NEPS
National Educational Panel Study

Research Data

# Starting Cohort 2: Kindergarten (SC2) SUF-Version 1.0.0 Data Manual [Supplement]: Weighting

*Christian Aßmann, Solange Koch, Hans Walter Steinhauer, and Sabine Zinn*

SPONSORED BY THE

Federal Ministry of Education and Research

# Weighting the Kindergarten Cohort of the National Educational Panel Study, Panel Cohort

Technical Report

Christian Aßmann, Solange Koch, Hans Walter Steinhauer, and Sabine Zinn*

National Educational Panel Study
University of Bamberg

*version: September 21, 2012*

*methods.neps@uni-bamberg.de

# 1 Structure of sample

Based on a short review of the survey and sampling design, this report provides the results of the response propensity analysis for the the sample of kindergarten children surveyed within the National Educational Panel Study (NEPS). This sample focuses on children attending kindergartens in Germany in the school year 2010/2011 who are expected to begin schooling in school year 2012/2013. These children are thus at about the age of four years as children in Germany have to start attending primary schools at an age of five to seven years according to their date of birth, see Aßmann et al. (2011) for details. Kindergartens and primary schools thus are linked, i.e. children trespass from kindergartens to schools.[1] This link can be used to get access to the population of kindergarten children by using an indirect sampling approach as introduced by Lavallée (2007). In addition, direct multi-stage sampling approaches fail as no frame is available for either kindergarten children or kindergarten institutions. As a byproduct, an overlap between the sample of kindergarten children and the sample of primary school children established in 2012/2013 may result from the indirect sampling approach. The principles of the implemented indirect sampling approach developed by Lavallée (2007) can be stated as follows, see also Kiesl (2010) and Aßmann et al. (2011). Assume a population $U_P$ that is linked in a defined manner to another population $U_K$. Further there exists a sampling frame for the population $U_P$ so that a sample $s_P$ can be drawn. Via the definition of a link $\theta_{pk}$ between elements $p \in U_P$ and $k \in U_K$ relating to a certain measurement and a reference year access is gained to a sample $s_K$ in the linked population $U_K$. Extensions of indirect sampling techniques to multi-stage sampling as well as their properties are discussed in Lavallée (2007) for two-stage sampling and in Kiesl (2010) for two-stage and three-stage sampling.

We describe the methods used to calculate the weights for the participating children in Section 2. Section 2 also briefly addresses corrections for potential systematic non-response. Section 3 describes the trimming procedure used. Section 4 concludes with some general remarks on the use of weights.

# 2 Design weights and response rate analysis

For the indirect sampling of kindergartens, primary schools need to be sampled in first place. This sample of primary schools will then in turn serve in school year 2012/2013 to establish a sample of first grade children. Therefore, let $U_P$ be the population of $M_P(= 16824)$ primary schools with at least one class in grade one. A sample of $m_P(= 300)$ primary schools is drawn via a systematic proportional to size (pps) scheme with the number of students in grade one $S^1$ as measure of size, whereas the total number of students in grade one (i.e. the total measure of size) is

$$S^1 = \sum_{p=1}^{M_P} S_p^1 = 710539, \tag{1}$$

[1] Note that 92% of all children aged from three to five years visit a day care facility before going to a primary school, see Statistisches Bundesamt (2010) and Statistisches Bundesamt (2011).

where $S_p^1$ denotes the number of children in grade one in school $p$, $p = 1, \ldots, M_P$. The corresponding frame reference year is the school year 2008/2009. The resulting inclusion probabilities of primary schools are given as

$$\pi_p = m_P \cdot \frac{S_p^1}{S^1}, \quad p = 1, \ldots, m_P. \tag{2}$$

For the drawn sample of $m_P$ schools, the corresponding mean and standard deviation of weights are given as 95.77 and 93.54, respectively. While the $m_P$ schools serve as the basis for the sample of first grade children, only a simple random subsample of $\tilde{m}_P(= 212)$[2] schools serves as the basis for the kindergarten sample with corresponding inclusion probabilities $\tilde{\pi}_p = \frac{\tilde{m}_p}{m_p}\pi_p$. Based on the sample of $\tilde{m}_P$ primary schools, each of these school is asked to list the kindergartens from which the children of their first grade came in school year 2009/2010. These lists form the sets $\Omega_p$, $p = 1, \ldots, \tilde{m}_P$, containing $|\Omega_p| = K_p$ kindergartens for each sampled school, and thus relate to a subset of the population of kindergartens $U_K$. Note that the sets $\Omega_p$ need not to be pairwise disjoint. This subset can be used to provide a sample of kindergartens and kindergarten children, respectively. The link function $\theta_{pk}$ between kindergartens and schools is defined as the number of children trespassing from kindergarten $k$ to school $p$, i.e. $\theta_{pk} > 0$ if children trespass from kindergarten $k$ to primary school $p$ and zero else. If all listed kindergartens are surveyed, this setup would render to the following indirect weights of a kindergarten $k$ given as[3]

$$w_k = \sum_{p=1}^{\tilde{m}_P} \frac{\theta_{pk}}{\theta_{+k}\tilde{\pi}_p}, \quad \text{with} \quad \theta_{+k} = \sum_{p=1}^{M_P} \theta_{pk}. \tag{3}$$

