NEPS
National Educational Panel Study

Research Data

Weighting the Student Sample of the National Educational Panel Study (Wave 1 to 6) Technical Report on SUF Version 6-0-0

Sabine Zinn, Christian Aßmann, Hans Walter Steinhauer, & Benno Schönberger

LIfBi
LEIBNIZ INSTITUTE FOR
EDUCATIONAL TRAJECTORIES

Copyrighted Material
Leibniz Institute for Educational Trajectories (LIfBi)
Wilhelmsplatz 3, 96047 Bamberg
Director: Prof. Dr. Hans-Günther Roßbach
Executive Director of Research: Dr. Jutta von Maurice
Executive Director of Administration: Dr. Robert Polgar
Bamberg, 2016

Weighting the Student Sample of the National Educational Panel Study (Wave 1 to 6)

Technical Report on SUF Version 6-0-0

Sabine Zinn, Christian Aßmann, Hans Walter Steinhauer, & Benno Schönberger

Corresponding address: methoden@lifbi.de

Version: April 2016

Dieser Bericht beinhaltet Ergebnisse, welche von *infas - Institut für angewandte Sozialwissenschaften GmbH* bereitgestellt wurden. Er erfüllt damit auch die Berichtspflichten des Erhebungsinstituts in Bezug auf die Studien B52, B55, und B59.

1 Introduction

This report refers to the Scientific Use File (SUF) `doi:10.5157/NEPS:SC5:6.0.0` of the survey “first-year undergraduate students in higher education” (Starting Cohort 5, SC5) conducted within the National Educational Panel Study (NEPS) in 2010, 2011, and 2012. The SC5 survey is part of the main cohort samples of the NEPS and focuses on central issues such as educational choices, competence development, the benefits of higher education, and entry into the job market. On the basis of a short review of the survey and the sampling design applied (in Section 2), this report presents information on the initial sample and results of the weighting procedures applied. Weighting for these students involves a step-by-step process. First, a correction of design weights was performed in order to adequately reflect the current numbers of students based on data from the Federal Statistical Office of Germany for the winter semester 2010/2011. Second, weights for participating students were calculated for seven studies, that is, for six survey waves. The studies B52[1], B55, and B59 were conducted via computer-assisted telephone interviews (CATIs).[2] The studies B54, B56, and B58 are online surveys. Table 1 depicts the attribution of the studies to the six panel waves. We describe the stage process of computing the different kinds of sampling weights in Section 3. Section 4 describes the procedure applied for trimming and standardizing the weights. Finally, Section 5 gives a summary on the sampling weights provided and some advice regarding their usage.

2 Population and Sampling Design

The target population is defined as all first-year students (German and non-German) enrolled for the first time in public or state-approved institutions of higher education in Germany who are aiming for a Bachelor’s degree, a state examination (*Staatsexamen*) in medicine, law, pharmacy, and teaching, a diploma or Master’s degree in Roman Catholic or Protestant theology or specific art and design degrees in the academic year of 2010/2011 . Students attending universities, universities technical or universities of applied sciences run by Federal Ministries or Federal States for members of their public services are excluded.[3] A stratified cluster sample was drawn from the defined population of first-year students at corresponding higher education institutions, see also Aßmann et al. (2011). We define a cluster as all students enrolled in a certain subject (of the sixty officially listed fields, see

[1] The study B53 involves competence tests that have been conducted in parallel to the telephone interviews of the B52 study. Thus, for reasons of convenience, both studies are pooled in Wave 1. Accordingly, results subsequently presented referring to the participation in study B52 also concern the B53 study.

[2] In this SUF, no sampling weights are provided for the B57 study dealing with competence tests conducted in Spring 2013. This is because, the SUF at hand comprises only parts of the B57 outcome. Weighting only a subset of the B57 sample might be precarious.

[3] In the beginning, the plan was to conduct a census among the students with a non-traditional admission certificate. However, difficulties during the recruiting process hindered this project. In detail this means that even though students with a non-traditional admission certificate were contacted separately, namely by conventional mail, a significant part of them was additionally recruited in the same way as students with traditional admission certificate, namely in courses targeted at first-year students. As a consequence, in the end it was impossible to disentangle both groups of students completely. Therefore, in the sampling process students with traditional and students with non-traditional admission certificate were not further differentiated.

Table 1: Attribution of Studies to Panel Waves.

Wave	**Study**	**Survey Time**
Wave 1	B52/B53	Winter 2010/11
Wave 2	B54	Autumn 2011
Wave 3	B55	Spring 2012
Wave 4	B56	Autumn 2012
Wave 5a	B59	Spring 2013
Wave 6	B58	Autumn 2013

Note: Wave 5 consists of the surveys of study B59 (Wave 5a) and B59 (Wave 5b). In the SUF at hand, no sampling weights are provided for Wave 5b; see footnote 2.

