the dependant variables $y^c = (y^s, y^p)$, the latent variables $\widetilde{y}_i^c = (\widetilde{y}_i^s, \widetilde{y}_i^p)$ and the characteristics $X_i^c = (X_i^s, X_i^p)$. The disturbance parameter is given by ε_i^c (see Greene, 2012, p. 778). Denoting the bivariate normal density by ϕ_2, the bivariate normal distribution function by Φ_2, $y^c = (y^s, y^p)$, $q^c = 2y^c - 1$, $\mu^c = (\mu^s, \mu^p) = (X^s\beta^s, X^p\beta^p)$, and $\theta = (\beta^s, \beta^p, \rho)$ the probabilities entering the likelihood are

$$\begin{aligned} \mathcal{L}(\theta) &= \mathrm{P}(Y^s = y_i^s, Y^p = y_i^p | X^s, X^p) = \mathrm{P}(Y^c = y_i^c | X^c) \\ &= \int_{D_{\mathcal{L}}^s}^{D_{\mathcal{U}}^s} \int_{D_{\mathcal{L}}^p}^{D_{\mathcal{U}}^p} \phi_2(\varepsilon^s, \varepsilon^p) d\varepsilon^s d\varepsilon^p = \int_{D_{\mathcal{L}}^c}^{D_{\mathcal{U}}^c} \phi_2(\varepsilon^c) d\varepsilon^c \\ &= \Phi_2(q^s\mu^s, q^p\mu^p) = \Phi_2(q^c\mu^c), \end{aligned} \tag{5.14}$$

with

$$\begin{aligned} D_i^c &= [D_i^s, D_i^p] \\ &= [(D_{i\mathcal{L}}^s, D_{i\mathcal{U}}^s) \times (D_{i\mathcal{L}}^p, D_{i\mathcal{U}}^p)] \\ &= \begin{cases} (-\infty, -\mu_i^s) \times (-\infty, -\mu_i^p), & \text{if } y_i^s = 0, y_i^p = 0 \\ (-\mu_i^s, +\infty) \times (-\mu_i^p, +\infty), & \text{if } y_i^s = 1, y_i^p = 1 \\ (-\infty, -\mu_i^s) \times (-\mu_i^p, +\infty), & \text{if } y_i^s = 0, y_i^p = 1 \\ (-\mu_i^s, +\infty) \times (-\infty, -\mu_i^p), & \text{if } y_i^s = 1, y_i^p = 0 \end{cases}, \end{aligned} \tag{5.15}$$

see Greene (2012, pp. 758f). Note that the set of regressors X_i^c does not necessarily have to be identical, that is, $X_i^s \neq X_i^p$. Furthermore for $\rho = 0$ the bivariate binary probit decomposes into two separate univariate binary probit models (Greene, 2012, p. 782).

The bivariate binary probit can also be extended to a bivariate binary probit model with random intercept. Therefore we will have to model two of the Equations given in (5.6). Again it is possible to decompose the error term as in Equation (5.5), see Greene (2012, pp. 784f). This changes

Equation (5.12) to

$$\begin{aligned} y_i^s &= \begin{cases} 1 & \text{if} \quad \widetilde{y}_i^s > 0, \\ 0 & \text{else} \end{cases} \qquad \text{with} \quad \widetilde{y}_{ij}^s = X_{ij}^s \beta^s + \underbrace{\alpha_j^s + \epsilon_{ij}^s}_{\varepsilon_{ij}^s} \quad \text{and} \\ y_i^p &= \begin{cases} 1 & \text{if} \quad \widetilde{y}_i^p > 0, \\ 0 & \text{else} \end{cases} \qquad \text{with} \quad \widetilde{y}_{ij}^p = X_{ij}^p \beta^p + \underbrace{\alpha_j^p + \epsilon_{ij}^p}_{\varepsilon_{ij}^p}. \end{aligned} \tag{5.16}$$

Here $\alpha_j = (\alpha_j^s, \alpha_j^p)$, with $\alpha_j \sim N(0, \Omega = \text{diag}(\omega_s^2, \omega_p^2))$ denotes the bivariate normal distributed random intercept for students and parents grouped in clusters $j = 1, \ldots, m$ of size n_j. Note that the random intercepts are uncorrelated because he off-diagonal elements of the covariance matrix are zero. The vector of disturbances given by ϵ_i^c is bivariate normal distributed with

$$\epsilon_{ij}^c = \begin{pmatrix} \epsilon_{ij}^s \\ \epsilon_{ij}^p \end{pmatrix} \sim N(0, \Sigma) \quad \text{with} \quad \Sigma = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}. \tag{5.17}$$

The model stated in Equation (5.16) therefore characterizes the joint probability of all couples $c \in \{s; p\}$ within cluster j, that is,

$$\text{P}(Y_{.j}^c = y_{.j}^c | X_{.j}^c, \theta), \tag{5.18}$$

where $Y_{.j}^c$ and $X_{.j}^c$ denote stacked vectors containing all information on the couple of students an parents in cluster j and $\theta = \{\beta_s, \beta_p, \rho, \text{diag}(\Omega)\}$ summarizes all parameters of the model. This formulation is similar to the formulation of the likelihood contribution for individual i of the random intercept probit given Equation (5.7). The difference is in the superscript denoting the student parent couple c. In order to provide individual participation probabilities serving as the basis for the derivation of adjustment factors one has to sum over the corresponding joint probabilities, that is

$$\text{P}(Y_{ij}^c = y_{ij}^c | X_{.j}^c, \theta) = \sum_{\Delta_j} \text{P}(Y_{.j}^c = y_{.j}^c | X_{.j}^c, \theta), \tag{5.19}$$

where $j = 1, \ldots, m$ and Δ_j denotes the set of combinations of participation decisions for all students and parents within cluster j. That is, the power set of the individual participation patterns of a student-parent couple $y^c = (y^s, y^o) = \{(0,0); (1,0); (0,1); (1,1)\}$ for $n_j - 1$ individuals. This set is required for each cluster j for marginalization of the considered individual participation probability. As $|\Delta_j|$ consists out of $4^{n_j - 1}$ combinations, computation becomes prohibitively burdensome for $n_j > 20$. To ensure computational feasibility the probabilities conditional on expected random intercepts

are considered, that is,

$$\mathrm{P}(Y^c_{iJ} = y^c_{iJ} | X^c_{\cdot j}, \theta, E[\alpha_j | Y^c_{\cdot j}, X^c_{\cdot j}, \theta]), \tag{5.20}$$

where $\widehat{\alpha}_j = E[\alpha_j | Y^c_{\cdot j}, X^c_{\cdot j}, \theta]$ is an estimate of the cluster-specific random intercept arising as conditional mean and provided as a byproduct of the estimation routine described below.

5.2.3 Parameter estimation

Summarizing all parameters of the bivariate binary probit model with random intercept given in Equation (5.16) as $\theta = \{\beta_s, \beta_p, \rho, \mathrm{diag}(\Omega)\}$, the corresponding likelihood contribution of cluster j results in a $2 \cdot n_j$ dimensional integral over the composite error term ε. Again this likelihood can be rearranged using Equation (5.8) and thus the likelihood contribution of cluster j results in a $2 \cdot n_j + 2$ dimensional integral over the decomposed error terms $\epsilon^c_{ij} = (\epsilon^s_{ij}, \epsilon^p_{ij})$ and the random intercepts $\alpha_j = (\alpha^s_j, \alpha^p_j)$. Using the decomposition of the error terms the complete likelihood can be written as

$$\begin{aligned}\mathcal{L}(\theta) = \prod_{j=1}^{m} \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} \left(\prod_{i=1}^{n_j} \left(\int_{D^s_{ij\mathcal{L}}(\alpha^s_j)}^{D^s_{ij\mathcal{U}}(\alpha^s_j)} \int_{D^p_{ij\mathcal{L}}(\alpha^p_j)}^{D^p_{ij\mathcal{U}}(\alpha^p_j)} \phi_2(\epsilon^s_{ij}, \epsilon^p_{ij}) \right) \right) \\ \phi_2(\alpha^s_j, \alpha^p_j) d\alpha^s_j d\alpha^p_j d\epsilon^s_{ij} d\epsilon^p_{ij}.\end{aligned} \tag{5.21}$$

Note that the term $\phi_2(\epsilon^s_{ij}, \epsilon^p_{ij})$ in the inner integral is not conditioned on the random intercepts. This is because the random intercepts enter via the integration limits given in Equations (5.24) and (5.25). Now summarizing the error terms as $\epsilon^c_{ij} = (\epsilon^s_{ij}, \epsilon^p_{ij})$ and the random intercepts as $\alpha_j = (\alpha^s_j, \alpha^p_j)$ the likelihood can be is shortened to

$$\mathcal{L}(\theta) = \prod_{j=1}^{m} \underbrace{\int_{-\infty}^{+\infty} \left(\prod_{i=1}^{n_j} \left(\int_{D^c_{ij\mathcal{L}}(\alpha_j)}^{D^c_{ij\mathcal{U}}(\alpha_j)} \phi_2(\epsilon^c_{ij}) \right) \right) \phi_2(\alpha_j) d\alpha_j}_{\mathcal{L}_j(\theta)} d\epsilon^c_{ij} \tag{5.22}$$

and explicitely given by

$$\begin{aligned}\mathcal{L}(\theta) = \prod_{j=1}^{m} \int_{-\infty}^{+\infty} \left(\prod_{i=1}^{n_j} \left(\int_{D^c_{ij\mathcal{L}}(\alpha_j)}^{D^c_{ij\mathcal{U}}(\alpha_j)} \frac{1}{2\pi} |\Sigma|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \epsilon^{c\prime}_{ij} \Sigma^{-1} \epsilon^c_{ij} \right\} d\epsilon^c_{ij} \right) \right) \\ \frac{1}{2\pi} |\Omega|^{-\frac{1}{2}} \exp\left\{ -\frac{1}{2} \alpha'_j \Omega^{-1} \alpha_j \right\} d\alpha_j,\end{aligned} \tag{5.23}$$

where n_j is the number of individuals i in cluster j, see Greene (2012, p. 785).[6] Because the distributions of the random intercepts are uncorrelated ($\alpha_j \sim N(0, \text{diag}(\Omega))$) they could also be rewritten as the product of two separate univariate normal distributions. The integration regions for the inner integral (including the random intercept, see Equation (5.25)[7]) over the bivariate normal distribution in Equation (5.21) are limited by

$$\begin{aligned} D^c_{ij} &= [D^s_{ij}, D^p_{ij}] \\ &= [(D^s_{ij\mathcal{L}}, D^s_{ij\mathcal{U}}) \times (D^p_{ij\mathcal{L}}, D^p_{ij\mathcal{U}})] \\ &= \begin{cases} (-\infty, -\mu^s_{ij}) \times (-\infty, -\mu^p_{ij}), & \text{if } y^s_{ij} = 0, y^p_{ij} = 0 \\ (-\mu^s_{ij}, +\infty) \times (-\mu^p_{ij}, +\infty), & \text{if } y^s_{ij} = 1, y^p_{ij} = 1 \\ (-\infty, -\mu^s_{ij}) \times (-\mu^p_{ij}, +\infty), & \text{if } y^s_{ij} = 0, y^p_{ij} = 1 \\ (-\mu^s_{ij}, +\infty) \times (-\infty, -\mu^p_{ij}), & \text{if } y^s_{ij} = 1, y^p_{ij} = 0 \end{cases} \end{aligned} \tag{5.24}$$

according to the different possible combinations of participation decisions of a student and its parent with

$$\mu^s_{ij} = X^s_{ij}\beta^s + \alpha^s_j \quad \text{and} \quad \mu^p_{ij} = X^p_{ij}\beta^p + \alpha^p_j. \tag{5.25}$$

The limits for the integrals in Equation (5.24) and (5.15) are identical except for the consideration of the random intercept term α_j in μ^c_{ij}. The likelihood of the model stated in Equation (5.21) can be calculated by the means of simulation.

To estimate the model the general idea is to use Monte Carlo integration and arrange the integral to take the form

$$\int_v g(v)f(v)dv, \tag{5.26}$$

where $f(v)$ denotes a regular density of a random variable v (for example a normal density) and $g(v)$ is a smooth function. The Monte Carlo approximation for the integral is expressed as a mean

$$\int_v g(v)f(v)dv = E_f[g(v)] \approx \frac{1}{S}\sum_{s=1}^{S} g(v_s), \tag{5.27}$$

where $v_s, s = 1, \ldots, S$ denote random draws from the density $f(v)$, see for example Jones, Maillardet, and Robinson (2009, p. 367). Based on the law of

[6] The R syntax for the implementation of the likelihood function as well es the estimation routine can be found in Appendix D. The code might help a reader familiar with R to understand the derivation of the likelihood.

[7] In the remaining equations we will skip the explicit conditioning on α_j so that $D^c_{ij}(\alpha_j) = D^c_{ij}$.

large numbers this approximation converges in probability to the expectation, see Greene (2012, p. 665).

Because the computation of the likelihood given in Equation (5.21) involves evaluations of the distribution function of the bivariate normal distribution, the approach developed by Geweke (1991), Hajivassiliou (1990) and Keane (1994) (GHK-simulator, documented in Geweke and Keane (2001)) can be adapted.[8] In general the GHK-simulator provides an approximation to the integral I over a K-variate normal distribution

$$\begin{aligned} I_K &= \int_{D_{\mathcal{L}}}^{D_{\mathcal{U}}} \phi_K = \int_{D_{\mathcal{L}}}^{D_{\mathcal{U}}} (2\pi)^{-\frac{K}{2}} |\Sigma|^{-\frac{1}{2}} \exp\left\{-\frac{1}{2}\epsilon'\Sigma^{-1}\epsilon\right\} d\epsilon \\ &\approx \frac{1}{S}\sum_{s=1}^{S}\prod_{l=1}^{K}[\Phi(\gamma_{sl\mathcal{U}}) - \Phi(\gamma_{sl\mathcal{L}})] = \widetilde{I} \end{aligned} \tag{5.28}$$

with $\epsilon \sim N(0, \Sigma)$ as the $K \times 1$ dimensional vector of error terms truncated in the region given by $D_{\mathcal{L}}$ and $D_{\mathcal{U}}$. The approximation is based on the property that any multivariate normal distribution can be factored into a set of corresponding conditional distributions given as univariate normal distributions. These univariate normal distributions are truncated at lower ($\gamma_{\mathcal{L}}$) and upper ($\gamma_{\mathcal{U}}$) truncation points corresponding to a transformation of the integration regions $D_{\mathcal{L}}$ and $D_{\mathcal{U}}$ using the Cholesky decomposition of Σ.