For illustration of this issue consider the population total of any variable $Y$ in population $U_K$ to be written as

$$t_Y = \sum_{k \in U_K} y_k = \sum_{k \in U_K} \left( \sum_{p \in U_P} \frac{\theta_{kp}}{\theta_{+k}} \right) y_k, \quad \text{where} \quad \sum_{p \in U_P} \frac{\theta_{kp}}{\theta_{+k}} = 1. \tag{4}$$

That is, the total of some variable in population $U_K$ can be written as total of the fractionized variable $\sum_{k \in U_K} \frac{\theta_{kp}}{\theta_{+k}} y_k$. This approach relates to the weight share method of Ernst (1989).

As the number of kindergartens per school is prohibitively large to allow for a complete survey a subsampling of kindergartens is added, i.e. a subsample $s_{K|s_P}$ of all kindergartens linked to sampled schools is drawn. Following Kiesl (2010), the sampling of kindergartens is done school wise using a probability proportional to size sampling without replacement. The measure of size is defined via the number of children trespassing from kindergarten $k$ to school $p$. Subsampling is done by drawing from each set $\Omega_p$ a sample of size $k_p$ according

[2] Due to the intended sample size of approximately 3000 kindergarten children 212 out of the 300 sampled primary schools have been found sufficient to provide a list of kindergartens.

[3] If a kindergarten sends children only to a single primary school, $w_k$ corresponds to the Horwitz-Thompson weight given as $\frac{1}{\tilde{\pi}_p}$.

to the following rule,

$$k_p = \begin{cases} 1 \text{ if } 0 < K_p \leq 6 \\ 2 \text{ if } 6 < K_p \leq 11 \\ 3 \text{ if } 11 < K_p \leq 19 \\ 4 \text{ else.} \end{cases} \tag{5}$$

As noted by Kiesl (2010), the resulting (conditional) inclusion probabilities of kindergartens depend on the sample of primary schools $s_P$, i.e. for a given kindergarten this probability can differ due to different subsamples $s_{K|s_P}$ and even different samples $s_P$. Calculation of inclusion probabilities is thus not feasible, but allows for the construction of weights providing an unbiased estimator of population totals. Following Kiesl (2010), the weights are given as

$$w_k = \sum_{p \in s_P} \frac{\theta_{pk}}{\theta_{+k}} \cdot \frac{I(k \in \Omega_p)}{\pi_p \cdot \pi_{k \in \Omega_p}}, \tag{6}$$

where $\pi_{k \in \Omega_p}$ refers to the sampling probability of kindergarten $k$ listed by primary school $p$ and $I(k \in \Omega_p)$ is 1 if $k \in \Omega_p$ and 0 else. As all kindergarten children of the defined age group[4] are asked to participate, these weights correspond directly to children's weights. If a kindergarten enters the sample only via a single primary school, i.e. sends children only to one particular sample school, no summation over all primary schools is necessary and the weights can be rendered directly to a two stage sampling approach.

Note that this multistage sampling process is naturally also affected by non-response on the level of schools as well as on the level of kindergartens. The refusal of primary schools to participate in NEPS is before all field work related to the sampling of kindergarten children. Therefore, a replacement strategy has been adapted to cope with non-response on the level of primary schools. As the sampling of schools has been based on implicit stratification according to federal states, regional classification and organizing institutions, a non-participating school has been replaced by a school identical to the originally drawn school with regard to these stratification variables. The sampled kindergartens may also refuse participation within the survey. To address this problem, for each sampled kindergarten a set of replacement kindergartens has been selected from the same kindergarten list.[5] Selection was based on smallest deviation between sampled kindergarten and replacement kindergartens with respect to the number of children trespassing between kindergarten and school. When a kindergarten refuses participation or within a defined range of time does not provide explicit consent, the replacement kindergartens were asked to participate. For each kindergarten originally sampled there are two replacement kindergartens (if available).

To address potential selectivity within the panel cohort sample at the level of children, logit models regressing the participation status on information available for the gross-sample of the kindergarten children have been estimated. The set of variables available

[4] I.e., children at about four years in 2010/2011.

[5] Such processing is feasible because the kindergartens listed by a single school are similar with respect to regional aspects.

includes year of birth, gender, language spoken in household, residence and occupational status of the parents. In addition, a kindergarten specific random effect is considered to allow for potential correlation among children attending the same kindergarten. The empirical results for the corresponding random intercept logit model on the considered determinants are shown in Table 1. The results suggest, that older children and children who do not speak German at home have a lower propensity to participate within the survey. The same observation is found for children with no information available on characteristics of the child (i.e. gender or year of birth) and no information on the child's environment (i.e. language spoken at home, residence status or occupational status of parents). However, since the number of cases within these categories is low, in the realized sample effects of selectivity are not to be considered severe.