Table 2) [4] at a particular higher education institution. For example, all students studying social sciences (*Sozialwissenschaften*) at the (*public*) University of Bamberg form one cluster. Within each cluster, all students are to be surveyed. The student cohort has been set up to incorporate an oversampling of teacher education students and students attending private higher education institutions, that is, private universities and private universities of applied sciences. This objective is addressed by setting up a first stratification level according to educational institution. This first stratification level defines four strata: Stratum h_1 comprises the clusters linked to teacher education at public universities. Stratum h_2 is set up to include all fields of study (for teacher education) at public universities, whereas stratum h_3 summarizes all fields of study offered by public universities of applied sciences. Finally, stratum h_4 comprises all degree programs offered by private universities or private universities of applied sciences. This level of stratification allows us to carry out an oversampling of teacher education students and students at private higher education institutions by using different sampling rates of clusters in the different strata. Overall, the plan was to establish a gross sample of 66,450 students[5]–15,950 students in stratum h_1, 26,500 students in stratum h_2, 16,800 students in stratum h_3, and 7,200 students in stratum h_4.

Given the heterogeneous distribution of students across the officially listed fields of study, sampling within the defined strata would result in a large sampling variation concerning the coverage of the range of subjects within the sample. Hence, a further level of stratification was introduced where strata are defined by groups of related subjects. This stratification was accompanied by an exclusion of clusters with less than thirty enrolled students in the academic year of 2008/2009. In summary, the sixty fields of study are pooled in several subject groups within the first-level stratum, see Table 2. Thus, strata s_1 to s_3 pool fields of study in stratum h_1, strata s_4 to s_{19} correspond to the first-level stratum h_2, $s_{20}, \ldots, s_{26}$ comprise fields of study within the stratum h_3, and pooling in the stratum h_4 is achieved by setting up a second-level strata s_{27} to s_{29}.

[4] In contrast to the definition provided by the Federal Statistical Office of Germany we separated three clusters of teacher training programmes from the fields of subjects and added them to the list.

[5] Assuming that a quarter of the sampled students refuse to participate, this yields approximately the intended net sample size of 16,500 students, see, for example, Aßmann et al. (2011).

Table 2: Allocation of the Sixty Listed Fields of Study to the Two Stratification Levels h_i and s_j $(i = 1, \cdots, 4, j = 1, \cdots, 29)$.

Code	Officially listed subject	h_1	h_2	h_3	h_4
1	*Sprach- und Kulturwissenschaften allgemein*	–	s_5	s_{20}	s_{27}
2	*Evangelische Theologie, Religionslehre*	–	s_4	s_{20}	s_{27}
3	*Katholische Theologie, Religionslehre*	–	s_4	s_{20}	s_{27}
4	*Philosophie*	–	s_4	–	s_{27}
5	*Geschichte*	–	s_4	–	s_{27}
6	*Bibliothekswissenschaft, Dokumentation, Publizistik*	–	s_5	s_{20}	s_{27}
7	*Allgemeine und vergleichende Literatur- und Sprachwissenschaft*	–	s_5	s_{20}	s_{27}
8	*Altphilologie (klassische Philologie), Neugriechisch*	–	s_4	–	s_{27}
9	*Germanistik (Deutsch, germanische Sprachen ohne Anglistik)*	–	s_6	–	s_{27}
10	*Anglistik, Amerikanistik*	–	s_7	–	s_{27}
11	*Romanistik*	–	s_7	–	s_{27}
12	*Slawistik, Baltistik, Finno-Ugristik*	–	s_7	–	s_{27}
13	*Außereuropäische Sprach- und Kulturwissenschaften*	–	s_7	s_{20}	s_{27}
14	*Kulturwissenschaften i.e.S.*	–	s_7	s_{20}	s_{27}
15	*Psychologie*	–	s_8	–	s_{27}
16	*Erziehungswissenschaften*	–	s_8	s_{21}	s_{27}
17	*Sonderpädagogik*	–	s_8	s_{21}	s_{27}
18	*Sport, Sportwissenschaft*	–	s_8	s_{20}	s_{27}
19	*Wirtschafts- und Gesellschaftslehre allgemein*	–	s_9	s_{20}	s_{27}
20	*Regionalwissenschaften*	–	s_9	s_{20}	s_{27}
21	*Politikwissenschaften*	–	s_9	s_{20}	s_{27}
22	*Sozialwissenschaften*	–	s_9	s_{20}	s_{27}
23	*Sozialwesen*	–	s_8	s_{21}	s_{27}
24	*Rechtswissenschaft*	–	s_{10}	s_{20}	s_{27}
25	*Verwaltungswissenschaft*	–	s_{10}	s_{20}	s_{27}
26	*Wirtschaftswissenschaften*	–	s_{11}	s_{22}	s_{28}
27	*Wirtschaftsingenieurwesen*	–	s_{11}	s_{22}	s_{28}
28	*Mathematik, Naturwissenschaften allgemein*	–	s_{14}	s_{23}	s_{29}
29	*Mathematik*	–	s_{12}	s_{23}	s_{29}
30	*Informatik*	–	s_{12}	s_{23}	s_{29}
31	*Physik, Astronomie*	–	s_{12}	–	s_{29}
32	*Chemie*	–	s_{13}	s_{23}	s_{29}
33	*Pharmazie*	–	s_{13}	s_{23}	s_{29}
34	*Biologie*	–	s_{14}	–	s_{29}
35	*Geowissenschaften (ohne Geographie)*	–	s_{14}	s_{23}	s_{29}
36	*Geographie*	–	s_{14}	–	s_{29}
37	*Gesundheitswissenschaften allgemein*	–	s_{15}	s_{23}	s_{29}
38a	*Humanmedizin ohne Zahnmedizin (ohne Approbation)*	–	s_{15}	–	s_{29}
38b	*Humanmedizin ohne Zahnmedizin (mit Approbation)*	–	s_{19}	–	s_{29}
39	*Zahnmedizin*	–	s_{15}	–	s_{29}
40	*Veterinärmedizin*	–	s_{15}	–	s_{29}
41	*Landespflege, Umweltgestaltung*	–	s_{15}	s_{23}	s_{29}
42	*Agrarwissenschaften, Lebensmittel- und Getränketechnologie*	–	s_{15}	s_{23}	s_{29}
43	*Forstwissenschaft, Holzwirtschaft*	–	s_{15}	s_{23}	s_{29}
44	*Ernährungs- und Haushaltswissenschaften*	–	s_{15}	s_{23}	s_{29}
45	*Ingenieurwesen allgemein*	–	s_{17}	–	s_{29}
46	*Bergbau, Hüttenwesen*	–	s_{17}	s_{26}	s_{29}
47	*Maschinenbau/Verfahrenstechnik*	–	s_{16}	s_{24}	s_{29}
48	*Elektrotechnik*	–	s_{17}	s_{25}	s_{29}
49	*Verkehrstechnik, Nautik*	–	s_{17}	s_{26}	s_{29}
50	*Architektur, Innenarchitektur*	–	s_{17}	s_{26}	s_{29}
51	*Raumplanung*	–	s_{17}	s_{26}	s_{29}
52	*Bauingenieurwesen*	–	s_{17}	s_{26}	s_{29}
53	*Vermessungswesen*	–	–	s_{26}	s_{29}
54	*Kunst, Kunstwissenschaft allgemein*	–	s_{18}	s_{20}	s_{27}
55	*Bildende Kunst*	–	s_{18}	s_{20}	s_{27}
56	*Gestaltung*	–	s_{18}	s_{20}	s_{27}
57	*Darstellende Kunst, Film und Fernsehen, Theaterwissenschaft*	–	s_{18}	s_{20}	s_{27}
58	*Musik, Musikwissenschaft*	–	s_{18}	s_{20}	s_{27}
59	*Außerhalb der Studienbereichsgliederung/Sonstige Fächer*	–	s_{18}	–	s_{27}
60a	*Lehramt: LA Grund+Haupt/LA Grund/LA Haupt/BA Sek I+Primar/ LA+BA Grundschule+SekI/LA Real/LA Real+BA Real+Haupt/ LA+BA Sonder+Förder*	s_1	–	–	–
60b	*Lehramt: LA Gym/BA Gym/BA allg./LA Oberstufe+Sek II/ LA+BA Berufl./LA Ober+Sek II+berufl.*	s_2	–	–	–
60c	*Lehramt: BA Lehramt allg.*	s_3	–	–	–