The likelihood stated in Equation (5.21) basically involves evaluation of n_j bivariate normal distributions ($K = 2$) within each cluster j. The GHK-simulator thus in our application approximates an integral over a rectangular region of the bivariate normal distribution for each individual i in cluster j. That is, we do not solve the $2 \cdot n_j$ dimensional integral over a multivariate normal distribution within an entire cluster j but instead we solve the n_j integrals for each couple within a cluster over the bivariate normal distribution. The bounds for the integration regions are given by D_{ij}^c. Now $s = 1, \ldots, S$ random draws $\widetilde{v}_s$ from truncated normals are generated using

$$\widetilde{v}_s = \Phi^{-1}[F\Phi(\gamma_{\mathcal{U}}^{(s)}) + (1 - F)\Phi(\gamma_{\mathcal{L}}^{(s)})], \tag{5.29}$$

[8] An illustrative example can be found in Appendix C. The example given differs slightly but should be helpful to understand what the importance sampler does.

with $F \sim U[0;1]$, see Greene (2012, p. 648). Whereas the truncation points are given by

$$\begin{aligned} \gamma_{\mathcal{U}}^{(s)} &= \left(\frac{D_{ij\mathcal{U},l} - L_{1:l-1,l} \cdot \widetilde{v}_{1:l-1,s}}{L_{l,l}} \right) \quad \text{and} \\ \gamma_{\mathcal{L}}^{(s)} &= \left(\frac{D_{ij\mathcal{L},l} - L_{1:l-1,l} \cdot \widetilde{v}_{1:l-1,s}}{L_{l,l}} \right). \end{aligned} \tag{5.30}$$

Here L denotes the Cholesky decomposition of Σ, that is $\Sigma = LL'$. Further $L_{1:l-1,l}, l = 1, \ldots, K$ denotes the vector of elements in columns l up to row $l-1$. Finally $\{\widetilde{v}_{l,s}\}_{s=1}^{S}$ denotes a matrix of S draws for each of the $l = 1, \ldots, K$ dimensions from a normal distribution truncated in the region $[\gamma_{\mathcal{L}}^{(s)}, \gamma_{\mathcal{U}}^{(s)}]$ corresponding to the transformation of D_{ij}^{c} according to Equation (5.30) and $\widetilde{v}_{1:l-1,s}$ denotes a vector stacking of these draws.

Using the above described GHK-simulator the approximated contribution of cluster j to the likelihood is given by

$$\begin{aligned} \widetilde{\mathcal{L}}_j(\theta) &= \mathrm{P}(Y_{\cdot j}^{c} = y_{\cdot j}^{c} | X_{\cdot j}^{c}, \theta) \\ &= \int_{-\infty}^{+\infty} \left(\prod_{i=1}^{n_j} \left(\int_{D_{ij\mathcal{L}}^{c}}^{D_{ij\mathcal{U}}^{c}} \phi_2(\epsilon_{ij}^{c}) \right) \right) \phi_2(\alpha_j) d\alpha_j d\epsilon_{ij}^{c}. \\ &= \int_{D_{ij\mathcal{L}}^{c}}^{D_{ij\mathcal{U}}^{c}} \int_{-\infty}^{+\infty} \left(\prod_{i=1}^{n_j} \phi_2(\epsilon_{ij}^{c}) \right) \phi_2(\alpha_j) d\alpha_j d\epsilon_{ij}^{c}. \end{aligned} \tag{5.31}$$

The integration regions D_{ij}^{c} enter the truncation regions $\gamma_{\mathcal{L}}^{(s)}$ and $\gamma_{\mathcal{U}}^{(s)}$ for the GHK-simulator and include the random intercepts α_j. The truncation regions are given for n_j all couples c within the cluster so that the likelihood contribution of cluster j arises as

$$\widetilde{\mathcal{L}}_j(\theta) = \frac{1}{S} \sum_{s=1}^{S} \left(\underbrace{\prod_{l=1}^{2n_j} \left[\Phi(\gamma_{\mathcal{U}}^{(s)}) - \Phi(\gamma_{\mathcal{L}}^{(s)}) \right]}_{\widetilde{I}(s|\alpha_j^{(s)}) = g(v)} \right). \tag{5.32}$$

The term $\widetilde{I}(s|\alpha_j^{(s)}) = g(v)$ here refers to the Monte Carlo approximation stated in Equation (5.27) using the GHK-simulator to approximate the integral given in Equation (5.28) for $K = 2$. Thus $f(x)$ is the remaining part of

the likelihood, so that the simulated likelihood finally arises as

$$\widetilde{\mathcal{L}}(\theta) = \prod_{j=1}^{m} \widetilde{\mathcal{L}}(\theta)_j = \prod_{j=1}^{m} \left(\frac{1}{S} \sum_{s=1}^{S} \widetilde{I}(s|\alpha_j^{(s)}) \right), \tag{5.33}$$

where the conditioning on $\alpha_j^{(r)}, r = 1, \ldots, R$ enters via the lower and upper bounds, that is, $D_{ij}^c = (D_{ij\mathcal{L}}^p, D_{ij\mathcal{U}}^s)$ corresponding to D_{ij}^c. Rearranging the likelihood this way for each individual i in cluster j only a two dimensional integral has to be evaluated using the GHK-simulator. R here denotes the number of random draws. The cluster specific expected random intercept $\widehat{\alpha}_j = E[\alpha_j | Y_{\cdot j}^c, X_{\cdot j}^c, \theta]$ (see Equation 5.20) needed for efficient computation of participation probabilities is numerically approximated by

$$\widehat{\alpha}_j \approx \widetilde{\alpha}_j = \frac{\frac{1}{R} \sum_{r=1}^{R} \alpha_j^{(r)} \left(\prod_{i=1}^{n_j} \frac{1}{S} \sum_{s=1}^{S} \widetilde{I}(s|\alpha_j^{(r)}) \right)}{\frac{1}{R} \sum_{r=1}^{R} \left(\prod_{i=1}^{n_j} \frac{1}{S} \sum_{s=1}^{S} \widetilde{I}(s|\alpha_j^{(r)}) \right)}, \tag{5.34}$$

see Greene (2004) or Train (2009, p. 263). Since the model described above is–by now–not available in *The R Project for Statistical Computing* (R Core Team, 2014) the estimation routine is implemented. The source code of the implementation in R is given in the Appendix D.

5.2.4 Simulation based evaluation

To assess the quality of the estimation procedure we need to check the statistical precision and the numerical precision. Both are assessed within two separate simulation studies. Both use $R = 1000$ replications. To check for statistical precision each of the replications include $m = 50$ clusters each of size $n_j = 30$, so that the total number of cases for each replication is $N = 1500$. The parameters of the model given in Equation (5.16) and the data generating process for each replication are specified as follows. The regression coefficients for the parent equation are $\beta^p = (1, 0.4, 0.6)'$ and the auxiliary variables are

$$\begin{aligned}
&X^p = (x_1^p, x_2^p, x_3^p) \qquad \text{with} \\
&x_1^p = (1, \ldots, 1)', \\
&x_2^p = (x_{2,1}^p, \ldots, x_{2,N}^p)' \sim N(0, 3) \qquad \text{and} \\
&x_3^p = (x_{3,1}^p, \ldots, x_{3,N}^p)' \sim N(0, 3).
\end{aligned}$$

For the equation of students regression parameters are $\beta^s = (-1, 0.5, -1.5)'$ and the auxiliary variables are

$$\begin{aligned} X^s &= (x_1^s, x_2^s, x_3^s) \qquad \text{with} \\ x_1^s &= (1, \ldots, 1)', \\ x_2^s &= (x_{2,1}^s, \ldots, x_{2,N}^s)' \sim N(0,3) \qquad \text{and} \\ x_3^s &= (x_{3,1}^s, \ldots, x_{3,N}^s)', \sim N(0,3). \end{aligned}$$

Further the error terms are correlated with $\rho = 0.4$ and the variances for the random intercepts are $\Omega = \text{diag}(\omega_s^2, \omega_p^2) = \text{diag}(0.8^2, 1.2^2)$. Within each replication the bivariate binary probit with random intercepts is estimated. Each of the replications is based on a different seed for the random number generator giving the X^p and X^s but on the same for $\{\widetilde{v}_{l,s}\}_{s=1}^S$. Thus variation is only induced by the different replications generating the data but not by the random numbers used in the estimation procedure. The results produced by the $R = 1000$ replications are then averaged. Table 5.2 gives these averaged results for statistical precision. The true parameters θ of the data generating process are given in the first column θ. The further columns report the average estimated parameter $\overline{\widehat{\theta}}$, the average standard deviation of the parameter estimates $ASE(\widehat{\theta})$, average bias $ABias(\widehat{\theta})$ as well as the average mean squared error $AMSE(\widehat{\theta})$. The four columns give standard deviations σ for the estimated parameters $\widehat{\theta}$ and their standard errors $SE(\widehat{\theta})$. The last column gives the coverage rates $\frac{\mathcal{I}(\theta \in CI)}{R}$ for 95%-confidence intervals. Standard errors are computed by inversion of the Hessian matrix.

Table 5.2: Statistical precision for $R = 1000$ replications.

	θ	$\overline{\widehat{\theta}}$	$ASE(\widehat{\theta})$	$ABias(\widehat{\theta})$	$AMSE(\widehat{\theta})$	$\sigma_{\widehat{\theta}}$	$\sigma_{SE(\widehat{\theta})}$	$\frac{\mathcal{I}(\theta \in CI)}{R}$
β_1^p	1.0	1.010	0.18019	0.00981	0.03838	0.19576	0.02371	0.92
β_2^p	0.4	0.402	0.02673	0.00186	0.00076	0.02758	0.00175	0.95
β_3^p	0.6	0.603	0.03478	0.00254	0.00132	0.03633	0.00257	0.95
β_1^s	-1.0	-0.950	0.14405	0.05042	0.02291	0.14278	0.01827	0.92
β_2^s	0.5	0.476	0.04300	-0.02366	0.00234	0.04220	0.00484	0.89
β_3^s	-1.5	-1.428	0.11202	0.07160	0.01725	0.11017	0.01410	0.88
ρ	0.4	0.364	0.11805	-0.03635	0.01440	0.11440	0.01929	0.95
ω_p	1.2	1.190	0.14403	-0.00981	0.02347	0.15295	0.02108	0.93
ω_s	0.8	0.667	0.15593	-0.13250	0.04276	0.15885	0.03257	0.87

Note: Simulation sample size for the GHK-simulator $S = 1000$.

The results show for all parameters a low bias. Coverage rates are as expected indicating good statistical precision. The largest bias is found for

$ABias(\omega_s) = -0.1325$ at a coverage rate of $\frac{\mathcal{I}(\omega_s \in CI)}{R} = 0.87$. This parameter also yields the lowest coverage. Table 5.2 shows that the average standard errors $ASE(\widehat{\theta})$ are close to the standard deviations of the estimates $\sigma_{\widehat{\theta}}$.

To check for numerical precision another simulation was performed. Randomly choosing one of the data produced in the previous replications for statistical precision we now estimate the bivariate binary probit with random intercept using $R = 1000$ replications. Within each of this replications different sets of random numbers $\{\widetilde{v}_{l,s}\}_{s=1}^{S}$, see Equation (5.30) are used. Thus variation is now induced by different sets of random numbers in the estimation procedure and not via the data generating process. Again results from the different replications are averaged and θ gives the same parameters for the data generating process as before in Table 5.3. The average estimate is given in column $\overline{\widehat{\theta}}$ and column $ASE(\widehat{\theta})$ reports the average standard error for the estimated $\widehat{\theta}$s. Columns $\sigma_{\widehat{\theta}}$ and $\sigma_{SE(\widehat{\theta})}$ give the according standard deviations for the estimates and their standard errors.

Table 5.3: Numerical precision for $R = 1000$ replications.

	θ	$\overline{\widehat{\theta}}$	$ASE(\widehat{\theta})$	$\sigma_{\widehat{\theta}}$	$\sigma_{SE(\widehat{\theta})}$
β_1^p	1.0	0.897	0.19962	0.05061	0.01366
β_2^p	0.4	0.453	0.03239	0.00099	0.00008
β_3^p	0.6	0.660	0.04276	0.00133	0.00011
β_1^s	-1.0	-1.155	0.20087	0.02403	0.00471
β_2^s	0.5	0.531	0.06134	0.00622	0.00101
β_3^s	-1.5	-1.566	0.16648	0.01823	0.00307
ρ	0.4	0.396	0.14995	0.02297	0.00732
ω_p	1.2	1.316	0.16846	0.03699	0.01468
ω_s	0.8	0.829	0.19304	0.03166	0.01040

Note: Simulation sample size for the GHK-simulator $S = 1000$.

The variation of estimates $\sigma_{\widehat{\theta}}$ and their standard errors $\sigma_{SE(\widehat{\theta})}$ induced by different sets of random numbers is within reasonable bounds. Comparing the average standard error $ASE(\widehat{\theta})$ and the standard deviation of the estimate $\sigma_{\widehat{\theta}}$ they do not overlay although for β_1^p the ratio is $\frac{\sigma_{\widehat{\theta}}}{ASE(\widehat{\theta})} = 0.254$. Similar results were produced when randomly choosing different data from the simulation for statistical precision.

We performed the simulation studies on a virtual machine with 4 Intel(R) Xeon(R) CPU X7550 with a 2.00GHz each equipped with 8 cores and a total of 32GB memory. The simulation for statistical precision lasted 24.5 hours and the simulation for numerical precision lasted 24.2 hours. The

estimation of the bivariate binary probit with random intercept is executed on a Intel(R) Core(TM) i3-3220 CPU with 3.30GHz and 8GB memory in less than 16 minutes. Times reported here are according to the settings given in Appendix D.

5.3 Application in grade 5 – re-weighting students and parents

Providing weights for this subgroup of special interest their response propensity was modeled having a bivariate probit model or a random intercept model at hand.

In Starting Cohort 3 of the NEPS (grade 5 students) the surveying and testing of students is accompanied by a telephone interview with one of the students' parents. In this parental interview some information given by the student are validated and additionally background information on the students environment are collected.

The decision processes described above result in the joint participation statuses shown in Table 5.1. The table gives the participation statuses for students and parents by wave. The panel cohort consists of 6112 students from which 5774 participated in wave 1 (participation rate: 94.47%) and 338 were classified as temporary drop-outs due to illness, bad weather conditions, etc. Students participation rates in SC3 by institution range from 30.77% up to 100% (with median of 96.67%). The parents of the students were less likely to participate in the CATI. Altogether 4151 parents participated in wave 1. The other 1961 parents did not participate in the first wave due to temporary drop-out or refusal. For the subgroup of 3974 of wave 1 participants an additional interview with one parent is available. In wave 2 there are fewer students and parents participating together. The subgroup consists of 3727 students and parents in wave 2.

In modeling the response propensities for students and parents the dependent variable is the binary participation status in the corresponding wave. Regressors included in the model comprise variables related to sampling characteristics such as stratification variables, variables on characteristics describing the socio demographics, for example sex or migration background and variables involving para data from call records.

Keep in mind that the variety of Federal-State-specific school systems as well as different transitions between elementary and secondary school institutions are respected via seven explicit strata, see Section 3.5. An additional

supplement of schools providing access to students with a turkish migration background or a migration related to the former Soviet Union is included as well and is the reference category.

Besides that information related to the socio demographic and family background include the age group of the student, gender, native language and migration background. According to year and month of birth the students are split by the median into a younger and an older half of the age group (reference category). Gender includes female and male, with male being the reference category. Native language consists of German and other (reference category) and the migration background is either turkish or related to the former Soviet Union (reference category). Information on the migration background of the students was available from the school records and provided by teachers in the forerun of the survey. A missing indicator is included for information on missing values in gender or age.

Besides that there is para data (Couper, 1998; Groves & Heeringa, 2006) available arising from test and telephone interview protocols during field work. The para data is available for those parents that were contacted. For parents the number of calls to the first contact is recorded as para data. It is included in the model as a dummy variable if the number of calls is less than four, with three calls being the median. Models of the second wave are conditioned on the first waves participation status of students and parents. Besides that a dummy is included indicating if a student has left the institutional context of a school and is followed up and surveyed individually. For all variables used in modeling the participation propensities Table B.7 gives the number of cases (n) an their corresponding proportions (p) for each category of the variable. Participation statuses are given for students, parents and students and parents participating jointly in each wave.