To obtain non-response adjusted weights, the inverse of the predicted response propensities are multiplied with the design weights $w_k$ of kindergarten children.

# 3 Weight trimming

To possibly increase the statistical efficiency of weighted analysis the adjusted weights have been trimmed. The general goal of weight trimming is to reduce the sampling variance and at the same time to compensate for potential increase in bias. Trimming has been 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 $\tau$ is computed whose occurrence probability is one percent, i.e., $1 - F_w(\tau) = 0.01$. Sampling weights in excess of $\tau$ 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 one. Subsequently, the distribution of the cohort sampling weights is summarized:

| Number of children | Min. | Lower Quart. | Median | Mean | Upper Quart. | Max. |
|---|---|---|---|---|---|---|
| 2996 | 0.073 | 0.489 | 0.752 | 1 | 1.152 | 4.113 |

# 4 Use of weights

Given the quite complex structure of the sample of the kindergarten children (first wave) no final recommendations are at hand concerning the use of design and adjusted weights. Although, there are no general results available how the use of design or adjusted weights render any possible analysis (see Rohwer (2011) for a general discussion) the use of weights may possibly help to highlight important features of the analysis under consideration not at least serving as a robustness check for the performed analysis. Adjusted cohort sampling

weights provided are labeled as `weight_design` and the standardized design weights are labeled as `weight_design_std`. The subsequent table lists all types of weights provided:

| Type of weight | Label |
|---|---|
| adjusted cohort sampling weight | `weight_design` |
| trimmed cohort sampling weight, standardized with mean one | `weight_design_std` |

The following syntax may be useful as a starting point when weights are incorporated into analysis using *Stata*:

```
* Put this before relevant command lines
svyset psu(su1) [pweight=pw]

* Commands
svy: command...
```

The `svyset` command is used to specify the sampling units used in clustering (via `psu`) and the sampling weights (via `pweight`). Here `su1` refers to an anonymous kindergarten ID as an indicator of the sampling units, `pw` gives the used weight variable. Note that given the above described multistage sampling process, the control for a single sampling stage only approximately reflects the whole sampling design. All commands following start with the prefix `svy`.

For further information on weighting please contact `methods.neps@uni-bamberg.de`.

## References


Aßmann, C., Steinhauer, H. W., Kiesl, H., Koch, S., Schönberger, B., Müller-Kuller, A., et al. (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* (Vol. 14, pp. 51–65). Wiesbaden: VS Verlag für Sozialwissenschaften.

Ernst, L. (1989). Weighting issues for longitudinal household and family estimates. *Panel Surveys*, 135-159.

Kiesl, H. (2010). Selecting kindergarten children by three stage indirect sampling. In American Statistical Association (Ed.), *Proceedings of the Survey Research Methods Section* (pp. 2730–2738). Alexandria.

Lavallée, P. (2007). *Indirect sampling.* New York: Springer.

Potter, F.J. (1990). A Study of procedures to identify and trim extreme sampling weights. In American Statistical Association (Ed.), *Proceedings of the Survey Research Methods Section* (pp. 225-230).

Statistisches Bundesamt. (2010). Bevölkerung und Erwerbstätigkeit, Bevölkerungsfortschreibung. *Fachserie 1, Reihe 1.3*.

Statistisches Bundesamt. (2011). Kinder und tätige Personen in Tageseinrichtungen und in öffentlich geförderter Kindertagespflege am 01.03.2011. *Statistiken der Kinder- und Jugendhilfe*.

Table 1: Response propensity model for kindergarten children.

| | |
|---|---|
| constant | 1.689** |
| | (0.638) |
| year of birth (2006) | −0.329 |
| | (0.636) |
| year of birth (before 2006) | −0.537 |
| | (0.638) |
| gender (male) | 0.010 |
| | (0.072) |
| language spoken at home (no German) | −0.653*** |
| | (0.101) |
| residence status (other) | −0.760 |
| | (0.700) |
| residence status (with single parent) | 0.207 |
| | (0.110) |
| residence status (with relatives) | 0.282 |
| | (0.507) |
| occupational status of parents (one employed) | −0.170* |
| | (0.084) |
| occupational status of parents (none employed) | −0.154 |
| | (0.149) |
| Missing values in variables related to the child's environment | −0.849*** |
| | (0.173) |
| Missing values in variables related to the child | −3.609*** |
| | (0.761) |
| random intercept at kindergarten level | 1.203 |
| | (1.097) |
| sample size | 4556 |

Notes: ***, ** and * denote significance at the 0.1%, 1% and 5% level respectively.