Selecting Kindergarten Children by Three Stage Indirect Sampling

Hans Kiesl*

Abstract
One cohort of the German National Educational Panel Study will consist of a sample of kindergarten children. No nationwide frame of kindergartens is available in Germany, contrary to the situation for primary schools. Following the works of Lavallée, we present a solution by indirect sampling, using links between kindergartens and primary schools. In the first stage, primary schools are randomly selected and all kindergartens are identified that have a link to the sampled schools. In the second stage, for every sampled school we take a random sample of kindergartens linked to this school. In the third stage, samples of children within the sampled kindergartens are selected. Using the links of the sampled kindergartens to all schools in the population, unbiased estimation for the population of children in kindergartens is possible. Our second stage sampling is in addition to the existing literature.

Key Words: indirect sampling, school sampling, educational survey

1. Introduction
The National Educational Panel Study (NEPS) is a new longitudinal educational survey in Germany with a complex sampling design. Six different samples from various age cohorts are drawn from the population and followed over time. One of the six target populations are children aged 4 (in 2010) who attend a kindergarten (or nursery school). Unfortunately, no complete list of all kindergartens in Germany, which could be used as a sampling frame, is available. Therefore it is not possible to draw a sample of kindergartens directly.

In situations like this, indirect sampling might be an alternative. Here, a sample $s_A$ is drawn from some other population $U_A$ whose units are linked to the units of the target population $U_B$; the sample $s_B$ from $U_B$ then consists of all units from $U_B$ that are linked to some unit in $s_A$. It is then possible to construct unbiased estimators from $s_B$; the basic reference for the theory of indirect sampling is Lavallée (2007). In our application, a (direct) sample from the population of primary schools in Germany (for which a sampling frame is available) is drawn; we then use links between kindergartens and primary schools to end up with an indirect sample of kindergartens. Since it is not feasible for budget reasons to use the complete indirect sample of kindergartens, we have to add another stage of subsampling; theory for this is quite straightforward, but has not been published in the literature yet.

The remainder of this paper is organized as follows. In section 2 we give a short overview of the theory of indirect sampling. In section 3 we present some theory for unbiased estimation after an additional subsampling stage. We then discuss in section 4 how this sampling and estimation procedure may be applied to construct a kindergarten sample for NEPS.

*Regensburg University of Applied Sciences, Fakultaet IM, Postfach 12 03 27, 93025 Regensburg, Germany, hans.kiesl@hs-regensburg.de
2. Review of Indirect Sampling

2.1 The Central Idea of Indirect Sampling

The idea underlying indirect sampling is rather simple. Consider the task to estimate the total \( t_Y \) of a variable \( Y \) in some population \( U_B \). If a sampling frame for \( U_B \) were available, we would (directly) sample from this frame, using any suitable sampling design with inclusion probabilities \( \pi_{ib}^B > 0 \) for every \( i \in U_B \). To get a design-unbiased estimator for \( t_Y \), we could use the Horvitz-Thompson-estimator (HT-estimator) \( \hat{t}_{Y,HT} = \sum_{sB} \frac{y_i}{\pi_{ib}^B} \), where the summation is over all units in the sample \( s_B \).

In certain applications no sampling frame for \( U_B \) is available. We might, however, have a sampling frame for a population \( U_A \), whose elements are somehow “linked” to the elements in the population of \( U_B \). A natural idea is then to draw a sample \( s_A \) from \( U_A \) (with known inclusion probabilities) and subsequently choose all elements from \( U_B \) that are linked to elements in \( s_A \) and define them as the sample \( s_B \) from \( U_B \). Because of the indirect selection of elements from \( U_B \), this procedure might be called indirect sampling.

The question remains, whether (and how) unbiased estimation for \( t_Y \) from \( s_B \) is possible, since the calculation of inclusion probabilities for \( s_B \) might be difficult or impossible under this setting.

Obviously, the properties of the “links” are crucial to this. If there are units in \( U_B \) that have no links to any unit in \( U_A \), they cannot get into the sample \( s_B \) and thus have inclusion probabilities of 0, which rules out the possibility of unbiased estimation. Suppose on the other hand that every unit in \( U_B \) is linked to exactly one unit in \( U_A \) (but units in \( U_A \) might be linked to more than one unit in \( U_B \)). In this case, the described procedure reduces to (one-stage) cluster sampling. We are thus mainly concerned with situations, where units in \( U_B \) might be linked to more than one element in \( U_A \).