For each wave we show the following settings. First separate univariate models without random intercepts are estimated for students (I) and parents (II). Second a joint bivariate model without random intercepts (III) is estimated. We then proceed with separate random intercept models for students (IV) and parents (V). Lastly we model joint participation of parents and students using the bivariate probit with random intercepts (VI) as stated in Equation (5.16). Each model has an additional suffix corresponding to wave 1 (a) and 2 (b), respectively. The values for the log-likelihood, AIC and BIC as well as the χ^2 of the likelihood ratio test for model comparison can be found in Table B.8. To test the bivariate model against two univariate models we use $\chi^2 = 2 \cdot (\ln \mathcal{L}_{1,2} - (\ln \mathcal{L}_1 + \ln \mathcal{L}_2))$, see Greene (2012, p. 782). Testing model specifications for models with random intercept is non standard, since the variance for the random intercept lies at the boundary of the parameter space. This is a violation of the standard regularity

conditions that causes the invalidity of the asymptotic χ^2-distribution of the likelihood ration test statistic. Gouriéroux, Holly, and Monfort (1982) derive an asymptotic distribution as a mixture of χ^2-distributions. This asymptotic distribution for testing the significance of in general k random coefficients using the likelihood ratio test takes the form $\sum_{df=0}^{k} w(k, df)\chi^2(df)$, where $w(k, df) = \frac{\binom{df}{k}}{2^k}$ and $\chi^2(df)$ denotes a χ^2-distribution with df degrees of freedom and $\chi^2(0)$ the unit mass at the origin. The resulting critical values are lower than those of a standard likelihood ratio test. Keeping this in mind when assessing the significance of random intercepts via standard likelihood ratio tests provides a test with a significance level reaching at most the announced one, see also Harvey (1989). Since the model specifications differ in only one parameter, that is ρ, the critical value for the test is $\chi^2(1) = 3.841$. The value for the asymptotic χ^2-distribution with one degree of freedom and one random intercept is $\frac{\binom{1}{1}}{2^1} \cdot 3.841 = 1.921$.

The likelihood ratio test clearly stress the importance of considering correlation as well as clustering. In wave 1 and 2 the model settings without clustering show that the consideration of correlation in error terms are fitting the data better at a significance level of 5% in wave 1 and at 0.1% at wave 2. This finding also applies to the consideration of clustering. The model settings that respect the cluster structure, that is, the random intercept models, fit the data better than those that do not. Besides that the bivariate model setting with random intercepts, as stated in Equation (5.16), fits the data significantly better than the two separate random intercept models do. This finding applies for wave 1 at a significance level of 1% and for wave 2 at a level of 0.1%.

Focusing on the subgroup of students and parents Table B.9 and B.10 provide the estimated models for the participation propensities of the students and parents in wave 1. Table B.11 and B.12 provide the models for wave 2.

Except for the random intercept the corresponding models are equal with respect to their covariates. In line with the literature reviewed in Section 4.2 we use the probit link function since it can be easily extended with random intercepts and moreover it allows for the estimation of the correlation parameter ρ in the bivariate model setting. In contrast we do not apply a sample selection model since our interest lies in the joint participation decisions of students and parents. The random intercept is for both (parents and students) specified on the school level. There is no random intercept on the interviewer level since parents that did not want to participate and did not provide contact information do not allow for this specification.

The variables used in the analysis are the strata relevant for sampling

representing also the school type the student is educated in ($N5$ refers mostly to primary schools and schools not educating students in grades higher than six), the migrational background (russian or turkish), the students' gender (male or female), their native language (German or other) and the age group of the student (younger half or older half).[9] Furthermore a dummy is included for missing information in personal characteristics of the student, that is, missing information in gender or age group. For the parents the number of contacts (if contact attempts were possible) was recorded. Half of the parents could be contacted using less than four contact attempts (with three being the median) and the other half needed to be contacted more often (with a mean 8 calls). Separation problems occur when using variables together that are missing for nonparticipants, because there is no information on this group. This problem is eased by information on nonparticipants of wave 1 participating in wave 2 of the panel. Furthermore, information not available yet (especially for parents, the students environment and competencies) may be available for weighting adjustments in future waves.

Tables B.9 and B.10 show the estimated coefficients for the different model specifications in wave 1. The main effects remain stable throughout all specifications, that is, they do not change in sign and magnitude. This is also true for the variance parameter of the random intercepts as well as for the correlation coefficient. Comparing the bivariate probit model without and with a random intercept the correlation increases slightly when considering the clustered structure of the data using random intercepts.

The same findings apply to wave 2. The models for wave 2 include 14 observations that are classified as final drop-outs. These students withdraw their panel consent between wave 1 and wave 2. A more detailed analyses of these 14 cases is, due to the small number, not possible. An estimation of the model with and without the final drop-outs did induce only small changes in the estimated coefficients. A small difference occurs in the variance parameters of the random intercept model and in the parental equation of the bivariate probit with random intercept. This parameter slightly reduces.

For wave 1 the bivariate probit with random intercept shows significantly negative effects for all secondary school types but FS. Negative effects are found for students being educated in a comprehensive school (Stratum IG) and schools offering several tracks of education (Stratum MB). The missing indicator is also significant and has a large influence on the participation decision. This is mostly due to the fact that information is missing for nonparticipants. A positive effect is found for students speaking German

[9] The students were categorized by their date of birth into the two groups. The younger half contains all students being younger than the median age.

as a native language in both participation decisions. Within the parental participation decision the school types have a positive effect for primary schools educating students in grade 5 (Stratum $N5$) and Gymnasien (Stratum GY). Parents having a child with a turkish migration background influences their participation decision positively. Lastly the dummy for the low number of contact attempts to the first contact indicates that parents that are easy to reach, that is, have a higher propensity of being at home and contacted, have a higher propensity to participate. This has to do with what Durrant and Steele (2009) call lifestyle characteristics, these characteristics influence the propensity of being at home.

For both, parents and students, a significantly and large variation in the level of the participation propensity across schools was found. Lastly there is little, but significant, correlation in the error terms of the model.

For wave 2 Tables B.11 and B.12 show the corresponding models describing the participation propensities. Students propensity is lowered strongly if they are educated in special schools (Stratum FS). The impact of the missing indicator reduces (compared to wave 1) in wave 2. A negative effect on the students participation decision is found for students being in the field of individual re-tracking. Students are handed to that field if the students cannot be surveyed and tested in their institutional context of the school. The participation status of the parents also positively influences the students propensity to participate, whereas the students own participation status is of negative sign. For parents the participation status of themselves and their child has a positive effect on wave 2 participation. The number of calls to the first contact being less than four is again a strong predictor for participation of the parents. The impact of the different school types remains stable and increases for comprehensive schools (Stratum IG).

Based on these models according adjustments are derived for the subgroup of students and parents participating jointly in wave 1 and wave 2 of SC3 in the NEPS.[10] Given the design weight d_i for student i the adjustment yields an adjusted weight $w_i^{s,p}$ using

$$w_i^{s,p} = d_i \cdot \widehat{\lambda}_i^{-1},$$

where $\widehat{\lambda}_i$ is the estimated participation propensity for the jointly participating couple of a student s and its parent p derived from the models VIa and VIb shown in Table B.10 and Table B.12.

[10] Again, we skip further indices to simplify notation.

Chapter 6

Concluding remarks

6.1 Summary

In surveying and testing students in Germany there are clear needs for complex sampling designs. The hierarchical and stratified school system in Germany is regarded using stratified multistage designs within the NEPS SC3 and SC4. To provide sufficient sample sizes for subgroups oversamplings are applied. To correct for aspects of this complex design in estimation the use of design weights is necessary. The panel entry weights provided along with the data do incorporate important features of the sampling design and the decision processes leading to the actual panel cohorts. Including them in the analysis will help to avoid bias in estimation of population parameters. Ignoring these design features would lead to biased results (Kreuter & Valliant, 2007).

Beside design features unit nonresponse and panel attrition are further sources of potential bias. This bias can be induced by non participation or drop-out of specific non-random subgroups and makes adjustments of the design weights necessary. Within the NEPS complex models to compensate for unit nonresponse in weighting adjustments are applied. The models are based on random intercept models to account for clustering at the school level. Further models for reweighting subgroups regard correlation in decision processes of students and their parents participating in the panel together. The models estimating participation decisions on different stages, that is, school and student level, for different waves as well as for different subgroups (for example student and parents) take the particularities of the design into account and reveal typical explaining factors of unit nonresponse. Further-

more they point at the need to consider important aspects of the design, cluster structures and correlations in modeling decision processes. Finally analyzing small subgroups, for example migrants, can yield inefficient estimates also adjusted weights are used (Little, 2004).

6.2 Critical assessment

6.2.1 Complex sampling designs

Also a simpler design might have been desirable for users of the data the stratified two-stage sampling design for regular schools was the only design that appropriately mirrors the federal and hierarchical structure of the school system in Germany.

Within the strata of special (and also of regular) schools systematic *pps* sampling could have been avoided using a stratification proportional to the size of schools and applying different sampling fractions to the strata. Using stratification by school size would have lead to an enormous number of strata, so that *pps* sampling was preferred.

Systematic sampling usually does not allow for computing second order inclusion probabilities. This makes classical variance estimation based on these inclusion probabilities impossible. Since there is a variety of other approaches (jackknife, bootstrapping, balanced repeated replication) allowing for variance estimation *pps* sampling allowed better controlling sample sizes than a simple random sampling.

Lastly the supplement related to students having a Turkish migration background or a migration background related to the former Soviet Union could have been based on names, as suggested by Schnell et al. (2013). Not having a list at hand made a two-stage approach necessary anyways. Using name based approaches would have put additional workload on the schools' staff. Furthermore the Ministries of Education were able to provide figures quantifying the (approximate) number of migrants in sampled schools this approach was implemented. So this approach did not put additional workload on the schools and sounded promising. In the aftermath schools sampled in this supplement should not have explicitly been asked to provide access to migrant groups only. Rather these schools should have been integrated in the normal design also including students with or without other migration backgrounds.

Replacing a nonparticipating and originally sampled school by a predefined replacement school assumes the schools to be identical or at least similar

in characteristics of interest. Besides that there is no inclusion probability for the replacement school and thus design weights cannot be computed. These two drawbacks were accepted to counteract against the reduction in sample size on the school level, because the study is not mandatory. The drawback of not having inclusion probabilities for replacement schools could have been eradicated by the pseudo weighting approach documented in Elliott (2009).

6.2.2 Modeling unit nonresponse and weighting adjustments

The literature on modeling unit nonresponse stresses the importance to properly account for differences between refusal and noncontact. The models applied at different stages and for different participation decisions have throughout aggregated the variety of manifestations of unit nonresponse to the level of nonparticipation. On the school level a school could explicitly refuse or not make a statement towards participation (implicit refusal) resulting in nonparticipation of the school. The same applies for the initial sample of students. They also provided either an explicit denial or just did not return their consent form. These differences have not been addressed by the models due to the small number of cases.

Adjustments on the school level alternatively could have been based on the measure of size or in line with other adjustments on response propensity reweighting. Finding only few explaining factors of school refusals we decided for cell weighting forming the cells by significant factors. It was based on the number of schools within the cells yielding almost the same adjustment factors as if based on the measure of size.

Adjustments on the initial sample of students used all information available on participants and nonparticipants and therefore result in five different adjustment models (three in SC3 and two in SC4, see Table B.5) since the sets of information differed.

In the final panel cohorts the models do not address differences in temporary or final drop-out. This is–by now–due to the small number of final dropouts that only occurred in SC3. In future adjustments multinomial models will have to be extended to allow for consideration of the cluster structure to more accurate account for the differences in participation, temporary and final drop-out.

One extension useful in the bivariate probit model with random effects would be allowing for two differently specified random effects. The random effect at the school level is accurate for students as they are surveyed in

schools but for parents being interviewed in a CATI a random effect at the interviewer level would be more accurate as discussed by O'Muircheartaigh and Campanelli (1999) and Durrant and Steele (2009).

Lastly there are yet no population weighting adjustments applied to the weights. This is because subgroups are by now based on non-matching definitions for which population weighting adjustments can be applied or due to the fact that the information needed is not yet available.

6.3 Outlook and future Research

With further progress of the panel cohorts models used in weighting adjustments will become more complex. In SC4 students will leave school and start vocational training. Others will leave their actual school or change school to get an university entrance certificate. Together with those staying in their school this makes three subgroups to consider, whereas it will be more than this three for sure. In SC3 students finally dropped out in wave 2. Also just being a small number it will increase in future waves. This will have to be regarded in modeling leading to a multinomial probit with random effects. Future research will depend on the educational pathways entered by SC4 students and participation patterns over different waves. Mostly these two characteristics together with the data available will give directions for extending existing models. Extensions do not only include switching from binary to multinomial models. Students leaving their initial cluster challenges multilevel models to allow for more flexible cluster structures over time. Additionally, new information arising in future waves will lead to updates of the models used in previous waves' adjustments and will further allow finding predictors for variables of interest to be included in weighting models.

Besides unit nonresponse future research in the NEPS will also have to account for re-weighting panel attrition. Raghunathan, Patil, and Shope (2000) show that separating the contact and cooperation components in weighting adjustments substantially reduces nonresponse bias in their panel context.

Allowing the use of the panel entry weights multiple imputation for unit nonresponse in complex survey designs will surely be a topic in future research already applied for example by Rässler and Schnell (2003) or Peytchev (2012). Lately this topic has been suggested by Little (2013). The quality of multiple imputations will be limited when applied in sample weighting adjustments by information available on respondents and nonrespondents and therefore suffer the same problem of sparse weakly correlated variables as weighting

adjustments. In contrast the richer information arising in the progress of a panel will enrich the possibilities to allow for multiple imputations of unit nonresponse emerging as temporary (wave-specific) or final drop-out (no information after a certain wave). Respecting unit nonresponse patterns in multiple imputation would then allow using the unit nonresponse adjusted design weights throughout the entire waves of a panel making wave-specific or longitudinal adjustments of weights obsolete. For a further discussion of wether to use weighting or imputation for compensation of nonresponse in panel surveys see Kalton (1986).

And lastly user needs will drive future research (not only) related to weighting adjustments since the NEPS is set up as a project to provide a research infrastructure for educational longitudinal research.

References

Adler, D., & Murdoch, D. (2010). *rgl: 3D visualization device system (OpenGL).* Retrieved 03.06.2014, from `http://cran.r-project.org/package=rgl`

Aliaga, A., & Ren, R. (2006). *Optimal Sample Sizes for Two-stage Cluster Sampling in Demographic and Health Surveys: DHS Working Papers No. 30.* Retrieved 17.01.2011, from `http://dhsprogram.com/pubs/pdf/WP30/WP30.pdf`

Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., Rohwer, G., Rässler, S., & Blossfeld, H.-P. (2011). Sampling designs of the National Educational Panel Study: challenges and solutions. In H.-P. Blossfeld, H. G. Roßbach, & J. v. Maurice (Eds.), *Education as a lifelong process: Zeitschrift für Erziehungswissenschaft* (Vol. 14, pp. 51–65). Wiesbaden: VS Verlag für Sozialwissenschaften. doi: 10.1007/s11618-011-0181-8

Aßmann, C., Steinhauer, H. W., & Rässler, S. (2012). Aspekte der Stichprobenziehung in der erziehungswissenschaftlichen Forschung. In S. Maschke & L. Stecher (Eds.), *Enzyklopädie der Erziehungswissenschaft Online (EEO), Fachgebiet Methoden der empirischen erziehungswissenschaftlichen Forschung* (pp. 1–15). Weinheim und Basel: Beltz Juventa. doi: 10.3262/EEO07120215

Baker, R., Brick, J. M., Bates, N. A., Battaglia, M., Couper, M. P., Dever, J. A., Gile, K. J., & Tourangeau, R. (2013). Summary Report of the AAPOR Task Force on Non-probability Sampling. *Journal of Survey Statistics and Methodology*, *1*(2), 90–143. doi: 10.1093/jssam/smt008

Baltes, P. B., Reese, H. W., & Lipsitt, L. P. (1980). Life-span developmental psychology. *Annual Review of Psychology*, *31*, 65–110.

