NEPS
National Educational Panel Study

Research Data

Sampling and Weighting the Sample of the Early Childhood Cohort of the National Educational Panel Study (Waves 1 and 2)

Technical Report on SUF SC1 Version 2-0-0

Christian Aßmann, Sabine Zinn, and Ariane Würbach

LIfBi
LEIBNIZ INSTITUTE FOR EDUCATIONAL TRAJECTORIES

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, 2015

Sampling and Weighting the Sample of the Early Childhood Cohort of the National Educational Panel Study (Waves 1 and 2)

Technical Report on SUF SC1 Version 2-0-0

Christian Aßmann, Sabine Zinn, and Ariane Würbach

Corresponding address: methoden@lifbi.de

November 2015

1 Prequel

This report documents the target population, the sampling, the sample size, and the weighting procedures of Wave 1 and 2 of Starting Cohort 1 (Early Childhood) of the National Educational Panel Study (NEPS).[1] In the following, Section 2 documents the target population of the Starting Cohort and the applied sampling design. In addition, the composition of the gross and net sample is provided. In Section 3, the derivation of the sampling weights are described in detail. This includes the computation of the design weights and the accordant nonresponse adjustments. Section 4 concludes with a summary of the sampling weights provided and comments on the use of weights when analyzing data.

2 Population, Sampling Design, and Sample Sizes

2.1 Population and Sampling Design

The target population of Starting Cohort 1 comprises children born in Germany from February 2012 to July 2012. To achieve valid measurements of infant development, surveyed children were required to be at least six months but no more than eight months old at the age of the survey. Access to this population had been gained via a register-based sample of addresses available at the level of municipalities. Based on data from the first half of 2009, the distribution of births in 2012 was expected to be highly unequal between municipalities; compare Table 1. To guarantee nevertheless a meaningful coverage of municipalities, the measure of size for selecting municipalities was determined to be proportional to the number of children born within these municipalities in the first half of 2009.[2] The selection of addresses was performed via a two-stage disproportional stratified sampling. As primary sampling units municipalities were drawn, explicitly stratified according to a classification of urbanization (BIK scale).[3] That is, all 6,472 municipalities concerned[4] were assigned to three strata:

1. Less than 50,000 inhabitants,
2. 50,000 to 500,000 inhabitants, and
3. 500,000 and more inhabitants.

As secondary sampling units addresses of newborns within the selected municipalities were sampled. Commonly, for administrative reasons within municipalities only multiples of a fixed quantum of addresses can be sampled. Therefore, the overall goal to sample addresses of individuals was achieved via sampling artificial units called sample points. For Starting Cohort 1, a quantum of $\bar{c} = 300$ per municipality had been sampled. This number was expected allowing for achieving the planned net sample size even with unlikely low participation rates.[5] Given this

[1] Wave 1 corresponds to the study B04 and Wave 2 of the study B05. Details on both studies are given in Bauer et al. (2013), and Aust and Bauer (2014a, 2014b).

[2] At time of sampling only data from 2009 was available.

[3] Note: no stratification according to Federal States was considered. Such stratification would increase the number of stratification cells vastly and the number of observations in each cell would be remarkably low.

[4] All German municipalities registered in 2009, with exclusion of municipalities having less than ten births in 2009.

[5] Further, (almost all) registration offices only provide samples of at most 50% of semiannual births.

Table 1: Distribution of births across municipalities

	Municipalities with at least ten births in 2009	Number of semiannual births in the first half of 2009	
	abs.	abs.	%
Less than 50,000 inhabitants	6,285	178,993	55.3
50,000 up to 500,000 inhabitants	173	81,854	25.3
500,000 or more inhabitants	14	62,674	19.4
Total	6,472	323,521	100.0

design, simulation studies were used to determine the number of required municipalities to reach a planned sample size of approximately 3,000 newborns. In the end, 90 sampling points in 84 municipalities were found to be sufficient to reach a planned sample size of approximately 3,000 newborns. Within each explicit level the same number of sampling points had been sampled and between levels a different number of sampling points, compare Table 1. In the selected municipalities, addresses were then sampled from the 2010 register data within two tranches. Dividing addresses into tranches facilitated accounting for the time span registration offices need to register current births and to sample addresses, always minding the infants' age range required. The first tranche considered comprises births from February to April, whilst the second tranche comprises births within the months of May to July. Opposed to Wave 1, Wave 2 comprises two subsamples: (i) all being participants from the parent interview with panel consent and (ii) a random sample of the targets with direct competence measurements. The latter was established by drawing via simple random sampling 34 municipalities from the 84 municipalities of Starting Cohort 1. Here, all individuals involved were asked for participation. This way administrative burden and costs involved in individual home-testing of children could be reduced.