The theory behind indirect sampling (although not yet called that way) started with Ernst (1989) in the context of longitudinal household surveys. A sample of households is drawn at time \( t_0 \) from all existing households (population \( U_A \)). Since the composition of households changes over time, the population of households changes as well. Depending on the follow-up rules of the longitudinal surveys (see Rendtel and Harms (2009) for a discussion on this topic), at time \( t_1 \) one ends up with a sample of households from \( U_B \), the population of all existing households at \( t_1 \). Links between households \( a \in U_A \) and \( b \in U_B \) might be defined by individuals living in household \( a \) at time \( t_0 \) and in household \( b \) at \( t_1 \).

In several papers (e.g. Lavallée 1995, Lavallée and Deville 2002) the theory was generalized to other situations. Lavallée (2007) is a comprehensive treatment of indirect sampling with theory and applications, although the notation gets a bit complicated. We review the basic theory in the next subsection, following the ideas of Lavallée, but trying to come up with a simplified notation.

2.2 Unbiased Estimation from Indirect Samples

Let \( U_A \) and \( U_B \) be two populations, and let \( \theta \) be a non-negative function (the “link function’) on \( U_A \times U_B \), i.e. for every \( a \in U_A \) and every \( b \in U_B \) we have \( \theta_{ab} \geq 0 \). We say that a link exists between \( a \in U_A \) and \( b \in U_B \), if and only if \( \theta_{ab} > 0 \). We then call \( \theta_{ab} \) the “weight” of the link between \( a \) and \( b \). (The question how best to define this link function depends on the application at hand.)

For every \( b \in B \), let \( \theta_{+b} := \sum_{a \in U_A} \theta_{ab} \) be the sum of the weights of all links from \( U_A \) to \( b \in B \). We assume that \( \theta_{+b} > 0 \) for all \( b \in B \), i.e. there exists a link to every \( b \in B \). (Otherwise unbiased estimation is impossible.)
The key observation is that the total of any variable \( Y \) in the population \( U_B \) might be written as follows:

\[
t_Y = \sum_{b \in U_B} y_b = \sum_{b \in U_B} \left( y_b \cdot \frac{\sum_{a \in U_A} \theta_{ab} \pi_a}{\pi_b} \right) = \sum_{a \in U_A} \sum_{b \in U_B} \theta_{ab} \pi_a y_b = \sum_{a \in U_A} \tilde{y}_a = \tilde{t}_Y,
\]

with \( \tilde{y}_a := \sum_{b \in U_B} \frac{\theta_{ab}}{\pi_b} y_b \).

Thus, the total of \( Y \) in population \( U_B \) can be written as the total of the variable \( \tilde{Y} \) in population \( U_A \). Since inclusion probabilities for every \( a \in s_A \) are known, the HT-estimator may be used for unbiased estimation of \( \tilde{t}_Y \) and thus also of \( t_Y \).

Let \( s_A \) be a sample of elements from \( U_A \), and let \( \pi_a \) be the inclusion probability of \( a \in s_A \). The sample \( s_B \) is defined as the set of all units in \( U_B \) that have a link to some element of \( s_A \); more formally \( s_B := \{ b \in U_B | \theta_{ab} > 0 \text{ for some } a \in s_A \} \). The HT-estimator for the total of \( \tilde{t}_Y \) is now also an unbiased estimator for the total of \( t_Y \), thus we call it the indirect sampling estimator \( \hat{t}_{Y,IS} \) for the total of \( t_Y \):

\[
\hat{t}_{Y,IS} := \hat{t}_Y = \sum_{a \in s_A} \frac{\tilde{y}_a}{\pi_a} = \sum_{a \in s_A} \sum_{b \in s_B} \frac{\theta_{ab} y_b}{\pi_a} = \sum_{a \in s_A} \sum_{b \in s_B} \frac{\theta_{ab} y_b}{\pi_a} = \sum_{b \in s_B} w_{bs} y_b
\]

with weights \( w_{bs} = \sum_{a \in s_A} \frac{\theta_{ab}}{\pi_a \pi_b} \). (Equality (*) is due to the fact that by definition of \( s_B \) we have \( \theta_{ab} = 0 \) for \( a \in s_A \) and \( b \notin s_B \).)