Bates, D., Maechler, M., & Bolker, B. (2012). *lme4: Linear mixed-effects models using S4 classes.* Retrieved 03.06.2014, from `http://CRAN.R-project.org/package=lme4`

Baumert, J., et al. (Eds.). (2002). *PISA 2000 - die Länder der Bundesrepublik Deutschland im Vergleich: [PISA-E].* Opladen: Leske + Budrich.

Biemer, P. P., Chen, P., & Wang, K. (2013). Using level-of-effort paradata in non-response adjustments with application to field surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *176*(1), 147–168. doi: 10.1111/j.1467-985X.2012.01058.x

Blossfeld, H.-P., Maurice, J., & Schneider, T. (2011). The National Educational Panel Study: need, main features, and research potential. In H.-P. Blossfeld, H. G. Roßbach, & J. v. Maurice (Eds.), *Education as a lifelong process: Zeitschrift für Erziehungswissenschaft* (Vol. 14, pp. 5–17). Wiesbaden: VS Verlag für Sozialwissenschaften. doi: 10.1007/s11618-011-0178-3

Blossfeld, H.-P., & Maurice, J. v. (2011). Education as a lifelong process. In H.-P. Blossfeld, H. G. Roßbach, & J. v. Maurice (Eds.), *Education as a lifelong process: Zeitschrift für Erziehungswissenschaft* (Vol. 14, pp. 19–34). Wiesbaden: VS Verlag für Sozialwissenschaften. doi: 10.1007/s11618-011-0179-2

Brick, J. M. (2013). Unit Nonresponse and Weighting Adjustments: A Critical Review. *Journal of Official Statistics*, *29*(3), 329–353.

Brick, J. M., & Kalton, G. (1996). Handling missing data in survey research. *Statistical Methods in Medical Research*, *5*(3), 215–238. doi: 10.1177/096228029600500302

Butler, J. S., & Moffitt, R. (1982). A Computationally Efficient Quadrature Procedure for the One-Factor Multinomial Probit Model. *Econometrica*, *50*(3), 761–764.

Cochran, W. (1977). *Sampling techniques* (3rd ed.). New York: Wiley.

Couper, M. P. (1998). Measuring Survey Quality in a CASIC Environment. In American Statistical Association (Ed.), *Proceedings of the Survey Research Methods Section* (Vol. 48, pp. 41–49). Retrieved 27.05.2013, from `http://www.amstat.org/sections/srms/Proceedings/papers/1998_006.pdf`

Dahl, D. B. (2012). *xtable: Export tables to LaTeX or HTML.* Retrieved 03.06.2014, from `http://CRAN.R-project.org/package=xtable`

Deming, W. E., & Stephan, F. F. (1940). On a Least Squares Adjustment of a Sampled Frequency Table When the Expected Marginal Totals are Known. *The Annals of Mathematical Statistics*, *11*(4), 427–444. doi: 10.1214/aoms/1177731829

Deville, J.-C., & Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. *Journal of the American Statistical Association*, *87*(418), 376–382.

Dragulescu, A. A. (2013). *xlsx: Read, write, format Excel 2007 and Excel 97/2000/XP/2003 files.* Retrieved 03.06.2014, from `http://CRAN.R-project.org/package=xlsx`

Durrant, G. B., & Steele, F. (2009). Multilevel modelling of refusal and non-contact in household surveys: evidence from six UK Government surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(2), 361–381. doi: 10.1111/j.1467-985X.2008.00565.x

Elder, G. H., Johnson, M. K., & Crosnoe, R. (2003). The Emergence and Development of Life Course Theory. In J. T. Mortimer & M. J. Shanahan (Eds.), *Handbook of the Life Course* (pp. 3–19). New York: Kluwer Academic/Plenum Publishers.

Elff, M. (2012). *memisc: Tools for Management of Survey Data, Graphics, Programming, Statistics, and Simulation.* Retrieved 03.06.2014, from `http://CRAN.R-project.org/package=memisc`

Elliott, M. R. (2009). Combining Data from Probability and Non-Probability Samples Using Pseudo-Weights. *Survey Practice*, *2*(6), 1–7. Retrieved 14.04.2014, from `http://surveypractice.org/index.php/SurveyPractice/article/view/185`

Esbensen, F.-A., Miller, M. H., Taylor, T., He, N., & Freng, A. (1999). Differential Attrition Rates and Active Parental Consent. *Evaluation Review*, *23*(3), 316–335. doi: 10.1177/0193841X9902300304

Gabler, S., Ganninger, M., & Münnich, R. (2012). Optimal allocation of the sample size to strata under box constraints. *Metrika*, *75*(2), 151–161.

Gambino, J. G. (2012). *pps: Functions for PPS sampling.* Retrieved 03.06.2014, from `http://CRAN.R-project.org/package=pps`

Genz, A., & Bretz, F. (2009). *Computation of multivariate normal and t probabilities*. Berlin: Springer.

Genz, A., Bretz, F., Miwa, T., mi, X., Leisch, F., Scheipl, F., & Hothorn, T. (2013). *mvtnorm: Multivariate Normal.* Retrieved 03.06.2014, from `http://CRAN.R-project.org/package=mvtnorm`

Geweke, J. (1989). Bayesian Inference in Econometric Models Using Monte Carlo Integration. *Econometrica*, *57*(6), 1317–1339.

Geweke, J. (1991). Efficient Simulation from the Multivariate Normal and Student-t Distributions Subject to Linear Constraints and the Evaluation of Constraint Probabilities. In American Statistical Association (Ed.), *Computing Science and Statistics: Proceedings of the 23rd Symposium on the Interface* (pp. 571–578). Alexandria.

Geweke, J., & Keane, M. (2001). Chapter 56: Computationally intensive methods for integration in econometrics. In J.J. Heckman & E.E. Leamer (Eds.), *Handbook of Econometrics* (Vol. 5, pp. 3463–3568). Elsevier. doi: 10.1016/S1573-4412(01)05009-7

Gouriéroux, C., Holly, A., & Monfort, A. (1982). Likelihood Ratio Test, Wald Test, and Kuhn-Tucker Test in Linear Models with Inequality Constraints on the Regression Parameters. *Econometrica*, *50*(1), 63–

80.

Greene, W. (2004). Convenient estimators for the panel probit model: Further results. *Empirical Economics*, *29*(1), 21–47. doi: 10.1007/s00181-003-0187-z

Greene, W. (2012). *Econometric analysis* (7th ed.). Boston: Pearson.

Groves, R. M. (1998). *Nonresponse in household interview surveys*. New York: Wiley.

Groves, R. M., & Heeringa, S. G. (2006). Responsive design for household surveys: tools for actively controlling survey errors and costs. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(3), 439–457. doi: 10.1111/j.1467-985X.2006.00423.x

Hájek, J., & Dupač, V. (1981). *Sampling from a finite population*. New York: M. Dekker.

Hajivassiliou, V. (1990). *Smooth Simulation Estimation of Panel Data LDV Models*. Unpublished doctoral dissertation, Department of Economics, Yale University.

Harvey, A. C. (1989). *Forecasting, structural time series models, and the Kalman filter*. Cambridge and New York: Cambridge University Press.

Hawkes, D., & Plewis, I. (2006). Modelling non-response in the National Child Development Study. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(3), 479–491. doi: 10.1111/j.1467-985X.2006.00401.x

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica*, *47*(1), 153–161.

Henry, K., & Valliant, R. (2012). Comparing Alternative Weight Adjustment Methods. In American Statistical Association (Ed.), *Proceedings of the Survey Research Methods Section* (pp. 4696–4710). Alexandria. Retrieved 03.06.2014, from `http://www.amstat.org/sections/srms/proceedings/y2012/files/306157_76012.pdf`

Holt, D., & Elliot, D. (1991). Methods of Weighting for Unit Non-Response. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *40*(3), 333–342.

Honaker, J., Owen, M., Imai, K., Lau, O., & King, G. (2013). *bprobit: Bivariate Probit Regression for Two Dichotomous Dependent Variables.* Retrieved 21.01.2014, from `http://cran.r-project.org/web/packages/ZeligChoice/vignettes/ZeligChoice-manual.pdf`

Horvitz, D. G., & Thompson, D. J. (1952). A Generalization of Sampling Without Replacement From a Finite Universe. *Journal of the American Statistical Association*, *47*(260), 663–685.

Iannacchione, V. G. (2003). Sequential Weight Adjustments for Location and Cooperation Propensity for the 1995 National Survey of Family

Growth. *Journal of Official Statistics*, *19*(1), 31–43.

Joncas, M. (2008). TIMSS 2007 Sampling Weights and Participation Rates. In J. F. Olson, M. O. Martin, & I. V. Mullis (Eds.), *TIMSS 2007 technical report* (pp. 153–192). Chestnut Hill: TIMSS & PIRLS International Study Center Lynch School of Education, Boston College.

Jones, O., Maillardet, R., & Robinson, A. (2009). *Introduction to scientific programming and simulation using R.* Boca Raton: CRC Press.

Kalton, G. (1986). Handling Wave Nonresponse in Panel Surveys. *Journal of Official Statistics*, *2*(3), 303–314.

Kalton, G., & Flores-Cervantes, I. (2003). Weighting methods. *Journal of Official Statistics*, *19*(2), 81–97.

Kalton, G., & Kasprzyk, D. (1986). The treatment of missing survey data. *Survey Methodology*, *12*, 1–16.

Kauermann, G., & Küchenhoff, H. (2011). *Stichproben: Methoden und praktische Umsetzung mit R.* Heidelberg: Springer.

Keane, M. P. (1994). A Computationally Practical Simulation Estimator for Panel Data. *Econometrica*, *62*(1), 95–116.

Kim, J. K., & Kim, J. J. (2007). Nonresponse Weighting Adjustment Using Estimated Response Probability. *The Canadian Journal of Statistics*, *35*(4), 501–514.

Kish, L. (1990). Weighting: Why, When, and How? In American Statistical Association (Ed.), *Proceedings of the Survey Research Methods Section* (pp. 121–130). Retrieved 03.06.2014, from `https://www.amstat.org/sections/SRMS/Proceedings/papers/1990_018.pdf`

Kish, L. (1992). Weighting for Unequal Pi. *Journal of Official Statistics*, *8*(2), 183–200.

Kish, L. (1995). *Survey sampling.* New York: Wiley.

Kreuter, F., & Olson, K. (2011). Multiple Auxiliary Variables in Nonresponse Adjustment. *Sociological Methods & Research*, *40*(2), 311–332. doi: 10.1177/0049124111400042

Kreuter, F., Olson, K., Wagner, J., Yan, T., Ezzati-Rice, T. M., Casas-Cordero, C., Lemay, M., Peytchev, A., Groves, R. M., & Raghunathan, T. E. (2010). Using proxy measures and other correlates of survey outcomes to adjust for non-response: examples from multiple surveys. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *173*(2), 389–407. doi: 10.1111/j.1467-985X.2009.00621.x

Kreuter, F., & Valliant, R. (2007). A survey on survey statistics: What is done and can be done in Stata. *The Stata Journal*, *7*(1), 1–21.

Kultusministerkonferenz. (2012). *Definitionenkatalog zur Schulstatistik 2012.* Retrieved 06.05.2014, from `http://www.kmk.org/fileadmin/pdf/Statistik/Defkat_2012.2_m_Anlagen.pdf`

Laaksonen, S. (2005). Does the choice of link function matter in response propensity modelling? *Model Assisted Statistics and Applications*, *1*(2), 95–100.

Lepkowski, J. M. (1989). Treatment of Wave Nonresponse in Panel Surveys. In D. Kasprzyk, G. J. Duncan, G. Kalton, & M. Singh (Eds.), *Panel surveys* (pp. 348–374). New York: Wiley.

Lepkowski, J. M., & Couper, M. P. (2002). Nonresponse in the second wave of longitudinal household surveys. In R. M. Groves (Ed.), *Survey nonresponse* (pp. 259–272). New York: Wiley.

Little, R. (2004). To Model or Not To Model? Competing Modes of Inference for Finite Population Sampling. *Journal of the American Statistical Association*, *99*(466), 546–556. doi: 10.1198/016214504000000467

Little, R. (2013). Discussion: Unit Nonresponse and Weighting Adjustments: A Critical Review. *Journal of Official Statistics*, *29*(3), 363–366.

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

Little, R., & Vartivarian, S. (2003). On weighting the rates in non-response weights. *Statistics in Medicine*, *22*(9), 1589–1599. doi: 10.1002/sim.1513

Little, R., & Vartivarian, S. (2005). Does Weighting for Nonresponse Increase the Variance of Survey Means? *Survey Methodology*, *31*(2), 161–168.

Lohr, S. L. (2010). *Sampling: Design and analysis* (2nd ed.). Boston: Cengage Learning.

Lugtig, P. (21.10.2013). *A great lecturer, and the contextuality of nonresponse.* Weblog and personal homepage. Retrieved 22.10.2013, from `http://www.peterlugtig.com/2013/10/a-great-lecturer-and-contextuality-of.html`

Maaz, K., Kreuter, F., & Watermann, R. (2006). Schüler als Informanten? Die Qualität von Schülerangaben zum sozialen Hintergrund. In J. Baumert, P. Stanat, & R. Watermann (Eds.), *Herkunftsbedingte Disparitäten im Bildungswesen: Differenzielle Bildungsprozesse und Probleme der Verteilungsgerechtigkeit* (pp. 31–59). Wiesbaden: VS Verlag für Sozialwissenschaften. doi: 10.1007/978-3-531-90082-7\textunderscore2

Madow, W. G. (1949). On the Theory of Systematic Sampling, II. *The Annals of Mathematical Statistics*, *20*(3), 333–354.

Madow, W. G. (1953). On the Theory of Systematic Sampling, III. Comparison of Centered and Random Start Systematic Sampling. *The Annals of Mathematical Statistics*, *24*(1), 101–106.

Madow, W. G., & Madow, L. H. (1944). On the Theory of Systematic Sampling, I. *The Annals of Mathematical Statistics*, *15*(1), 1–24.

Martin, M. O., Mullis, I. V. S., & Kennedy, A. M. (2007). *PIRLS 2006 technical report*. Chestnut Hill: TIMSS & PIRLS, International Study Center, Lynch School of Education, Boston College.

Mehrotra, P. C., Srivastava, A. K., & Tyagi, K. K. (1987). On Unequal Cluster Sampling for Fixed Sample Size. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *36*(4), 385–391.

Meixner, S., Schiller, D., Maurice, J. v., & Engelhardt-Wölfler, H. (2011). Data protection issues in the National Educational Panel Study. In H.-P. Blossfeld, H. G. Roßbach, & J. v. Maurice (Eds.), *Education as a lifelong process: Zeitschrift für Erziehungswissenschaft* (Vol. 14, pp. 301–313). Wiesbaden: VS Verlag für Sozialwissenschaften. doi: 10.1007/s11618-011-0191-6

Meyer, D., Zeileis, A., & Hornik, K. (2006). The Strucplot Framework: Visualizing Multi-way Contingency Tables with vcd. *Journal of Statistical Software*, *17*(3), 1–48.