2.2 Sampling Sizes

Starting from a gross sample size of 8,483 persons established via sampling of addresses within 84 municipalities, the realized sample size in Wave 1 is 3,481 (corresponding response rate 41%), whereas the realized sample size of Wave 2 parent interviews is 2,849 (corresponding to a participation rate of 82%).[6] Wave 2 comprises a direct measurements gross sample of 2,021 persons, of whom 1,510 interviews have been realized.[7] The accordant gross and net samples sizes are also given in Table 2.

3 Derivation of Design Weights

Calculation of design weights derives directly from the sampling design, that is, from the sampling probabilities. In more detail: the pure design weights are calculated as inverse inclusion

[6] 42 participants gave no panel consent in B04 and in the meantime 8 participants have withdrawn their panel consent.

[7] Please note that the number of direct measurements eligible for evaluation can be lower.

Table 2: Gross and net sample sizes of Wave 1 and 2

	Gross sample	Net sample
Wave 1	8,483	3,481
Wave 2 (parent)	3,481	2,849
Wave 2 (direct measurem.)	2,021	1,510

probabilities, respecting the disproportional stratification. That is, assuming an individual inclusion probability π, its corresponding design weight is $1/\pi$. First stage sampling was performed based on an allocation of a fixed number s_l of sample points to each stratum $l = 1, \ldots, 3$. Each stratum comprises m_l municipalities summing up to $\sum_{l=1}^{L} m_l = 6,472$ (see Table 1). Because each sampling point corresponds to a fixed quantum of addresses, $s_l\bar{c} = 300s_l$ gives the number of addresses from which to sample within stratum l. A total of s_l municipalities was sampled from each stratum l with replacement, where each municipality is sampled proportional to size (pps). The respective measure of size is given as

$$N_{ml}/N_l, \quad m = 1, \ldots, m_l,$$

with N_{ml} denoting the number of semiannual births observed in the first half of 2009 within municipality m within stratum l, and N_l denotes the total number of addresses available in stratum l. Beware that this approach allows for a repeated sampling of individual municipalities. This implies assigning to a municipality m multiple sampling points, say s_{ml}. Then, a total of $\bar{c}s_{ml}$ addresses is sampled from the available N_{ml}. Thus, the sampling probability of an individual address i in stratum l in municipality m can be given as

$$\pi_{ilm} = \frac{\bar{c}s_{ml}}{N_{ml}} \frac{s_l N_{ml}}{N_l} = \frac{\bar{c}s_{ml}s_l}{N_l}.$$

(For reasons of clarity, subsequently the indices l and m are omitted. Thus, π_{ilm} simplifies to π_i.)

4 Weighting Adjustments for Wave Participation

Systematic refusals may arise and for this, the response and nonresponse processes of sampled individuals, has to be accounted for. Thus, for reasons of usability, commonly design weights are adjusted to account for nonresponse in the survey. For this purpose, the units' probabilities to participate in each survey wave are employed.[8] To highlight possible effects of participation on the sample, a comparison of the gross sample and the realized sample of Wave 1 is shown in Table 3. Note that only few characteristics of the newborns are known in advance from the registration offices. The descriptives are thus restricted to this set of model parameters. Only minor differences exist between the gross and the realized sample.

[8] In Starting Cohort 1 the target population are newborns but the respondents are their legal guardians. Hence, in this particular case it would be more appropriate to use the term realization probability instead of participation probability. Nevertheless, realization probability is not commonly used in this respect, therefore it is waived.