Table 3: Number of Clusters Sampled and Realized in Each Stratum.

Stratum		Number of clusters	
1^{st} level	2^{nd} level	sampled	realized
h_1	s_1	21	18
	s_2	26	25
	s_3	7	9
h_2	s_4	10	11
	s_5	9	9
	s_6	8	9
	s_7	16	10
	s_8	18	20
	s_9	17	18
	s_{10}	8	8
	s_{11}	18	21
	s_{12}	24	23
	s_{13}	11	12
	s_{14}	17	15
	s_{15}	10	8
	s_{16}	5	9
	s_{17}	14	12
	s_{18}	12	9
	s_{19}	6	7
h_3	s_{20}	15	14
	s_{21}	12	13
	s_{22}	35	35
	s_{23}	31	28
	s_{24}	15	20
	s_{25}	13	9
	s_{26}	24	23
h_4	s_{27}	21	13
	s_{28}	29	19
	s_{29}	21	17

Note: Discrepancies between the number of sampled and realized clusters are caused by (1) whole clusters dropping out and (2) incorrect information of students about their main subject. We use poststratification to correct for these deficiencies.

The number of clusters to be drawn within each stratum h_1 to h_4 was determined such that the sample distribution of students across the fields of study resembled the one in the population. At the same time, the intended oversampling could be incorporated in a straightforward way and homogeneous inclusion probabilities were probable to realize. In particular, the number of clusters m_i sampled within stratum h_i is calculated according to

$$m_i = \frac{\tilde{n}_i}{\frac{1}{K_i}\sum\limits_{k=1}^{K_i} N_{ik}}, \tag{1}$$

dividing the planned sample size $\tilde{n}_i$ in stratum h_i by the average cluster size in terms of the number of first-year students N_{ik} in the academic year of 2008/2009 for all clusters $k = 1, \ldots K_i$ in h_i. (Here, K_i denotes the total number of clusters in stratum h_i.) In sum, we obtain $m_1 = 54$ clusters to be sampled for stratum h_1 and $m_4 = 71$ clusters to be sampled for stratum h_4. For strata h_2 and h_3, where no oversampling was carried out, a total of 348 clusters to be sampled has been found sufficient to generate the planned gross sample size. Here, clusters are allocated proportionally to the overall number of clusters in both strata, resulting in $m_2 = 203$ clusters to be sampled in stratum h_2 and $m_3 = 145$ clusters in stratum h_3. For each substratum the number of clusters m_{ij} to be sampled from stratum h_i, $i = 1, \ldots, 4$ are calculated according to