Meyer, D., Zeileis, A., & Hornik, K. (2013). *vcd: Visualizing Categorical Data.* Retrieved 03.06.2014, from `http://CRAN.R-project.org/package=vcd`

Münnich, R. T. (2008). Varianzschätzung in komplexen Erhebungen. *Austrian Journal of Statistics*, *37*(3), 319–334.

Münnich, R. T., Sachs, E. W., & Wagner, M. (2012). Numerical solution of optimal allocation problems in stratified sampling under box constraints. *AStA Advances in Statistical Analysis*, *96*(3), 435–450. doi: 10.1007/s10182-011-0176-z

Nicoletti, C., & Peracchi, F. (2005). Survey response and survey characteristics: microlevel evidence from the European Community Household Panel. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *168*(4), 763–781. doi: 10.1111/j.1467-985X.2005.00369.x

OECD. (2012). *PISA 2009 Technical Report*. Paris: OECD Publishing.

Olson, J. F., Martin, M. O., & Mullis, I. V. (Eds.). (2008). *TIMSS 2007 technical report*. Chestnut Hill: TIMSS & PIRLS International Study Center Lynch School of Education, Boston College.

O'Muircheartaigh, C., & Campanelli, P. (1999). A Multilevel Exploration of the Role of Interviewers in Survey Non-Response. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, *162*(3), 437–446.

Owen, M., Imai, K., Lau, O., & King, G. (2012). *ZeligChoice: Zelig Choice Models.* Retrieved 03.06.2014, from `http://CRAN.R-project.org/package=ZeligChoice`

Peytchev, A. (2012). Multiple Imputation for Unit Nonresponse and Measurement Error. *Public Opinion Quarterly*, *76*(2), 214–237. doi: 10.1093/poq/nfr065

Pike, G. R. (2008). Using Weighting Adjustments to Compensate for Survey Nonresponse. *Research in Higher Education*, *49*(2), 153–171. doi: 10.1007/s11162-007-9069-0

Porter, S. R., & Whitcomb, M. E. (2005). Non-response in student surveys: The Role of Demographics, Engagement and Personality. *Research in Higher Education*, *46*(2), 127–152. doi: 10.1007/s11162-004-1597-2

R Core Team. (2013). *foreign: Read Data Stored by Minitab, S, SAS, SPSS, Stata, Systat, Weka, dBase, ..* Retrieved 03.06.2014, from `http://CRAN.R-project.org/package=foreign`

R Core Team. (2014). *R: A Language and Environment for Statistical Computing.* Vienna, Austria. Retrieved from `http://www.R-project.org/`

Raghunathan, T., Patil, S., & Shope, J. T. (2000). *Weighting Adjustments for Attrition: An Evaluation* (No. 093). Michigan. Retrieved 22.01.2014, from `http://www.isr.umich.edu/src/smp/Electronic%20Copies/93.pdf`

Rässler, S., & Riphahn, R. (2006). Survey Item Nonresponse and its Treatment. In O. Hübler (Ed.), *Modern econometric analysis* (pp. 215–230). Berlin: Springer. doi: 10.1007/3-540-32693-6\textunderscore15

Rässler, S., & Schnell, R. (2003). *Multiple Imputation for Unit-Nonresponse versus Weighting including a comparison with a Nonresponse Follow-Up Study.* Retrieved 19.09.2012, from `http://www.statistik.wiso.uni-erlangen.de/forschung/d0065.pdf`

Rendtel, U., & Harms, T. (2009). Weighting and Calibration for Household Panels. In P. Lynn (Ed.), *Methodology of longitudinal surveys* (pp. 265–286). Chichester: Wiley.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41–55. doi: 10.1093/biomet/70.1.41

Rubin, D. B. (1987). *Multiple Imputation for Nonresponse in Surveys.* New York: John Wiley & Sons.

Rust, K., & Johnson, E. G. (1992). Chapter 2: Sampling and Weighting in the National Assessment. *Journal of Educational and Behavioral Statistics*, *17*(2), 111–129. doi: 10.3102/10769986017002111

Rust, K., Krawchuk, S., & Monseur, C. (2013). PISA Student Nonresponse Adjustment Procedures. In M. Prenzel (Ed.), *Research on PISA* (pp. 87–102). New York: Springer. doi: 10.1007/978-94-007-4458-5\textunderscore6

Särndal, C.-E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, *33*(2), 99–119.

Särndal, C.-E., Swensson, B., & Wretman, J. H. (2003). *Model assisted*

survey sampling. New York: Springer.

Schaich, E., & Münnich, R. T. (2001). *Mathematische Statistik für Ökonomen*. München: Vahlen.

Schnell, R., Gramlich, T., Bachteler, T., Reiher, J., Trappmann, M., Smid, M., & Becher, I. (2013). Ein neues Verfahren für namensbasierte Zufallsstichproben von Migranten. *Methoden-Daten-Analysen*, *7*(1), 5–33. doi: 10.12758/mda.2013.001

Skinner, C. J., & D'Arrigo, J. (2011). Inverse probability weighting for clustered nonresponse. *Biometrika*, *98*(4), 953–966. doi: 10.1093/biomet/asr058

Snijders, T. A. B., & Bosker, R. J. (2012). *Multilevel analysis: An introduction to basic and advanced multilevel modeling* (2nd ed.). Los Angeles: Sage.

Steele, F., & Durrant, G. B. (2011). Alternative Approaches to Multilevel Modelling of Survey Non-Contact and Refusal. *International Statistical Review*, *79*(1), 70–91. doi: 10.1111/j.1751-5823.2011.00133.x

Steinhauer, H. W., Blossfeld, H.-P., & Maurice, J. v. (2012). Das Nationale Bildungspanel: Methodische Ansätze. In B. Dippelhofer-Stiem & S. Dippelhofer (Eds.), *Enzyklopädie der Erziehungswissenschaft Online* (pp. 1–20). Weinheim und Basel: Beltz Juventa. doi: 10.3262/EEO20120235

Swinton, J. (2009). *The xtable gallery*. Retrieved 25.05.2012, from `http://cran.r-project.org/web/packages/xtable/vignettes/xtableGallery.pdf`

Tillé, Y. (1996). Some Remarks on Unequal Probability Sampling Designs without Replacement. *Annals of Economics and Statistics / Annales d'Économie et de Statistique*(44), 177–189.

Tillé, Y. (2006). *Sampling algorithms*. New York: Springer.

Train, K. (2009). *Discrete choice methods with simulation* (2nd ed.). Cambridge: Cambridge University Press.

Trivellato, U. (1999). Issues in the Design and Analysis of Panel Studies: A Cursory Review. *Quality & Quantity*, *33*(3), 339–352.

Uthayakumaran, N. (1998). Additional Circular Systematic Sampling Methods. *Biometrical Journal*, *40*(4), 467–474.

Valliant, R., Dever, J. A., & Kreuter, F. (2013). *Practical tools for designing and weighting survey samples*. New York, Heidelberg: Springer.

Wagner, S. M., Rau, C., & Lindemann, E. (2010). Multiple Informant Methodology: A Critical Review and Recommendations. *Sociological Methods & Research*, *38*(4), 582–618. doi: 10.1177/0049124110366231

Wolter, K. (2007). *Introduction to Variance Estimation*. New York: Springer.

Wood, A. M., White, I. R., & Hotopf, M. (2006). Using number of failed contact attempts to adjust for non-ignorable non-response. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *169*(3), 525–542. doi: 10.1111/j.1467-985X.2006.00405.x

Xie, Y. (2012). *knitr: A general-purpose package for dynamic report generation in R.* Retrieved 2012.09.18, from `http://CRAN.R-project.org/package=knitr`

Yuan, Y., & Little, R. (2007). Model-based estimates of the finite population mean for two-stage cluster samples with unit non-response. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *56*(1), 79–97. doi: 10.1111/j.1467-9876.2007.00566.x

Appendix A

List of Abbreviations and Nomenclature

List of Abbreviations

ABias	Average bias
AGS	Amtlicher Gebietsschlüssel
AIC	Akaike Information Criterion
AMSE	Average mean square error
ASE	Average standard error
BDSG	Bundesdatenschutzgesetz
BHPS	British Household Panel Study
BIC	Bayesian Information Criterion
CATI	Computer Assisted Telephone Interview
CI	Confidence interval
CIRP	Cooperative Institutional Research Program
EFS	Expenditure and Food Survey
FRS	Family Resources Survey
FW	School type / stratum Freie Waldorfschule
GHK	Geweke, Hajivassiliou, Keane
GHS	General Household Survey
GREG	General Regression
GS	School type Grundschule
GG	Grundgesetz
GY	School type / stratum Gymnasium
HS	School type / stratum Hauptschule
HT	Horvitz-Thompson
ID	Identifier

IEA	International Association for the Evaluation of Educational Achievement
IG	School type / stratum Integrierte Gesamtschule
LFS	Labour Force Survey
LR	Likelihood ratio
MAR	Missing at random
MB	School type / stratum Schule mit mehreren Bildungsgängen
MCAR	Missing completely at random
MIG	Migrants supplement
mos	Measure of size
MOS	Total measure of size
N5	Stratum schools providing classes in grade 5 but none in 9
NCDS	National Child Development Study
NMAR	Not missing at random
NEAP	National Assessment of Educational Progress
NEPS	National Educational Panel Study
NSSE	National Survey of Student Engagement
NTS	National Travel Survey
OECD	Organisation for Economic Co-operation and Development
PIRLS	Progress in International Reading Literacy Study
PISA	Programme for International Student Assessment
pps	probability proportional to size
PSU	Primary sampling units
RS	School type / stratum Realschule
SC	Starting cohort
SE	Standard error
srs	Simple random sampling
SSU	Secondary sampling units
SU	School type Schulartunabhängige Orientierungsstufe
TIMSS	Trends in International Mathematics and Science Study
UK	United Kingdom

Nomenclature

(S, p)	Sample design
α	Random intercept
β	Regression coefficients
χ^2	Chi square distribution
Δ	Set of all possible participation patterns
δ	Adjustment factor
ϵ	Disturbance
$\forall$	for all
$\gamma_{\mathcal{L}^{(s)}}$	Lower truncation
$\gamma_{\mathcal{U}^{(s)}}$	Upper truncation
λ_i	Response propensity for element i
$\mathcal{I}$	Indicator function
$\mathcal{L}$	Likelihood
μ	Mean
Ω	Covariance Matrix
ω	Standard deviation
Φ	Distribution function of the normal distribution
ϕ	Density function of the normal distribution
Φ_2	Distribution function of the bivariate normal distribution

ϕ_2	Density function of the bivariate normal distribution
π	Inclusion probability
ρ	Correlation
Σ	Covariance Matrix
σ	Standard deviation
τ	General estimator
τ	Test group size
θ	Parameter vector
ε	Error term
ϱ	Covariance
$\widetilde{v}$	Trajectory of draws from a normal distribution
$\widetilde{y}$	Latent variable
C	Cluster (e.g., schools or classes)
c	Couple of a student s and a parent p
C_j	Number of classes in school j
d	Design weight
D_i^c	Integration regions for individual i
D_{ij}^c	Integration regions for individual i in cluster j
df	Degrees of freedom
F	Uniform distributed random numbers
f	Index for a weighting cell
$f(v)$	Regular density
$g(v)$	Smooth function
H	Number of strata
h	Index for strata

I Value of an integral

i Index for individuals

K Number of dimensions for the multivariate normal distribution

k Interval length

k Number of parameters

L Cholesky decomposition of Σ

l Index for columns

M Number of clusters in the population

m Number of clusters in the sample

MOS Total measure of size

MOS Total measure of size

mos Measure of size

mos Measure of size

MOS_i Cumulative measure of size up to element i

mos_i Measure of size for element i

N Population size

n Sample size

$N(\mu, \sigma)$ Normal distribution with mean μ and standard error σ

n_0 Minimum net sample size

N_h Stratum-specific population size

n_h Stratum-specific sample size

n_j Number of elements i in cluster j

n_{net} Net sample size

P Probability

p Parent

p	Participation rate
p	p-distribution
R	Number of random draws
R	Number of replications
r	Random start
S	Set of samples
S	Simulation sample size for the GHK-simulator
s	Index for simulation sample size
s	Sample
s	Student
S_j	Number of students in school j
T	Total number of waves in the panel
t	Number of test groups
t_0	Maximum number of test groups
U	Universe / Population
u_i	Unit / Element i
w	Adjusted design weight
x, X	Auxiliary information
y, Y	Characteristic of interest
y_{ij}	Participation status of student i in cluster j

Appendix B

Tables

To not disturb the reading of this thesis some of the tables mentioned in the text are put separately in this section of the appendix. Tables are given in order of appearance within the text.

Table B.1: Distributions for net sample sizes n_{net} for different participation rates p by strata when sampling $m^I = 480$ PSUs.

h	x_{Min}	$x_{0.05}$	μ_x	$x_{0.95}$	x_{Max}	σ_x
			Distribution of net sample size n_{net} for $p = 0.50$			
GY	4018.62	4063.32	4151.31	4226.93	4254.63	50.49
HS	1727.99	1831.08	1936.19	2024.61	2071.76	59.70
IG	971.61	1025.78	1102.16	1197.33	1256.84	52.24
MB	635.16	666.48	729.97	783.57	806.45	36.15
RS	2697.52	2744.16	2827.16	2925.91	2980.97	55.23
Sample	10365.51	10559.20	10746.78	10933.08	11085.83	113.42
			Distribution of net sample size n_{net} for $p = 0.55$			
GY	4420.49	4469.65	4566.44	4649.62	4680.09	55.54
HS	1900.79	2014.19	2129.81	2227.07	2278.94	65.67
IG	1068.77	1128.36	1212.38	1317.06	1382.52	57.46
MB	698.67	733.13	802.96	861.92	887.10	39.77
RS	2967.27	3018.57	3109.87	3218.50	3279.06	60.76
Sample	11402.06	11615.12	11821.46	12026.39	12194.42	124.76
			Distribution of net sample size n_{net} for $p = 0.60$			
GY	4822.35	4875.98	4981.57	5072.31	5105.55	60.59
HS	2073.59	2197.30	2323.43	2429.53	2486.12	71.64
IG	1165.93	1230.93	1322.59	1436.80	1508.21	62.69
MB	762.19	799.78	875.96	940.28	967.74	43.38
RS	3237.02	3292.99	3392.59	3511.09	3577.16	66.28
Sample	12438.61	12671.04	12896.14	13119.69	13303.00	136.10
			Distribution of net sample size n_{net} for $p = 0.65$			
GY	5224.21	5282.31	5396.70	5495.00	5531.01	65.64
HS	2246.39	2380.41	2517.05	2632.00	2693.29	77.61
IG	1263.10	1333.51	1432.81	1556.53	1633.89	67.91
MB	825.71	866.43	948.96	1018.64	1048.38	47.00
RS	3506.77	3567.40	3675.30	3803.69	3875.26	71.80
Sample	13475.16	13726.96	13970.82	14213.00	14411.58	147.44
			Distribution of net sample size n_{net} for $p = 0.70$			
GY	5626.07	5688.65	5811.83	5917.69	5956.48	70.69
HS	2419.19	2563.51	2710.67	2834.46	2900.47	83.58
IG	1360.26	1436.09	1543.02	1676.26	1759.57	73.13
MB	889.22	933.08	1021.96	1096.99	1129.03	50.61
RS	3776.52	3841.82	3958.02	4096.28	4173.35	77.32
Sample	14511.71	14782.88	15045.50	15306.31	15520.17	158.78
			Distribution of net sample size n_{net} for $p = 0.75$			
GY	6027.94	6094.98	6226.96	6340.39	6381.94	75.74
HS	2591.99	2746.62	2904.28	3036.92	3107.64	89.55
IG	1457.42	1538.66	1653.24	1795.99	1885.26	78.36
MB	952.74	999.73	1094.95	1175.35	1209.67	54.23
RS	4046.28	4116.23	4240.73	4388.87	4471.45	82.85
Sample	15548.27	15838.80	16120.17	16399.62	16628.75	170.12
			Distribution of net sample size n_{net} for $p = 0.80$			
GY	6429.80	6501.31	6642.09	6763.08	6807.40	80.79
HS	2764.79	2929.73	3097.90	3239.38	3314.82	95.52
IG	1554.58	1641.24	1763.46	1915.73	2010.94	83.58
MB	1016.25	1066.37	1167.95	1253.71	1290.32	57.84
RS	4316.03	4390.65	4523.45	4681.46	4769.55	88.37
Sample	16584.82	16894.72	17194.85	17492.92	17737.33	181.46