Table 3: Comparison of gross sample and realized sample of Wave 1

	Gross sample		Realized interviews	
	abs.	%	abs.	%
Federal state				
Schleswig-Holstein	325	3.8	125	3.6
Hamburg	645	7.6	249	7.2
Niedersachsen	461	5.4	205	5.9
Bremen	217	2.6	84	2.4
Nordrhein-Westfalen	2,330	27.5	893	25.7
Hessen	689	8.1	314	9.0
Rheinland-Pfalz	181	2.1	72	2.1
Baden-Württemberg	835	9.8	346	9.9
Bayern	1,029	12.1	461	13.2
Saarland	150	1.8	37	1.1
Berlin	867	10.2	358	10.3
Brandenburg	41	0.5	17	0.5
Mecklenburg-Vorpommern	153	1.8	75	2.2
Sachsen	353	4.2	158	4.5
Sachsen-Anhalt	168	2.0	72	2.1
Thüringen	39	0.5	15	0.4
BIK categories				
Less than 50,000 inhabitants	1,291	15.2	577	16.6
50,000 up to 500,000 inhabitants	3,517	41.5	1,345	38.6
500,000 or more inhabitants	3,675	43.3	1,559	44.8
Citizenship based on register data				
German	7.459	87.9	3.181	91.4
Non-German	555	6.5	151	4.3
Unknown	469	5.5	149	4.3
Gender				
Male	4,390	51.8	1,774	51.0
Female	4,090	48.2	1,707	49.0
Unknown	3	0.0	0	-
Total	8,483	100.0	3,481	100.00

4.1 Modeling Wave 1 Participation

The participation probability π_1 of Wave 1 is estimated by means of logistic regression. In the considered case, only a small set of explanatory variables is available, but with some additional information from the contact history. To control for accessibility the number of attempts to contact a respondent is used. That is, information for nonresponse adjustment refers predominantly to the given characteristics of the newborns. The set of variables incorporated within the regression and the resulting odds ratios are given in Table 4. Overall, the regression points to only modest selectivity with respect to the gross sample. Individuals with non-German citizenship show a slightly lower probability of participation in the survey than individuals with German citizenship. The resulting nonresponse adjusted weight for Wave 1 is

$$w_{1i} = \frac{1}{\pi_i \pi_1}.$$

Table 4: Results of the logistic regression model for Wave 1 participation

Value	Reference Category	Odds Ratio	p-Value
Attempts to contact person	1 up to 3 attempts		
4 up to 6 attempts		1.880	<0.001
7 up to 10 attempts		1.181	0.102
11 or more attempts		1.053	0.790
Birth month	April		
February		1.108	0.166
March		0.950	0.478
May		0.906	0.183
June/July		0.879	0.088
Gender	Female		
Male		0.940	0.171
Citizenship	German		
Non-German		0.511	<0.001
Federal state	Nordrhein-Westfalen		
Schleswig-Holstein		1.059	0.645
Hamburg		0.851	0.120
Niedersachsen		1.308	0.012
Bremen		0.908	0.531
Hessen		1.303	0.003
Rheinland-Pfalz		1.094	0.577
Baden-Württemberg		1.041	0.637
Bayern		1.123	0.160
Saarland		0.525	0.001
Berlin		0.927	0.425
Brandenburg		0.976	0.941
Mecklenburg-Vorpommern		1.377	0.061
Sachsen		1.106	0.410
Sachsen-Anhalt		1.156	0.383
Thüringen		0.776	0.456
BIK categories	Less than 50,000 inhabitants		
50,000 up to 500,000 inhabitants		0.817	0.006

500,000 or more inhabitants		1.052	0.517
Intercept		0.669	<0.001
Pseudo R^2 (McFadden)	0.024		
Number of cases	8,483		

4.2 Modeling Wave 2 Participation

By design, further wave participation was determined to be conditional on the willingness of Wave 1 participants to continue participation.[9] All participants of the Wave 1 interview gave their consent to participate in future waves. Thus, to specify individual participation probabilities, panel consent was not required to be modeled (since all participants were positive about continuing in the survey). The participation probabilities of the two subsamples of Wave 2 are denoted as $\pi_{2,parent}$ and $\pi_{2,target}$. Again, logistic regression models are used to quantify the participation probabilities. Variables used for this purpose are highlighted in Table 5 and Table 6. Table 5 shows in addition the results of the logistic regression of the participation probability of the parents. The results concerning factors influencing the participation of the target in the direct measurements are provided in Table 6. In conclusion, the estimates of the logistic regression models show only modest selectivity effects in the net samples. Influential factors refer to citizenship or migration background as well as to education of parents. The corresponding nonresponse adjusted weights for the Wave 2 are

$$w_{2i,parent} = \frac{1}{\pi_i \pi_1 \pi_{2,parent}} \quad \text{and} \quad w_{2i,target} = \frac{1}{\pi_i \pi_1 \pi_{2,target}}.$$