$$m_{ij} = m_i \frac{K_{ij}}{K_i}, \tag{2}$$

where K_{ij} denotes the total number of clusters in the second-level stratum s_j in the first-level stratum h_i. Table 3 gives the corresponding numbers. Within each stratum h_i and s_j the m_{ij} clusters are sampled by simple random sampling without replacement so that the inclusion probability for cluster k_{ij} is given by

$$\pi_{ij} = \frac{m_{ij}}{K_{ij}}. \tag{3}$$

Inserting equation (2) yields

$$\pi_{ij} = \frac{m_i}{K_i} \tag{4}$$

and the corresponding design weight d_i is given by the inverse of that inclusion probability

$$d_i = \frac{K_i}{m_i} = \begin{cases} \frac{90}{54} & = 1.667 \quad \text{for} \quad i = 1 \\ \frac{1276}{203} & = 6.286 \quad \text{for} \quad i = 2 \\ \frac{923}{145} & = 6.366 \quad \text{for} \quad i = 3 \\ \frac{134}{71} & = 1.887 \quad \text{for} \quad i = 4. \end{cases} \tag{5}$$

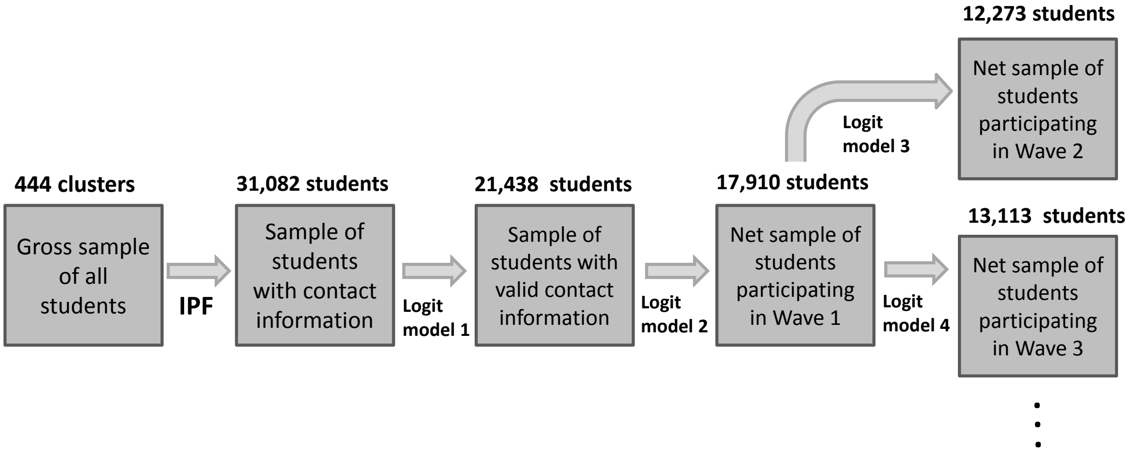
To handle institutional nonparticipation, the following replacement strategy was implemented. If a university refuses to participate, all fields of study sampled at this specific university are lost. Hence, only those institutions are eligible for replacement that maintain the original sample composition with regard to the sampled departments and subjects. For each combination of sampled subjects at a particular higher education institution, all institutions offering the same combination of subjects within the frame are listed, irrespective of whether the institutions have already been sampled or not. Institutions not sampled are given preferential consideration in the choice of replacement candidates. Given that several replacement institutions offer the same combination of subjects to be replaced, the replacement institution is defined as the one with the smallest difference in numbers of enrolled students compared to the nonparticipating institution.

These steps were carried out on the basis of information on first-year students from the winter semester 2008/2009 (provided by the Federal Statistical Office of Germany). At the point of planning the sampling and recruitment procedures, these were the most current data available for the population of students. As (during the planning process) the absolute number of first-year students had risen from 2008/2009 to 2009/2010 by about 6.5%, a further rise in 2010/2011 seemed probable. This fact was taken into account by incrementing the 2008/2009 data by 10% in order to have a good estimate of the actual number of students for the sampling process in 2010.

In order to achieve high response rates, two different contact modes were employed to approach the sampled students: First, all students were informed about the NEPS and invited to participate in den panel study via conventional mail. Besides this, several institutions facilitated a second way of contact by the personal information and recruitment in courses targeted at or mandatory for first-year students. In a pilot study, this twofold recruitment process yielded higher participation rates, as well as a higher panel attendance. In total, 31,082 first-year students could be contacted via this procedure. The following section outlines the performed weighting adjustments.