Table B.2: Schüler-Teilnahme-Liste / students participation list

Schülerliste NEPS-Haupterhebung K5/K9
Herbst/Winter 2010 Jg. 9

Schulname: Musterschule

Schul-ID: 255201
Klasse: 9c
Testgruppen-ID: 25520131

Haupterhebung K5/K9
Angaben zur Grundgesamtheit

Information im Klartext: Bundesland, Schulform, Jahrgang

Schülerliste NEPS-Haupterhebung K5/K9
Herbst/Winter 2010 Jg. 9

Schulname: Musterschule

Schul-ID: 255201
Klasse: 9c
Testgruppen-ID: 25520131

[1]	[2]	[3]	[4]	[5]	[6]		[7]	[8]	[9]	[10]	[11]	[12]	[13]	[14]	[15]	[16]		[17]		[18]		[19]
Lfd. Nr.	Name des Schülers (Nachname, Vorname)	Schüler-ID	Staatsangehörigkeit = deutsch ja/nein / deutsch	Deutsch als vorwiegend zu Hause gesprochene Sprache (ja/nein)	Noten aus dem letzten Zeugnis in Deutsch	Mathematik	Sonderstatus Förderbedarf (ja/nein)	Diagnostizierte Lese-/Rechtschreibschwäche (ja/nein)	Geschlecht (m/w)	Geburtsmonat und -jahr (MM.JJ)	Teilnahmebereitschaft (ja/nein)	Elternzustimmung (ja/nein)	Schulleitungseinwilligung (ja/nein)	Übertrag aus Spalte 9 Geschlecht (m/w)	Übertrag aus Spalte 10 Geburtsmonat und -jahr (MM.JJ)	Übertrag aus Spalte 8 Noten aus dem letzten Zeugnis in Deutsch	Mathematik	Instrumente-Nr. FB	TH	Teilnahmestatus Haupttest (TA/E) FB	TH	Nachtest (TA/E) FB	TH	Schüler-ID
001	*Akcan, Deniz*	25520119-001	2	nein	3	2	nein	nein	m	12.94	ja	ja	ja	m	12.94	3	2	1	1						25520119-001
002	*Bonngi, Claudia*	25520119-002	1	ja	2	2	nein	nein	w	03.95	ja	ja	ja	w	03.95	2	2	2	2						25520119-002
003	*Freese, Verena*	25520119-003	1	ja	5	3	nein	ja	w	08.95	nein	nein	ja					1	3						25520119-003
004	*Frisch, Lukas*	25520119-004	1	ja	4	4	nein	nein	m	11.94	nein	nein	nein					2	4						25520119-004
005	*Günther, Marzia*	25520119-005	1	ja	1	2	nein	nein	w	12.95	nein	ja	nein					1	1						25520119-005
006	*Halberstadt, Julius*	25520119-006	1	ja	2	3	nein	nein	m	02.95	nein	nein	nein					2	2						25520119-006
007	*Liebig, Sebastian*	25520119-007	1	ja	4	5	ja	ja	m	01.95	ja	ja	ja	m	01.95	4	5	1	3						25520119-007
008	*Landmann, Anne-Kathrin*	25520119-008	1	ja	3	1	nein	nein	w	02.95	ja	ja	ja	w	02.95	3	1	2	4						25520119-008
009	*Marx, Alexander*	25520119-009	1	ja	3	4	nein	nein	m	12.94	nein	nein	nein					1	1						25520119-009
010	*Muller, Renee*	25520119-010	2	ja	3	3	nein	nein	w	10.94	ja	ja	ja	w	10.94	3	3	2	2						25520119-010

Zum Verbleib in der Schule!

Versand im vorfrankierten und adressierten Umschlag an das DPC ohne Absenderinformation

Zur Aushändigung an den Testleiter!

Table B.3: Results of random intercept models for school participation (by strata).

	Strata						
	$N5$	GY	HS	IG	MB	RS	FS
Constant	−1.283	−1.065	−1.868***	−0.275	−0.799	−0.999	0.137
	(1.041)	(0.598)	(0.473)	(0.529)	(1.095)	(0.714)	(0.224)
Number of students	0.001	0.001	0.001	0.001	−0.002	−0.000	
in school	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)	(0.001)	
Number of students	0.010	−0.001	−0.001	−0.002	0.000	−0.004	
in age group	(0.020)	(0.002)	(0.002)	(0.003)	(0.005)	(0.003)	
Number of students							−0.002
in grade 9 (squared)							(0.021)
Organizing institution	0.762	0.127	1.043**	−0.063	−0.348	−0.012	
public	(0.662)	(0.207)	(0.372)	(0.377)	(0.668)	(0.304)	
Urbanization	−0.027	0.233	0.164	−0.179	−0.132	0.088	−0.341
rural	(0.661)	(0.287)	(0.168)	(0.459)	(0.365)	(0.295)	(0.311)
Urbanization	−0.790	−0.069	0.085	−0.242	−0.430	−0.215	−0.138
urban	(0.543)	(0.139)	(0.123)	(0.294)	(0.342)	(0.158)	(0.201)
Number of classes	−0.020	0.025	−0.038	−0.026	−0.107	−0.014	
in grade 5	(0.087)	(0.041)	(0.057)	(0.051)	(0.130)	(0.054)	
Number of classes		0.021	0.025	−0.079	0.137	0.108	
in grade 7		(0.069)	(0.032)	(0.048)	(0.078)	(0.096)	
Cohorts sampled							−0.082
five and nine							(0.183)
Effort in recruitment							0.708**
up to 8							(0.253)
Effort in recruitment		0.807***	0.435	1.011**	1.192**	0.295	
up to 51		(0.220)	(0.478)	(0.317)	(0.418)	(0.538)	
Random intercept							
ω Federal States	0.000	0.084	0.256	0.205	0.379	0.145	0.000
Schools per stratum	54	415	672	218	110	337	203

Notes: ***, **, and * denote significance at the 0.1%, 1%, and 5% level, respectively. Standard errors are given in parenthesis. To model school participation, the `glmer` function with a probit link provided by `lme4` package (Bates et al., 2012) in R (R Core Team, 2014) was used.

Table B.4: Results of random intercept model for the participation of schools contacted for the supplement of migrants. Standard deviations are given in parentheses.

	MIG
Constant	-1.538
	(0.798)
Number of students	0.001
in school	(0.001)
School type	-0.039
GY	(0.695)
School type	0.407
HS	(0.657)
School type	-0.368
IG	(0.711)
School type	0.951
MB	(0.737)
School type	-0.327
RS	(0.669)
Stratum	-0.400
Russian	(0.383)
Stratum	0.193
Turkish	(0.322)
Urbanization	-4.502
rural	(146.097)
Urbanization	-0.008
urban	(0.437)
Random intercept	
ω Federal States	0.472
Number of schools	198

Notes: ***, **, and * denote significance at the 0.1%, 1%, and 5% level, respectively. Standard errors are given in parenthesis. To model school participation, the `glmer` function with a probit link provided by `lme4` package (Bates et al., 2012) in R (R Core Team, 2014) was used.

Table B.5: Models estimating the individual participation propensity used to derive adjustment factors for sample weighting adjustment of the initial sample.

	SC3			SC4	
	regular schools	special schools	migrant supplement	regular schools	special schools
Intercept	0.416	1.395**	0.636*	0.922***	1.257***
	(0.255)	(0.469)	(0.296)	(0.176)	(0.309)
Age group	0.109*	0.196	−0.123	0.145***	0.316*
younger half	(0.048)	(0.184)	(0.118)	(0.034)	(0.139)
Age group	−0.140	0.220	0.271	−0.099*	0.003
older half	(0.106)	(0.208)	(0.192)	(0.048)	(0.121)
Gender	−0.033	−0.364*	−0.156	−0.172***	−0.276**
male	(0.046)	(0.159)	(0.104)	(0.032)	(0.104)
Language spoken at home	−0.293**	−0.987***	−0.438**	−0.128*	−0.238
no German	(0.095)	(0.287)	(0.136)	(0.064)	(0.194)
Test group size	−0.003	−0.042	0.008	−0.012*	−0.027*
in number of students	(0.008)	(0.022)	(0.006)	(0.006)	(0.013)
Competence or grade in math	0.275***	0.518**	0.209	0.153***	0.422**
1 to 3 (A to C)	(0.061)	(0.196)	(0.128)	(0.034)	(0.133)
Reading competence or grade in German	0.150*	−0.029	0.068	0.182***	−0.062
1 to 3 (A to C)	(0.064)	(0.190)	(0.130)	(0.037)	(0.139)
Special educational needs	0.043		−0.152	−0.649***	
yes	(0.178)		(0.276)	(0.151)	
Missing indicator	−3.826***			−3.166***	
personal characteristics	(0.503)			(0.228)	
Missing indicator	−0.587***	−0.327	0.737*	0.110	−0.344
migration characteristics	(0.178)	(0.703)	(0.309)	(0.124)	(0.389)
Missing indicator	−1.013***	−4.392***	−2.624***	−1.234***	−2.699***
competence characteristics	(0.173)	(1.257)	(0.350)	(0.226)	(0.406)
Migrational background			−1.054***		
Turkish			(0.190)		
Random effect					
ω school level	0.615	1.112	0.736	0.978	1.083
Sample size#	9081	962	845	23327	2146
Initial sample size	9458	1036	983	24229	2372

Notes: ***, ** and * denote significance at the 0.1%, 1% and 5% level, respectively. Standard errors are given in parenthesis. # Number of students providing valid participation consent forms. To model individual participation, the `glmer` function with a probit link provided by `lme4` package (Bates et al., 2012) in R (R Core Team, 2014) was used.

Table B.6: Models estimating the individual participation propensity used to derive adjustment factors for sample weighting adjustment of wave 1 and 2, respectively.

	Starting Cohort 3		Starting Cohort 4	
	Wave 1	Wave 2	Wave 1	Wave 2
Intercept	1.233***	3.901***	1.814***	2.079***
	(0.270)	(0.508)	(0.091)	(0.195)
Stratum $N5$ (SC3)	−0.368	−0.486		
	(0.299)	(0.524)		
Stratum FS (SC4)	0.117	−2.129***	−0.207*	−1.674***
	(0.302)	(0.450)	(0.088)	(0.171)
Stratum GY (SC4)	−0.298	−0.542		
	(0.278)	(0.450)		
Stratum HS (SC4)	−0.331	−0.560	−0.146*	−0.163
	(0.288)	(0.439)	(0.067)	(0.172)
Stratum IG (SC4)	−0.749*	−0.194	−0.108	0.152
	(0.310)	(0.705)	(0.082)	(0.250)
Stratum MB (SC4)	−0.636*	−0.649	−0.117	−0.120
	(0.302)	(0.528)	(0.094)	(0.248)
Stratum RS (SC4)	−0.192	−0.444	−0.070	−0.205
	(0.284)	(0.450)	(0.069)	(0.180)
Age group	−0.055	0.189	0.066	0.284***
younger half	(0.067)	(0.127)	(0.045)	(0.070)
Gender	0.061	0.184	−0.070	0.025
female	(0.063)	(0.110)	(0.038)	(0.064)
Migration background	−0.169	−0.571		
turkish	(0.312)	(0.492)		
Native language	1.140***	0.415**	0.433***	0.276**
German	(0.068)	(0.148)	(0.049)	(0.088)
Nationality			−0.169*	−0.001
German			(0.070)	(0.111)
Class size			−0.058	0.094
less than 25			(0.050)	(0.095)
Missing indicator for			−1.323***	−0.705***
migration characteristics			(0.074)	(0.126)
Missing indicator for	−1.148***	−1.143***	−2.259***	−0.056
personal characteristics	(0.116)	(0.196)	(0.329)	(0.477)
Student participating in		−0.392		0.566***
wave 1		(0.268)		(0.106)
Individual re-tracking in		−3.498***		
wave 2		(0.181)		
Random intercept				
ω school level	0.311	0.500	0.276	0.844
Sample size	6112	6098	16425	16425

Notes: ***, **, and * denote significance at the 0.1%, 1%, and 5% level, respectively. Standard errors are given in parenthesis. To model individual participation, the `glmer` function with a probit link provided by `lme4` package (Bates et al., 2012) in R (R Core Team, 2014) was used.

Table B.7: Number of cases (n) and proportion (p) for variables in models by wave.

	Wave 1		Wave 2	
	n	p	n	p
Participation status: participant				
students	5774	0.9447	5790	0.9473
parents	4151	0.6792	3820	0.6250
students & parents	3974	0.6502	3727	0.6098
Gender				
female	2895	0.4737	2895	0.4737
male	3146	0.5147	3146	0.5147
missing	71	0.0116	71	0.0116
Nationality				
German	5112	0.8364	5112	0.8364
other	344	0.0563	344	0.0563
missing	656	0.1073	656	0.1073
Native language				
German	5225	0.8549	5225	0.8549
other	717	0.1173	717	0.1173
missing	170	0.0278	170	0.0278
Number of calls				
four or More	2318	0.3793	2231	0.3650
less than four	2435	0.3984	2400	0.3927
no calls	1359	0.2223	1481	0.2423
Tracking status				
individual re-tracking	0	0.0000	444	0.0726
in school	6112	1.0000	5668	0.9274
Age group				
older half	3253	0.5322	3253	0.5322
younger half	2670	0.4368	2670	0.4368
missing	189	0.0309	189	0.0309
Sampling stratum				
MIG	242	0.0396	242	0.0396
N5	458	0.0749	458	0.0749
FS	587	0.0960	587	0.0960
GY	2372	0.3881	2372	0.3881
HS	677	0.1108	677	0.1108
IG	284	0.0465	284	0.0465
MB	352	0.0576	352	0.0576
RS	1140	0.1865	1140	0.1865
Migration background				
russian	68	0.0111	68	0.0111
turkish	174	0.0285	174	0.0285
missing	5870	0.9604	5870	0.9604
Missing indicator for				
personal characteristics	194	0.0317	226	0.0370

Table B.8: $\ln \mathcal{L}$, AIC and BIC for considered model specifications.

model specifications			$\ln \mathcal{L}$	information criteria AIC	BIC
		Wave 1			
no clustering					
separate	students	Ia	−1049.599	2125.197	2212.531
	parents	IIa	−3149.145	6320.290	6394.188
		Ia+IIa	−4198.744		
joint		IIIa	−4195.952	8441.904	8609.854
LR test	Ia+IIa vs. IIIa	χ^2	5.584*		
clustering					
separate	students	IVa	−1039.797	2107.595	2201.647
	parents	Va	−3125.542	6275.084	6355.700
		IVa+Va	−4165.339		
joint		VIa	−4161.778	8377.555	8558.942
LR test	IVa+Va vs. VIa	χ^2	7.122**		
		Wave 2			
no clustering					
separate	students	Ib	−526.426	1082.853	1183.623
	parents	IIb	−1985.893	3997.786	4085.120
		Ib+IIb	−2411.555		
joint		IIIb	−2483.226	4845.998	5047.538
LR test	Ib+IIb vs. IIIb	χ^2	143.342***		
clustering					
separate	students	IVb	−510.032	886.052	993.503
	parents	Vb	−1983.092	3961.509	4055.529
		IVb+Vb	−2493.124		
joint		VIb	−2465.790	4816.750	5031.653
LR test	IVb+Vb vs. VIb	χ^2	54.668***		

Notes: ***, **, and * denote significance at the 0.1%, 1%, and 5% level, respectively and $\chi^2(1) = 3.841$.