Table 5: Results of the logistic regression model for Wave 2 participation in the parent interview

Value	Reference Category	Odds Ratio	*p*-Value
Birth month	February		
March		0.787	0.088
April		0.973	0.865
May		1.056	0.717
June/July		1.134	0.425
Gender	Female		
Male		1.014	0.889
Citizenship child	German		
Non-German		0.497	0.002
Number of siblings	none		
1		0.875	0.483
2 or more		0.935	0.784
Relationship of interviewed person and child	Father or other person		
Mother		1.171	0.652
Year of birth interviewed person	1981-1985		

[9]The accordant consent was requested at the end of the Wave 1 interview.

Before 1975		1.296	0.088
1976-1980		1.298	0.041
Later than 1986		0.696	0.010
Self-reported heath rating of interviewed person concerning child	Good		
Very good		1.094	0.462
Fair to poor		1.343	0.468
Living together with partner	No		
Yes		1.666	0.038
Size of household	3 persons		
2 persons		0.871	0.639
4 persons		1.092	0.643
5 or more persons		0.901	0.683
Migration background interviewed person	No		
Yes		0.563	<0.001
Migration background of parents of interviewed person	No		
Yes		0.857	0.321
Highest degree of interviewed person	Mittlere Reife/else		
None/Hauptschulabschluss		0.748	0.050
Abitur/Fachhochschulreife		1.483	0.001
Employment status of interviewed person	Unemployed		
Full time		0.624	0.094
Part time		1.092	0.632
Employment status of partner of interviewed person	Full time		
Part time		0.999	0.998
Unemployed		1.081	0.702
Net household income	2,500 up to 4,000 Euro		
Less than 1,500 Euro		1.021	0.922
1,500 up to 2,500 Euro		0.932	0,610
4,000 up to 6,000 Euro		1.207	0,248
6,000 Euro or more		1.041	0.877
Not available		0.685	0.069
Social benefits (AlgII/Sozialgeld) interviewed person	No		
Yes		0.921	0.661
Social benefits (AlgII/Sozialgeld) partner of interviewed person	No		
Yes		0.536	0.006
Completed competence tests Wave 1	No		
Yes, looking at pictures		1.145	0.463
Yes, playing with mother		1.399	0.064
Yes, playing with interviewer		1.259	0.242
BIK categories	Less than 50,000 inhabitants		
50,000 up to 500,000 inhabitants		1.051	0.716
500,000 or more inhabitants		1.559	0.002
Federal region	West (without Berlin)		

East (including Berlin)		0.760	0.033
Pseudo R^2 (McFadden)	0.101		
Number of cases	3,431		

Table 6: Results of the logistic regression model for Wave 2 participation in direct measurements (target)

Value	Reference Category	Odds Ratio	p-Value
Birth month	February		
March		0.944	0.729
April		0.875	0.461
May		0.802	0.199
June/July		0.875	0.447
Gender	Female		
Male		0.977	0.836
Citizenship child	German		
Non-German		1.106	0.758
Number of siblings	None		
1		0.984	0.947
2 or more		0.834	0.535
Relationship of interviewed person and child	Father or other person		
Mother		1.084	0.848
Year of birth interviewed person	1981-1985		
Before 1975		1.195	0.292
1976-1980		0.976	0.859
Later than 1986		0.731	0.079
Self-reported heath rating of interviewed person concerning child	Good		
Very good		1.158	0.306
Fair to poor		0.945	0.901
Living together with partner	No		
Yes		1.595	0.129
Size of household	3 persons		
2 persons		0.670	0.298
4 persons		0.974	0.912
5 or more persons		1.113	0.727
Migration background interviewed person	No		
Yes		0.545	<0.001
Migration background of parents of interviewed person	No		
Yes		0.797	0.175
Highest degree of interviewed person	Mittlere Reife/else		
None/Hauptschulabschluss		0.692	0.052
Abitur/Fachhochschulreife		1.358	0.025