3 Derivation of Sampling Weights

To mirror the recruitment and participation process within the weighting adjustments, consecutive modeling of the decision and participation process is performed, see Figure 1. The first modeling step involves the correction of the stratum-specific design weights d_i in relation to the nonresponse occurring from the gross sample of students (in the clusters previously determined) to the set of students who provided (any kind of) contact information. The second modeling step corrects for nonresponse occurring from the sample of persons with contact information (of any kind) to the sample of persons with valid contact information–that is, to the gross sample of Wave 1 (corresponding to the CATI of the study B52). All further modeling steps correct for the nonresponse among the recruited students in the distinct survey waves (i.e., in the studies B52, B54, B55, B56, B57, and B59). Note that as participation in the first telephone interview (i.e., in the study B52) forms the indispensable backbone of the panel study, the panel cohort is defined as the set of students who participated in Wave 1. In total, (currently) the panel cohort comprises 17,910 students. Consequently, all computations related to nonresponse adjustments in further waves refer to this set of students minus the number of students who refuse to



444 clusters
Gross sample of all students
IPF
31,082 students
Sample of students with contact information
Logit model 1
21,438 students
Sample of students with valid contact information
Logit model 2
17,910 students
Net sample of students participating in Wave 1
Logit model 3
12,273 students
Net sample of students participating in Wave 2
Logit model 4
13,113 students
Net sample of students participating in Wave 3

Figure 1: Steps of Consecutive Modeling of the Decision and Participation Process.

participate further in the panel or are subject to the so-called *2-years rule* of NEPS.[6] With regard to the first step, an iterative proportional fitting (IPF) mechanism originally described by Deming and Stephan (1940) was implemented. The IPF uses mathematical scaling to ensure that a multidimensional table of data is adjusted so that its row and column totals correspond to constrained row and column totals obtained from alternative sources.[7] We apply the procedure to determine weighting factors for the 31,082 students who provided contact information, on the basis of current frame information on student numbers and attributes from the winter semester 2010/2011–when sampling took place. The respective variables were gender, German versus non-German students, public versus private higher education institutions, universities versus universities of applied sciences as well as an indicator variable for the subject.[8] The weighting factors derived in this way are multiplied to the design weights d_i referring to the first-level strata h_1 to h_4, yielding sampling weights w^0_{ijs} for all students s in the first-level stratum h_i and in the second-level stratum s_j who have their provided contact information.

The second modeling step determines the propensity of students to actually participate in Wave 1.[9] Therefore, first the loss occurring from the sample of students with contact information (i.e, the recruited sample) to the sample of students with valid contact information (i.e., the gross sample of Wave 1) is modeled. Thereafter, the decision of all contacted students to actually participate is specified. The variables considered here are gender,

[6] If a student does not participate in NEPS for 3 consecutive years, he/she is marked as a final drop out; compare Sixt & Aßmann (2013). In the student panel, such cases have not occurred so far.

[7] To this end, values of the original table are gradually adjusted through repeated calculations to fit row and column constraints.

[8] The corresponding data were taken from the Federal Statistical Office of Germany (Statistisches Bundesamt, 2011).

[9] Unfortunately, some students participated twice in one study or even asked to friends to participate in their place. All such cases that could be detected up to now were deleted from the panel cohort. However, it might be that in the sample there are still one or the other case. Accordingly, in future the size of the panel cohort sample might (slightly) change. All numbers reported here refer to the time of the preparation of this report.

nationality (German, foreign, unknown), type of institution (university, *Fachhochschule*, abroad/not specified), year of birth, intended university degree (Bachelor, *Staatsexamen Lehramt*, other) and type of contact (personal or postal). Note that for only 26,913 of the 31,082 students who provided any kind of contact information enough (valid) data were available to include them into the analysis. Only 18,030 of the 21,438 students who were asked to participate in the first wave could be considered in the second model because they provided sufficient information on the considered variables. The empirical analysis is performed under the assumption that data are missing completely at random. Table 6 and 7 (given in the Appendix A) document the results of the corresponding models. [10] Overall, we find that women and German students have a higher propensity than men and foreign students to give their valid contact information and to participate in the first survey wave. Likewise, students aiming for a teacher training programme and students who were contacted by mail show a higher tendency to give valid information and to participate in Wave 1.

On the basis of the outcome of the two logit models presented, adjustment factors for all students participating in Wave 1 can be computed. Multiplying these by the weights w_{ijs}^0 yields the (cross-sectional) weights w_{ijs}^1 of students to attend Wave 1. We correct for potential deviation of the weights distribution from the distribution of first-year students in winter semester 2010/2011 in the distinct fields of study[11] and in the first strata by poststratification, and align the weights w_{ijs}^1 accordingly.

Participation modeling of Wave 2 and all further waves (i.e., studies B54, B55, B56, B58, and B59) is based on the panel cohort (i.e., the sample of Wave 1), see Figure 1. Here, the following variables are included as fixed effects: participation in previous waves, type of institution, public or private institution, gender, mother tongue, educational degree of parents, intended higher education degree, enjoyment of studies, migration background (measured by generation status), household size, presence of kids, year of birth, reading ability (measured by NEPS tests in the study B53), whether a student has a traditional higher education entrance qualification (i.e., students with a school leaving certificate qualifying for higher education), and whether he/she has changed the degree program. Some variables are time-dependent (such as the presence of kids) and updated on the basis of data from the current study. Differences due to different fields of study and institutions are accounted for by including relevant random effects. Item-nonresponse in the data has been tackled by multiple imputation–concretely, by multivariate imputation by chained equations (van Buuren and Groothuis-Oudshoorn (2011). Here, besides the variables already included in the model, also Federal State and the different levels of stratification are considered. The results of the corresponding logit models are given in Tables 9 to 12 (see Appendix A). All tables report only significant (fixed) effects. In conclusion, the participation in previous waves is a very strong indicator for the propensity to participate in future waves. Furthermore, as already noted before, women and German students are usually more likely to take part the survey than men and foreign students. As expected,