Table B.9: Alternative models estimating the individual participation propensity of students and parents for SC3 in wave 1.

	No clustering		Clustering	
	Students (Ia)	Parents (IIa)	Students (IVa)	Parents (Va)
Intercept	1.195***	−0.708***	1.234***	−0.753***
	(0.243)	(0.167)	(0.270)	(0.184)
Stratum $N5$	−0.361	0.584**	−0.367	0.643**
	(0.264)	(0.179)	(0.299)	(0.204)
Stratum FS	0.098	−0.021	0.116	0.001
	(0.271)	(0.175)	(0.303)	(0.196)
Stratum GY	−0.297	0.594***	−0.299	0.654***
	(0.248)	(0.169)	(0.278)	(0.189)
Stratum HS	−0.339	0.238	−0.331	0.266
	(0.256)	(0.174)	(0.288)	(0.196)
Stratum IG	−0.772**	0.472*	−0.750*	0.484*
	(0.265)	(0.185)	(0.310)	(0.218)
Stratum MB	−0.602*	0.169	−0.635*	0.218
	(0.266)	(0.181)	(0.302)	(0.208)
Stratum RS	−0.207	0.389*	−0.192	0.447*
	(0.253)	(0.171)	(0.284)	(0.193)
Migration background	−0.179	0.407*	−0.170	0.446*
turkish	(0.281)	(0.194)	(0.313)	(0.214)
Native language	1.099***	0.440***	1.140***	0.454***
German	(0.064)	(0.050)	(0.068)	(0.052)
Age group	−0.072		−0.055	
younger half	(0.063)		(0.067)	
Gender	0.057		0.061	
female	(0.060)		(0.063)	
Missing indicator for	−1.098***		−1.147***	
personal characteristics	(0.111)		(0.116)	
Number of calls		1.299***		1.324***
less than 4		(0.043)		(0.044)
Random intercept				
ω school level			0.311	0.261
$\ln \mathcal{L}$	−1049.599	−3149.145	−1039.797	−3125.542
AIC	2125.197	6320.290	2107.595	6275.084
BIC	2212.531	6394.188	2201.647	6355.700
Sample size	6112	6112	6112	6112

Notes: ***, **, and * denote significance at the 0.1%, 1%, and 5% level, respectively. Standard errors are given in parenthesis. To model individual participation, the `glmer` and `glm` functions with a probit link provided by `lme4` (Bates et al., 2012) and `stats` package in R (R Core Team, 2014) was used.

Table B.10: Results for the bivariate probit models without and with random intercept estimating the individual participation propensities for students and parents for SC3 in wave 1.

	Bivariate probit – no clustering		Bivariate probit – clustering	
	Parents	Students	Parents	Students
	(IIIa)		(VIa)	
Intercept	-0.709^{***}	1.188^{***}	-0.762^{***}	1.197^{***}
	(0.167)	(0.242)	(0.186)	(0.268)
Stratum $N5$	0.586^{**}	-0.352	0.657^{**}	-0.340
	(0.179)	(0.263)	(0.207)	(0.297)
Stratum FS	-0.020	0.105	0.010	0.139
	(0.175)	(0.270)	(0.197)	(0.299)
Stratum GY	0.596^{***}	-0.287	0.667^{***}	-0.271
	(0.169)	(0.247)	(0.191)	(0.276)
Stratum HS	0.239	-0.329	0.274	-0.297
	(0.174)	(0.255)	(0.198)	(0.286)
Stratum IG	0.473^{*}	-0.768^{**}	0.502^{*}	-0.747^{*}
	(0.185)	(0.264)	(0.221)	(0.305)
Stratum MB	0.170	-0.596^{*}	0.228	-0.613^{*}
	(0.181)	(0.265)	(0.210)	(0.301)
Stratum RS	0.390^{*}	-0.201	0.456^{*}	-0.164
	(0.171)	(0.252)	(0.195)	(0.283)
Migration background	0.409^{*}	-0.175	0.458^{*}	-0.157
turkish	(0.194)	(0.280)	(0.216)	(0.311)
Age group		-0.077		-0.062
younger half		(0.063)		(0.066)
Native language	0.440^{***}	1.099^{***}	0.452^{***}	1.132^{***}
German	(0.050)	(0.064)	(0.052)	(0.069)
Gender		0.061		0.064
female		(0.060)		(0.062)
Missing indicator for		-1.080^{***}		-1.120^{***}
personal characteristics		(0.111)		(0.119)
Number of calls	1.297^{***}		1.315^{***}	
less than 4	(0.043)		(0.044)	
Correlation	0.097^{*}		0.122^{**}	
ρ students parents	(0.049)		(0.044)	
Random intercept				
ω school level			0.261	0.302
$\ln \mathcal{L}$	-4195.952		-4161.778	
AIC	8441.904		8377.555	
BIC	8609.854		8558.942	
Sample size	6112.000		6112.000	

Notes: ***, **, and * denote significance at the 0.1%, 1%, and 5% level, respectively. Standard errors are given in parenthesis. To model individual participation decisions, the `zelig` function with `bprobit` link provided by `ZeligChoice` package (Owen, Imai, Lau, & King, 2012) in R (R Core Team, 2014) was used. Correlation parameter from the bivariate probit model without random intercept is transformed according to Honaker, Owen, Imai, Lau, and King (2013).

Table B.11: Alternative models estimating the individual participation propensity of students and parents for SC3 in wave 2.

	No clustering		Clustering	
	Students (Ib)	Parents (IIb)	Students (IVb)	Parents (Vb)
Intercept	3.214^{***}	-2.205^{***}	3.769^{***}	-2.251^{***}
	(0.40*S*2)	(0.236)	(0.489)	(0.242)
Stratum *N*5	−0.201	0.565^{*}	−0.345	0.587^{*}
	(0.377)	(0.231)	(0.517)	(0.240)
Stratum *FS*	-1.591^{***}	0.030	-1.861^{***}	0.047
	(0.352)	(0.228)	(0.438)	(0.235)
Stratum *GY*	−0.453	0.702^{**}	−0.624	0.731^{**}
	(0.354)	(0.220)	(0.437)	(0.227)
Stratum *HS*	−0.492	0.225	−0.533	0.230
	(0.348)	(0.225)	(0.436)	(0.234)
Stratum *IG*	−0.398	0.724^{**}	−0.530	0.755^{**}
	(0.446)	(0.239)	(0.590)	(0.252)
Stratum *MB*	−0.391	0.417	−0.489	0.443
	(0.405)	(0.235)	(0.517)	(0.245)
Stratum *RS*	−0.450	0.535^{*}	−0.632	0.557^{*}
	(0.351)	(0.222)	(0.441)	(0.230)
Migration background	−0.418	−0.049	−0.490	−0.060
turkish	(0.400)	(0.247)	(0.486)	(0.255)
Native language	0.246^{*}	0.121	0.335^{*}	0.117
German	(0.116)	(0.066)	(0.136)	(0.068)
Student participating in	−0.386	0.216^{*}	−0.483	0.229^{*}
wave 1	(0.214)	(0.102)	(0.252)	(0.103)
Age group	0.100		0.127	
younger half	(0.097)		(0.115)	
Gender	0.118		0.138	
female	(0.087)		(0.101)	
Missing indicator for	-0.999^{***}		-1.135^{***}	
personal characteristics	(0.155)		(0.181)	
Individual re-tracking in	-2.640^{***}		-3.197^{***}	
wave 2	(0.100)		(0.140)	
Number of calls		0.503^{***}		0.513^{***}
less than 4		(0.048)		(0.049)
Parent participating in		2.337^{***}		2.364^{***}
wave 1		(0.051)		(0.052)
Random intercept				
ω school level			0.533	0.173
$\ln \mathcal{L}$	−526.426	−1985.893	−510.032	−1983.092
AIC	1082.853	3997.786	1052.064	3994.184
BIC	1183.623	4085.120	1159.553	4088.236
Sample size	6112	6112	6112	6112

Notes: ***, **, and * denote significance at the 0.1%, 1%, and 5% level, respectively. Standard errors are given in parenthesis. To model individual participation, the `glmer` and `glm` functions with a probit link provided by `lme4` (Bates et al., 2012) and `stats` package in R (R Core Team, 2014) was used.

Table B.12: Results for the bivariate probit models without and with random intercept estimating the individual participation propensities for students and parents for SC3 in wave 2.

	Bivariate probit – no clustering		Bivariate probit – clustering	
	Parents	Students	Parents	Students
	(IIIb)		(VIb)	
Intercept	-2.163^{***}	3.018^{***}	-2.212^{***}	3.368^{***}
	(0.235)	(0.395)	(0.239)	(0.469)
Stratum $N5$	0.525^{*}	-0.262^{*}	0.554^{*}	-0.479
	(0.230)	(0.370)	(0.237)	(0.466)
Stratum FS	0.001	-1.590^{***}	0.023	-1.836^{***}
	(0.226)	(0.347)	(0.230)	(0.412)
Stratum GY	0.661^{**}	-0.439^{*}	0.696^{**}	-0.635
	(0.219)	(0.349)	(0.223)	(0.406)
Stratum HS	0.186^{*}	-0.519^{*}	0.198	-0.609
	(0.224)	(0.343)	(0.229)	(0.405)
Stratum IG	0.685^{**}	-0.332^{*}	0.717^{**}	-0.540
	(0.238)	(0.452)	(0.247)	(0.525)
Stratum MB	0.376^{*}	-0.403^{*}	0.411	-0.567
	(0.234)	(0.398)	(0.242)	(0.471)
Stratum RS	0.499^{*}	-0.461^{*}	0.524^{*}	-0.709
	(0.221)	(0.346)	(0.226)	(0.408)
Migration background	-0.081	-0.376^{*}	-0.088	-0.517
turkish	(0.246)	(0.396)	(0.250)	(0.452)
Age group		0.083^{*}		0.092
younger half		(0.096)		(0.101)
Native language	0.122^{*}	0.243^{*}	0.119	0.296^{*}
German	(0.066)	(0.115)	(0.067)	(0.124)
Gender		0.163^{*}		0.177
female		(0.087)		(0.091)
Missing indicator for		-0.933^{***}		-1.006^{***}
personal characteristics		(0.153)		(0.169)
Student participating in	0.217^{*}	-0.379^{*}	0.230^{*}	-0.419
wave 1	(0.101)	(0.209)	(0.104)	(0.227)
Individual re-tracking in		-2.589^{***}		-2.899^{***}
wave 2		(0.099)		(0.162)
Number of calls	0.493^{***}		0.501^{***}	
less than 4	(0.048)		(0.048)	
Parent participating in	2.337^{***}	0.308^{***}	2.361^{***}	0.327^{***}
wave 1	(0.051)	(0.087)	(0.052)	(0.093)
Correlation	0.415^{**}		0.434^{***}	
ρ students parents	(0.158)		(0.052)	
Random intercept				
ω school level			0.181	0.347
$\ln \mathcal{L}$	-2483.226		-2465.790	
AIC	5026.452		4995.580	
BIC	5227.993		4995.580	
Sample size	6112.000		6112.000	

Notes: ***, **, and * denote significance at the 0.1%, 1%, and 5% level, respectively. Standard errors are given in parenthesis. To model individual participation decisions, the `zelig` function with `bprobit` link provided by `ZeligChoice` package (Owen et al., 2012) in R (R Core Team, 2014) was used. Correlation parameter from the bivariate probit model without random intercept is transformed according to Honaker et al. (2013).

Appendix C

Illustrating the GHK-simulator

We implement the GHK-simulator (Geweke (1989), Hajivassiliou (1990) und Keane (1994)) in R (R Core Team, 2014) to approximate the probability of a bivariate normal distribution ($K = 2$)

$$f(x_1, x_2) = \frac{1}{2\pi\sigma_{x_1}\sigma_{x_2}\sqrt{1-\rho^2}} e^{-\frac{1}{2}\left(\frac{\left(\frac{x-\mu_{x_1}}{\sigma_{x_1}}\right)^2+\left(\frac{x-\mu_{x_2}}{\sigma_{x_2}}\right)^2-2\rho\left(\frac{x-\mu_{x_1}}{\sigma_{x_1}}\right)\left(\frac{x-\mu_{x_2}}{\sigma_{x_2}}\right)}{1-\rho^2}\right)},$$

with parameters $\mu_{x_1} = 0, \mu_{x_2} = 0, \sigma_{x_1} = 1, \sigma_{x_2} = 1$ and $\rho = 0.5$ (Schaich & Münnich, 2001). This is computing the volume enclosed by the density function within the range $x_1 \in [-4, 0]$ and $x_2 \in [-4, 4]$. The density of the bivariate normal distribution is given in Figure C.1a.

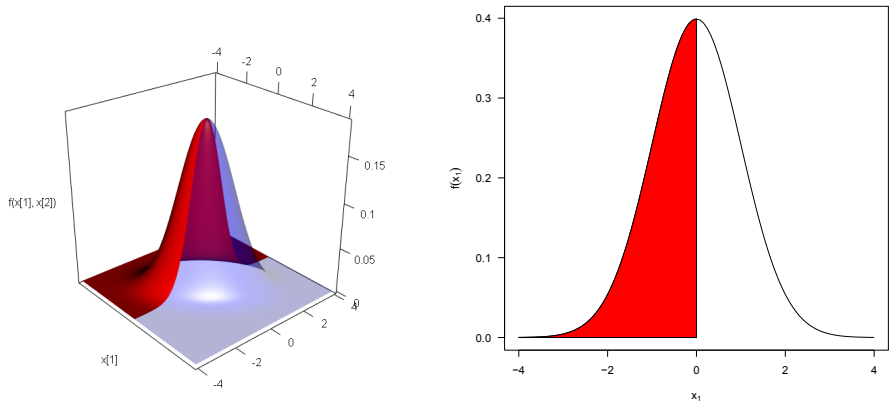
The GHK-simulator will approximate the volume under the red part of the bivariate normal distribution.

$$\text{Prob}[a_1 < x_1 < b_1, a_2 < x_2 < b_2] \approx \frac{1}{R}\sum_{r=1}^{R}\prod_{k=1}^{K} Q_{rk} \tag{C.1}$$

using univariate probabilities Q_{rk} Greene (2012, p. 668f.). Therefore we factor Σ using the Cholesky decomposition, that is $\Sigma = LL'$, where L is the lower triangular matrix. Its elements l_{km} equal 0 for $m > k$ and for $k = m = 1$ it is σ_{x_1}.

```
### Correlation matrix Simga
Sigma

##      [,1] [,2]
## [1,]  1.0  0.5
## [2,]  0.5  1.0
```



(a) Density of the bivariate normal distribution.