Employment status of interviewed person	Unemployed		
Full time		0.708	0.335
Part time		0.891	0.557
Employment status of partner of interviewed person	Full time		
Part time		0.708	0.136
Unemployed		1.062	0.796
Net household income	2,500 up to 4,000 Euro		
Less than 1,500 Euro		1.065	0.815
1,500 up to 2,500 Euro		1.090	0.591
4,000 up to 6,000 Euro		1.007	0.964
6,000 Euro or more		1.658	0.099
Not available		0.524	0.005
Social benefits (AlgII/Sozialgeld) interviewed person	No		
Yes		1.148	0.549
Social benefits (AlgII/Sozialgeld) partner of interviewed person	No		
Yes		0.674	0.153
Completed competence tests Wave 1	No		
Yes, looking at pictures		1.372	0.137
Yes, playing with mother		2.575	<0.001
Yes, playing with interviewer		1.396	0.141
Pseudo R^2 (McFadden)	0.095		
Number of cases	2,021		

5 Calibration to External Benchmark Totals

For the considered population, only few relevant and valid benchmark totals are available. The typically used source for calibration to benchmark totals–the German microcensus as 1% sample of the total German population–lacks in precision for the considered population. However, a reasonable population total is given via the total number of births per month per municipality. When detailed information from the German statistical office based on the German census 2011 becomes available, accordingly calibrated weights can be provided. In meantime, weights calibrated to corresponding figures available for 2009 are provided upon request.

6 Provision and Use of Weights

To ease statistical analysis, all weights are provided in a trimmed and standardized form. Trimming was conducted at the 95th percentile in order to remove outliers. Standardized weights have mean one and sum up to the number of participants in the corresponding wave. Table 7 summarizes all types of weights that are provided and their accordant label. In Table 8 some summary statistics for the standardized weights are presented.

Table 7: Types of weights provided

Type of weight	Standardized with mean one
Nonresponse adjusted weight Wave 1	`w_t1`
Nonresponse adjusted weight Wave 2, parent	`w_p2`
Nonresponse adjusted weight Wave 2, target/child	`w_t2`

Table 8: Summary statistics for standardized weights

Label of weight	Number of newborns	Min.	Lower Quart.	Median	Mean	Upper Quart.	Max.
`w_t1`	3,481	0.038	0.095	0.132	1.000	0.275	5.956
`w_p2`	2,849	0.233	0.392	0.494	1.000	0.733	9.248
`w_t2`	1,510	0.289	0.454	0.560	1.000	0.807	9.267

No general recommendations are at hand concerning the use of design and nonresponse adjusted weights. Whether and how weights should be used depends on the analysis considered. The use of weights is recommended in descriptive analysis. However, there are no general results available on how to use nonresponse adjusted design weights in statistical inference, see Rohwer (2011) for a general discussion. The use of weights may possibly help to highlight important features of the analysis under consideration, not least serving as a robustness check for the analysis performed. Generally, models have to be tested for their dependence on the sampling design. Concretely, this means that the user has to ensure that the way of sampling has no or only a negligible effect on the model results or/and that the sampling design is considered in the model definition adequately. A general description of how to test and account for the sampling design is given in Snijder and Bosker (2012). Two possible strategies exist to include weights in the analysis. First, in the model-based approach, all variables employed for construction of the weights are included as explanatory variables into the model under consideration. In the second (design-based) approach design information and weights are directly included into the model. As a guideline, we recommend the first strategy.

For further information on weighting please contact `methoden@lifbi.de`

References

[1] Aust, F., & Bauer, A. (2014a). *Methodenbericht NEPS Startkohorte 1 - Haupterhebung 2013 B05 (CATI-Erhebung)*, Bonn: infas Institut für angewandte Sozialwissenschaft GmbH.

[2] Aust, F., & Bauer, A. (2014b). *Methodenbericht NEPS Startkohorte 1 - Haupterhebung 2013 B05 (CAPI-Erhebung)*, Bonn: infas Institut für angewandte Sozialwissenschaft GmbH.

[3] Bauer, A., Bech, K., Gilberg, R. & Kleudgen, M. (2013). *Methodenbericht NEPS Startkohorte 1 - Haupterhebung 2012/2013 B05*, Bonn: infas Institut für angewandte Sozialwissenschaft GmbH.

[4] Snijder, T., & Bosker, R. (2012). *Multilevel anaylsis: An introduction to basic and advanced multilevel modeling*, 2nd ed., Los Angeles [o.a.]: Sage Publications, 216-246.

[5] Rohwer, G. (2011, April). Use of probabilistic models of unit nonresponse, Version 1. Retrieved from http://www.stat.ruhr-uni-bochum.de/papers/dun.pdf.