[10] The estimation of these two models and the related data preparation were conducted by Martin Kleudgen and Reiner Gilberg from *infas - Institut für angewandte Sozialwissenschaften GmbH.*

[11] The following ten categories were considered: *Spach-/ Kulturwissenschaften, Rechts-/ Wirtschafts-/ Sozialwissenschaften, Mathematik/ Naturwissenschaften, Humanmedizin, Agrar-/ Forst-/ Ernährungswissenschaften, Ingenieurwissenschaften, Kunst, Lehramt.*

students who have changed their degree program and students who do not enjoy studying show a low tendency of participation. A further impediment to participation is migration background and low reading ability. Students with kids show a higher propensity to participate if they are interviewed via telephone and a lower propensity in case of online surveys and testing (cf. results for study B58). Similarly, household size (more than one person) is an indicator for a positive attitude towards participating in telephone interviews and a negative one for participating in online surveys and testing.

On the basis of all estimated models participation probabilities are predicted and adjustment factors are derived.[12] By means of these adjustment factors, cross-sectional sampling weights w^r_{ijs} for participating in the single survey waves $r = 2, \cdots, 7$ and longitudinal sampling weights w^u_{ijs}, $u \subseteq \{1, \cdots, 7\}$, (e.g., for always participating or for participating in all CATI interviews) can be computed. However, as the set of possible participation patterns becomes highly complex with an increasing number of survey waves conducted, the set of longitudinal weights provided is restricted to only successive waves and/or to the survey mode–that is, CATI or online.

4 Trimming and Standardizing Weights

To possibly increase the statistical efficiency of weighted analysis, the adjusted weights were trimmed. The general goal of weight trimming is to reduce sampling variance and, at the same time, to compensate for potential increase in bias. Trimming was performed using the so-called "Weight Distribution" approach (Potter, 1990). Here, design weights are assumed to follow an inverse beta distribution with a cumulative distribution function F_w. Parameters of the sampling weight distribution are estimated using the sampling weights, and a trimming level τ is computed, whose occurrence probability is 1%, that is, $1 - F_w(\tau) = 0.01$. Sampling weights in excess of τ are trimmed to this level and the excess is distributed among the untrimmed weights. The parameters for the sampling weight distribution are then again estimated using the trimmed adjusted weights, and a revised trimming level $\tilde{\tau}$ is computed. The trimmed adjusted weights are compared to the revised level $\tilde{\tau}$. If any weights are in excess of $\tilde{\tau}$, they are trimmed to this level, and the excess is distributed among the untrimmed weights. This procedure is iteratively repeated until no weights are in excess of a newly revised trimming level. To ease statistical analysis, the trimmed sampling weights are standardized with mean 1.

5 Summary of Weights and Advice Regarding the Usage of Weights

The weights are provided 'purely' and–to ease statistical analysis–in a trimmed and standardized form. Table 4 lists the types of weights provided for SUF release version 5-0-0 and Table 5 gives some summary statistics of the weights provided.

No general recommendation for the usage of sampling weights can be given. Whether, and if so how, weights have to be used depends on the problem to be studied. However, it is

[12] Adjustment factors are defined as the inverse participation probabilities.

Table 4: Types of Weights Provided.

Type of weight	Label
Weights of strata	w_h
Weights of students participating in B52	w_t1
Weights of students participating in B54	w_t2
Weights of students participating in B55	w_t3
Weights of students participating in B56	w_t4
Weights of students participating in B59	w_t5a
Weights of students participating in B58	w_t6
Weights of students participating in B52 and B55	w_t13
Weights of students participating in B52, B55, & B59	w_t135a
Weights of students participating in B54 and B56	w_t24
Weights of students participating in B54, B56, & B58	w_t246
Weights of students participating in B52, B54, B55, B56, B59, & B58	w_t12345a6

Table 5: Summary Statistics for (Trimmed and Standardized) Weights.

Label of weight	Number of students	Min.	Lower Quart.	Median	Mean	Upper Quart.	Max.
w_t1	17,910	0.009	0.329	0.997	1.000	1.328	3.386
w_t2	12,273	0.009	0.332	0.978	1.000	1.337	3.441
w_t3	13,113	0.010	0.321	0.981	1.000	1.344	3.512
w_t4	11,202	0.009	0.322	0.901	1.000	1.325	3.784
w_t6	10,183	0.011	0.320	0.871	1.000	1.301	3.897
w_t5a	12,694	0.009	0.316	0.922	1.000	1.333	3.727
w_t24	9,351	0.009	0.318	0.951	1.000	1.342	3.619
w_t246	7,424	0.019	0.314	0.923	1.000	1.342	3.781
w_t13	13,113	0.010	0.321	0.981	1.000	1.344	3.512
w_t135a	10,995	0.010	0.315	0.951	1.000	1.359	3.653
w_t12345a6	5,875	0.182	0.303	0.973	1.000	1.391	3.670

commonly recommended to apply sampling weights when conducting descriptive statistics. For analytical analysis, models have to be tested for their dependence on the sampling design. Specifically, this means that the user has to ensure that the way of sampling has no or only negligible effect on the model results or/and that the sampling design is adequately considered in the model specification. A general description of how to test and account for the sampling design is given in, for example, Snijder and Bosker (2012). As a guideline, we recommend to include all variables employed for constructing the (used set of) weights as explanatory variables into the model under consideration.