(b) Marginal conditional distribution.

Figure C.1: Bivariate normal distribution (via `persp3d` from the `rgl` package (Adler & Murdoch, 2010)) and its' marginal distribution.

```
### Cholesky decomposition
l <- t(chol(Sigma))
l

##      [,1]  [,2]
## [1,]  1.0 0.000
## [2,]  0.5 0.866
```

In the first step of the following recursion the probability of the marginal distribution with regard to $x_1 \in [-4, 0]$ is computed. The probability is given by the red area under the distribution in Figure C.1b.

$$Q_{1k} = \Phi(b_1/l_{11}) - \Phi(a_1/l_{11}), \tag{C.2}$$

with $a_1 = -4$ and $b_1 = 0$ denoting the lower and upper truncation points and it is $l_{11} = \sigma_1 = 1$.

```
### for k = 1 (first dimension)
###
### bounds
A[,1] <- a[1]/l[1,1]
B[,1] <- b[1]/l[1,1]
```

```
### univariate probabilities for dimension k=1
Q[,1] <- pnorm(B[,1]) - pnorm(A[,1])
```

Since the conditional distribution is a univariate normal distribution the probabilities Q_{r1} can easily be computed. The first recursion ends with drawing R random numbers ϵ_{r1} from a truncated (standard) normal distribution. A_{rk} and B_{rk} give the truncation points for sampling from the distribution. For sampling from truncated normal distributions see Greene (2012, p. 647f.).

```
### Random numbers for a truncated
### standard normal distribution for k = 2
epsilon <- qnorm(pnorm(a[1]) +
                 (pnorm(b[1]) - pnorm(a[1])) * runif(R)
                 )
summary(epsilon)

##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  -3.890  -1.140  -0.672  -0.801  -0.329  -0.001
```

To approximate the probability Prob$[-4 < x_1 < 0, -4 < x_2 < 4]$ the truncation points for the second recursion are given by

$$A_{rk} = \left[a_k - \sum_{m=1}^{k-1} l_{km}\epsilon_{rm}\right] l_{kk}^{-1} \quad \text{and} \quad B_{rk} = \left[b_k - \sum_{m=1}^{k-1} l_{km}\epsilon_{rm}\right] l_{kk}^{-1}. \tag{C.3}$$

So that the probabilities Q_{rk} are given as the difference of the distribution function at the upper and lower truncation points

$$Q_{rk} = \Phi(B_{rk}) - \Phi(A_{rk}). \tag{C.4}$$

```
### for k =2
###
### bounds
A[,2] <- (a[2] - l[2,1] * epsilon)/l[2,2]
B[,2] <- (b[2] - l[2,1] * epsilon)/l[2,2]
### univariate probabilities of the marginal
### truncated distribution for dimension k=2
Q[,2] <- pnorm(B[,2]) - pnorm(A[,2])
### random numbers from a truncated
```

```
### standard normal distribution for k = 3
epsilon <- qnorm(pnorm(a[2]) +
                 (pnorm(b[2]) - pnorm(a[2])) * runif(R)
                 )
```

For the bivariate setting the example would end here. The extension to a trivariate normal distribution is illustrated below.

```
### for k = 3
###
### bounds
A[,3] <- (a[2] - (l[3,1] * epsilon + l[3,2] + epsilon))/l[3,3]
B[,3] <- (b[2] - (l[3,1] * epsilon + l[3,2] + epsilon))/l[3,3]

# equal to
# (rowSums(matrix(epsilon,ncol=1) %*% l[3,1:2,drop=F])/l[3,3]

### univariate probabilities of the marginal
### truncated distribution for dimension k=3
Q[,3] <- pnorm(B[,2]) - pnorm(A[,2])
```

The probability in Figure C.1a can be approximated by

$$\text{Prob}[-4 < x_1 < 0, -4 < x_2 < 4] \approx \frac{1}{1000} \sum_{r=1}^{1000} \prod_{k=1}^{2} Q_{rk} = 0.4999. \tag{C.5}$$

This yields the following result which is identical to using the `mvtnorm` package in R, see Genz et al. (2013) and Genz and Bretz (2009).

```
### GHK-approximation
sum(apply(Q,1,prod))/R

## [1] 0.4999

### standard R implementation
pmvnorm(a, b, Mu, Sigma)

## [1] 0.4999
## attr(,"error")
## [1] 1e-15
## attr(,"msg")
## [1] "Normal Completion"
```

Appendix D

R code

```
logLikBPRE <- function(param, yy1, yy2, xx1, xx2, k1, k2, m,
                               n_j, S, crn1, crn2, unicrn){
  ## param: initial parameters
  ## 1 corresponds to students (s)
  ## 2 corresponds to pardents (p)
  ## k1 / k2: number of auxiliary variables in equation 1 / 2
  ## m: number of clusters
  ## n_j: number of inidviduals / size of clusters
  ## S: simulation sample size for GHK-simulator
  ## crn1 / crn2: normal random numbers for random intercept
  ## in equation 1 / 2
  ## unicrn: uniform random numbers / halton sequences for truncated
  ## normals

  ## definition of model parameters
  beta1     <- param[1:k1]           # parameters for equation 1
  beta2     <- param[(k1+1):(k1+k2)] # parameters for equation 2
  rho       <- param[k1+k2+1]        # correlation parameter
  sig1      <- param[k1+k2+2]        # sd for random intercept 1
  sig2      <- param[k1+k2+3]        # sd for random intercept 2
  sig       <- matrix(1, 2, 2)       # corvariance matrix
  sig[1,1]  <- 1
  sig[2,2]  <- 1
  sig[1,2]  <- rho
  sig[2,1]  <- rho
  L         <- t(chol(sig)) # Lower triangular of Cholesky
                            # decomposition
```

```
alpha1    <- sig1 * crn1  # random intercept 1
alpha2    <- sig2 * crn2  # random intercept 2

## estimation of the LogLikelihood
likeli <- rep(NA, m) # vector for likelihood contributions
                     # of cluster j
for(j in 1:m){ # looping through the j = 1, ..., m schools
  gammaLower <- matrix(NA,S,2*n_j[j]) # Equation 5.30
  gammaUpper <- matrix(NA,S,2*n_j[j]) # Equation 5.30
  uniInd <- ((j-1)*S+1):(j*S)
  for(i in 1:n_j[j]){ # looping through the i = 1, ..., n_j
                      # individuals in school j
    ## mu_ij = {mu1, mu2} referring to Equation 5.25
    mu_ij <- -cbind(xx1[i, ,j] %*% beta1 + alpha1[j, ],
                    xx2[i, ,j] %*% beta2 + alpha2[j, ])
    ## upper integration limits, Equations 5.24
    DUpper <- cbind((yy1[j, i] *  1000 + mu_ij[,1]),
                    (yy2[j, i] *  1000 + mu_ij[,2]))
    ## lower integration limits, Equations 5.24
    DLower <- cbind(((1-yy1[j, i]) * (-1000) + mu_ij[,1]),
                    ((1-yy2[j, i]) * (-1000) + mu_ij[,2]))
    ## random numbers form truncated normal, Equation 5.29
    vhat  <- matrix(NA,S,2)
    ind <- (i-1)*2+1
    gammaLower[,ind] <- DLower[,1]/L[1,1] # Equation 5.18
    gammaUpper[,ind] <- DUpper[,1]/L[1,1] # Equation 5.18
    ## random numbers form truncated normal
    vhat[,1] <- qnorm(unicrn[uniInd,1] *
                      pnorm(gammaUpper[,ind]) +
                      (1-unicrn[uniInd,1]) *
                      pnorm(gammaLower[,ind])
    )
    ## truncation for normal distributions, Equation 5.30
    gammaLower[,ind+1] <- (DLower[,2]-L[2,1]*vhat[,1])/L[2,2]
    gammaUpper[,ind+1] <- (DUpper[,2]-L[2,1]*vhat[,1])/L[2,2]
  }
  ## likelihood contribution of cluster j according
  ## to Eqation 5.32
  likeli[j] <- mean(
                apply(
```

```
                    pnorm(gammaUpper)-pnorm(gammaLower),
                      1,
                      prod)
                    )
  }
  logLikelihood <- -sum(log(likeli))
  return(logLikelihood)
}
```

```
library(numDeriv)
BPREoptim <- function(DataRaw, y1, x1, su1, y2, x2, seed=NULL){
  ## DataRaw: dataframe containing all information
  ## the rest of the arguments are the variable names
  nObs <- nrow(DataRaw) # total number of observations
  Y1  <- DataRaw[ ,y1]  # participation variable eq 1
  X1  <- DataRaw[ ,x1]  # includes a vector of 1
  SU1 <- DataRaw[ ,su1] # grouping variable
  Y2  <- DataRaw[ ,y2]  # participation variable eq 2
  X2  <- DataRaw[ ,x2]  # includes a vector of 1
  S     <- 1000         # simulation sample size
                        # for GHK-simulator
  k1    <- length(x1)   # number of variables equation 1
  k2    <- length(x2)   # number of variables equation 2

  ## prepare and arrange data
  m <- length(unique(SU1))      # number of clusters
  n_j <- as.vector(table(SU1))  # number of individuals in j

  ## matrices and arrays containing (in-)dependant variables
  yy1 <- matrix(NA, m, max(n_j)) # dep. variable equation 1
  yy2 <- matrix(NA, m, max(n_j)) # dep. variable equation 2
  ## independend variabled equations 1 and 2
  xx1 <- array(NA, dim=c(max(n_j), length(x1), m))
  xx2 <- array(NA, dim=c(max(n_j), length(x2), m))

  ## filling empty matrices and arrays by clusters
  for(j in 1:m){
    ## matrix with y1 and y2 for individuals per cluster,
    ## dimension m x max(n_j)
    pos <- which(SU1 == unique(SU1)[j])
    yy1[j, 1:n_j[j]] <- as.vector(Y1[pos])
    yy2[j, 1:n_j[j]] <- as.vector(Y2[pos])
    ## array with X1 and X2 for individuals per cluster,
    ## dimension n_j x k x m (3D)
    xx1[1:n_j[j], 1:length(x1), j] <- as.matrix(DataRaw[pos, x1])
    xx2[1:n_j[j], 1:length(x2), j] <- as.matrix(DataRaw[pos, x2])
  }
  ## random numbers for simulations
  if(is.null(seed)){
```

```
  ## random numbers for evaluation of numerical precision
  ## no seed -> different random numbers in each replication
  ## common random numbers for truncated normals
  unicrn<- matrix(runif(2*S*m), S*m, 2)
  ## normal random numbers for random intercept
  crn1   <- matrix(rnorm(S*m), m, S)
  crn2   <- matrix(rnorm(S*m), m, S)
}else{
  ## random numbers for evaluation of statistical precision
  ## fixed seed -> no variation in random numbers
  set.seed(seed)
  unicrn<- matrix(runif(2*S*m), S*m, 2)
  crn1  <- matrix(rnorm(S*m), m, S)
  crn2  <- matrix(rnorm(S*m), m, S)
}

## initial parameters, lower and upper bounds for optimization
theta <- c(rep(  , (k1+k2)),  0,      0.1,     0.1)
LB    <- c(rep(-5, (k1+k2)), -0.99,   0.001,   0.001)
UB    <- c(rep( 5, (k1+k2)),  0.99,   5,       5)
## optimization
ergMin <- nlminb(start=theta,
                 objective = logLikBPRE,
                 gradient = NULL,
                 hessian = NULL,
                 yy1, yy2, xx1, xx2, k1, k2,
                 m, n_j, S, crn1, crn2, unicrn,
                 control = list(trace=1,
                                iter.max=500,
                                rel.tol=10^-6),
                 lower = LB, upper = UB)
## log-likelihood
logLik <- -ergMin$objective
## hessian matrix
Hesse  <- numDeriv:::hessian(func = logLikBPRE,
                             x = ergMin$par,
                             method = 'Richardson',
                             method.args = list(),
                             yy1, yy2, xx1, xx2, k1, k2,
                             m, n_j, S, crn1, crn2, unicrn)
```

```
  ## estimated parameters
  ## all model parameters (beta1, beta2, rho, sigma1, sigma2)
  theta   <- ergMin$par
  nPar    <- length(theta) # number of estimated parameters
  ## standard errors, t-values, p-values
  seTheta <- sqrt(diag(solve(Hesse)))
  tValue  <- theta/seTheta
  pValue  <- 2*(1-pt(abs(tValue), nObs-length(tValue)))
  sig <- rep(NA, nPar) # significance labels
  sig[pValue >= 0.1]  <- ''
  sig[pValue < 0.1]   <- '.  '
  sig[pValue < 0.05]  <- '*  '
  sig[pValue < 0.01]  <- '** '
  sig[pValue < 0.001] <- '***'

  ## putting it all together
  Final <- data.frame(theta, seTheta, tValue, pValue, sig,
                      row.names=c(paste(x1, ':1', sep=''),
                                  paste(x2, ':2', sep=''),
                                  'rho', 'sigma1', 'sigma2'))
  colnames(Final) <- c('Estimate', 'StdError',
                       'tValue', 'pValue', '')

  AIC    <- 2*nPar-2*logLik
  BIC    <- nPar*log(nObs)-2*logLik
  ## output list
  OutList <- list('Coefficients' = Final,
                  'logLik'       = logLik,
                  'AIC'          = AIC,
                  'BIC'          = BIC,
                  'N'            = nObs,
                  'm'            = m,
                  'Hessian'      = Hesse,
                  'Optimization' = ergMin)

  return(OutList)
}
```

Appendix E

R session information

- R version 3.1.1 (2014-07-10), `x86_64-w64-mingw32`
- Base packages: base, datasets, graphics, grDevices, grid, methods, splines, stats, stats4, tools, utils
- Other packages: boot 1.3-11, foreign 0.8-61, knitr 1.6, lattice 0.20-29, lme4 1.1-7, MASS 7.3-33, Matrix 1.1-4, memisc 0.96-9, mvtnorm 1.0-0, pps 0.94, Rcpp 0.11.2, rgl 0.93.1098, rJava 0.9-6, sandwich 2.3-1, stringr 0.6.2, vcd 1.3-2, VGAM 0.9-4, xlsx 0.5.7, xlsxjars 0.6.0, xtable 1.7-3, Zelig 4.2-1, ZeligChoice 0.8-1
- Loaded via a namespace (and not attached): car 2.0-20, colorspace 1.2-4, evaluate 0.5.5, formatR 0.10, highr 0.3, minqa 1.2.3, nlme 3.1-117, nloptr 1.0.4, nnet 7.3-8, zoo 1.7-11

Analyses presented in this thesis are mostly based on R version 3.1.1 (R Core Team, 2014). Further information on *The R Project for Statistical Computing* can be found on the website `www.r-porject.org`. This document is written using `knitr` (combining LaTeX and `R`), see Xie (2012). Data are read using `read.dta` or `read.xlsx` functions provided by packages `foreign` (R Core Team, 2013) and `xlsx` (Dragulescu, 2013). Tables of dimensions higher than two are created using the `structable` function from the `vcd` package (Meyer, Zeileis, and Hornik (2013) and Meyer, Zeileis, and Hornik (2006)). Tables are exported to LaTeX using the packages `memisc` (Elff, 2012) and `xtable` (Dahl (2012) and Swinton (2009)). Additionally used packages are cited at appropriate positions in text or table notes.