The *survey* package of Stata allows defining the survey design of the sample at hand, and thus conducting design-based inference in an appropriate way (Kreuter and Valliant, 2007). The accordant command for the SC5 sample is

```
gen f_h = w_h^{-1}
svyset ID_cl [pweight=w_t1], strata(stratum) fpc(f_h)
```

In this command, `f_h` gives the sampling rate used as final population correction factor, ID_cl determines the cluster membership of a sampled student, and w_t1 describes the corresponding survey weight (to be part of the SC5 sample). The term stratum is self-explanatory. All subsequent analysis has to be preceded by the prefix svy. Also the statistical software R provides a survey package to deal with design-based inference, see Lumley (2004). Here, the definition of a design object is similar to the one asked for in Stata.

References

[1] Aßmann, C., and Sixt, M. (2013). Umgang mit (temporären) Ausfällen im Nationalen Bildungspanel (NEPS) - Startkohorten-übergreifende Regelung. Internal NEPS Discussion Paper. *(Available on request from the corresponding author.)*

[2] Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., Rohwer, G., Rässler, S., and Blossfeld, H.-P. (2011). Sampling designs of the National Educational Panel Study: Challenges and solutions. In H.-P. Blossfeld, H.-G. Roßbach, & J. von Maurice (Eds.), Education as a lifelong process: The German National Educational Panel Study (NEPS) (pp. 51-65).

[3] Deming, W. E., and 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, Vol. 11, No. 4 (Dec., 1940), pp. 427-444*, Published by: Institute of Mathematical Statistics, Stable URL: http://www.jstor.org/stable/2235722.

[4] Kreuter, F., and Valliant, R. (2007), A survey on survey statistics: What is done and can be done in Stata, Stata Journal, 7(1), 1.

[5] Lumley, T. (2004), Analysis of complex survey samples, Journal of Statistical Software, 9(1), 1-19.

[6] Potter, F.J. (1990), A study of procedures to identify and trim extreme sampling weights, Stata Journal, 7(1), 1.

[7] Snijder, T., and Bosker, R. (2012). Multilevel anaylsis: An introduction to basic and advanced multilevel modeling. In (2nd ed., p. 216-246). Sage Publications.

[8] van Buuren, S., and Groothuis-Oudshoorn, K. (2011). MICE: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45 (3), 1-67.

[9] Statistisches Bundesamt (2011), Fachserie 11 Reihe 4.1: Bildung und Kultur: Studierende an Hochschulen: Wintersemester 2010/2011, Stable URL: https://www.destatis.de/GPStatistik/servlets/MCRFileNodeServlet/DEHeft_derivate_00006845/2110410117004.pdf.

A Results of Nonresponse Modeling

Table 6: Modeling Dropouts from the Sample of Students who Provided any Kind of Contact Information to the Sample of Students with Valid Contact Information, i.e., to the Gross Sample of Wave 1 (Corresponding to the Study B52).

Variable	**Reference Category**	**Estimated**	**P-Value**
Gender	female		
male		-0.152	0.000
not specified		1.179	0.009
Nationality	German		
foreign		-0.198	0.003
unknown		-0.498	0.279
Type of institution	university		
Fachhochschule		0.067	0.047
not specified/abroad		0.292	0.000
Year of birth	1989 or earlier		
1990 - 1995		-0.057	0.049
not specified		-1.187	0.000
Intended degree	Bachelor		
Staatsexamen		0.154	0.004
Lehramt		0.324	0.000
other, unknown		-0.412	0.000
Type of contact (WS 2010/11)	personal		
postal		0.758	0.000
Number of cases	26,913		

Notes: (i) The calculations were performed by *infas - Institut für angewandte Sozialwissenschaften GmbH.* (ii) Among the 31,082 first-year students who could be contacted, only 26,913 students provided valid information on the variables considered in this model. We assume no selection bias by omitting the set of students with invalid or partial information. Nonetheless, at a later stage we use poststratification to correct for potential bias.

Table 7: Modeling Dropouts from the Gross Sample of Wave 1 (i.e., Study B52) to Current Participation.

Variable	**Reference Category**	**Estimated**	**P-Value**
Gender	female		
male		-0.109	0.040
not specified		0.072	0.937
Nationality	German		
foreign		-0.732	0.000
unknown		-0.826	0.413
Type of institution	university		
Fachhochschule		-0.136	0.030
not specified/abroad		-0.580	0.393
Year of birth	1989 or earlier		
1990 - 1995		-0.007	0.896
not specified		0.171	0.724
Intended degree	Bachelor		
Staatsexamen		0.018	0.855
Lehramt		0.093	0.161
other, unknown		-0.256	0.196
Type of contact (WS 2010/11)	personal		
postal		0.382	0.000
Instrument	CATI		
without telephone number		0.080	0.172
Attempts to contact target	1 to 3 attempts		
4 to 6 attempts		0.136	0.092
7 to 10 attempts		0.083	0.443
More than 10 attempts		-2.189	0.000
Number of cases	18,030		

Notes: (i) The calculations were performed by *infas - Institut für angewandte Sozialwissenschaften GmbH.* (2) Among the 21,438 first-year students who could be contacted, only 18,030 students provided valid information on the variables considered in this model. We assume no selection bias by omitting the set of students with invalid or partial information. Nonetheless, at a later stage we use post-stratification to correct for potential bias.

Table 8: Modeling Participation in Wave 2 (i.e., Study B54).

Variable	**Reference Category**	**Estimated**	**P-Value**
Type of institution	*Fachhochschule*		
university		0.099	0.004
Gender	female		
male		-0.129	0.000
Change of study programme	no		
yes		-0.157	0.000
Type of university	private		
public		-0.146	0.053
Intended degree Bachelor	yes		
no		-0.072	0.022
Enjoys studying	yes		
moderate		-0.158	0.000
no		-0.210	0.001
Migration background	no		
first generation		-0.142	0.001
second/third generation		-0.095	0.002
Household size	one		
two or more		-0.106	0.000
Reading ability	bad		
moderate		0.138	0.007
good		0.238	0.001
very good		0.294	0.001
Random effect (std. dev.)			
cluster		0.149	
Number of cases	17,910		

Table 9: Modeling Participation in Wave 3 (i.e., Study B55).

Variable	Reference Category	Estimated	P-Value
Participation in B54	no		
yes		0.596	0.000
Gender	female		
male		0.103	0.000
German is first language	no		
yes		0.241	0.000
Father has university degree	no		
yes		0.058	0.006
Intended degree Bachelor	yes		
no		0.155	0.000
Enjoys studying	yes		
moderate		-0.075	0.015
no		-0.157	0.010
Migration background	no		
first generation		-0.077	0.098
second/third generation		-0.071	0.028
Household size	one		
two or more		0.092	0.000
Kids in the household	no		
yes		0.288	0.000
Year of birth	<1988		
1988/1989		0.026	0.438
1990		0.104	0.001
>1990		0.078	0.024
Random effect (std. dev.)			
cluster		0.065	
Number of cases	17,910		

Table 10: Modeling Participation in Wave 4 (i.e., Study B56).

Variable	**Reference Category**	**Estimated**	**P-Value**
Participation in B54	no		
yes		1.050	0.000
Participation in B55	no		
yes		0.713	0.000
Gender	female		
male		-0.077	0.001
German is first language	no		
yes		0.206	0.001
Mother has university degree	no		
yes		0.047	0.034
Intended degree Bachelor	yes		
no		0.113	0.000
Enjoys studying	yes		
moderate		-0.087	0.006
no		-0.070	0.272
Reading ability	bad		
moderate		0.132	0.014
good		0.217	0.004
very good		0.265	0.002
Year of birth	<1988		
1988/1989		0.063	0.058
1990		0.028	0.389
>1990		0.058	0.096
Random effect (std. dev.)			
cluster		0.105	
Number of cases	17,910		

Table 11: Modeling Participation in the Interview of Wave 5 (i.e., Study B59).

Variable	**Reference Category**	**Estimated**	**P-Value**
Participation in B54	no		
yes		0.230	0.000
Participation in B55	no		
yes		1.173	0.000
Participation in B56	no		
yes		0.597	0.000
Gender	female		
male		0.082	0.001
German is first language	no		
yes		0.187	0.004
Intended degree Bachelor	yes		
no		0.121	0.000
Migration background	no		
first generation		-0.067	0.180
second/third generation		-0.084	0.015
Household size	one		
two or more		0.105	0.000
Kids in the household	no		
yes		0.226	0.000
Year of birth	<1988		
1988/1989		0.084	0.020
1990		0.117	0.001
>1990		0.117	0.002
Random effect (std. dev.)			
cluster		0.104	
Number of cases	17,910		

Table 12: Modeling Participation in Wave 6 (i.e., Study B58).

Variable	**Reference Category**	**Estimated**	**P-Value**
Participation in B54	no		
yes		0.518	0.000
Participation in B55	no		
yes		0.147	0.000
Participation in B56	no		
yes		1.021	0.000
Participation in B59	no		
yes		0.603	0.000
Mother has university degree	no		
yes		0.065	0.005
Enjoys studying	yes		
moderate		-0.078	0.016
no		-0.092	0.166
Migration background	no		
first generation		-0.150	0.001
second/third generation		-0.060	0.070
Kids in the household	no		
yes		-0.170	0.001
Year of birth	<1988		
1988/1989		0.051	0.146
1990		0.114	0.001
>1990		0.095	0.007
Random effect (std. dev.)			
cluster			
Number of cases	17,910	0.116	