Note that to evaluate \( \hat{t}_{Y,IS} \), we need to know \( \theta_{ab} \) for every \( a \in s_A \) and \( b \in s_B \), but we also need to know \( \theta_{+b} \) for every \( b \in s_B \). Note also that the weights \( w_{bs} \) are in general different from the inclusion probabilities \( \pi_b \), and they are sample dependent (therefore the subindex \( s \)): the weight of unit \( b \in s_B \) depends on which units in \( U_A \) that are linked to \( b \) are actually in the sample \( s_A \). Since a linear homogenous estimator of the form \( \sum_{b \in s_B} w_{bs} y_b \) is unbiased for \( t_Y \) if and only if \( E(w_{bs}) = 1 \) for every \( b \), it follows immediately that \( E(w_{bs} | b \in s_B) = 1/\pi_b \), i.e. the indirect sampling weights of a unit \( b \in s_B \) are in expectation equal to the Horvitz-Thompson-weights. (As a consequence, for units \( b \in s_B \) that have exactly one link to \( U_A \) the indirect sampling weight is equal to the Horvitz-Thompson-weight; thus in case of cluster sampling the indirect sampling estimator coincides with the HT-estimator).

Up to now, we assumed that \( U_B \) is the set of final sampling units. Now suppose that the elements of \( U_B \) are actually clusters of individual units, i.e. \( U_B \) is the set of primary sampling units (but note that the links are still defined between units in \( U_A \) and \( U_B \)). The value \( y_b \) of a cluster \( b \in U_B \) is now itself a total of individual values. Let \( y_{bi} \) be the value of the variable \( Y \) of the \( i \)-th element in cluster \( b \), let \( e_b \) be the set of all elements in cluster \( b \). Then, \( y_b = \sum_{i \in e_b} y_{bi} \). Suppose that we draw only a subsample from \( e_b \); in this case we do not observe \( y_{bi} \), but we can estimate it from the sample.

To be more precise, consider a second stage of sampling, independently within the clusters \( b \) of the first stage sample \( s_B \). The sample of cluster \( b \) is called \( s_b \). Let \( \pi_{i|b} \) be the (conditional) inclusion probability of unit \( i \) in cluster \( b \), given that \( b \in s_B \). Then \( y_b \) can be estimated by \( \hat{y}_b = \sum_{i \in s_b} \frac{y_{bi}}{\pi_{i|b}} \), and an estimator for \( t_Y \) may be defined as follows:

\[
\hat{t}_{Y,IS,sub} := \sum_{a \in s_A} \sum_{b \in s_B} \frac{\theta_{ab}}{\pi_a} \frac{\hat{y}_b}{\pi_b} = \sum_{a \in s_A} \sum_{b \in s_B} \frac{\theta_{ab}}{\pi_a} \sum_{i \in s_b} \frac{y_{bi}}{\pi_{i|b}} = \sum_{b \in s_B} \sum_{i \in s_b} w_{bi} y_{bi}
\]

with weights \( w_{bi} = \sum_{a \in s_A} \frac{\theta_{ab}}{\pi_a \pi_{i|b} \pi_b} \).
Lavallée (2007, section 5.1) calls this procedure two stage indirect sampling and shows that $\hat{t}_{Y,\text{IS,sub}}$ is unbiased for $t_Y$.

### 3. Subsampling of Indirect Samples

The indirect sampling estimators in Lavallée (2007) and in the previous section assumed that sample $s_B$ consists of all units of $U_B$ that are linked to the units in the direct sample $s_A$ from population $U_A$. In some applications, where $U_B$ is much larger than $U_A$ or where the number of links is highly variable among the units in $U_A$, this might result in a very large, or at least quite unpredictably sized sample $s_B$. From a practical point of view, it might be desirable to draw only a subsample of $s_B$ as the final sample from $U_B$. In the following, we discuss two different ways to do this, and call them three stage and three phase indirect sampling, respectively.

#### 3.1 Three Phase Indirect Sampling

Consider first the simple case that $U_B$ consists of final sampling units (we return to the case in which $U_B$ is a population of PSUs below). The indirect sampling procedure results in an indirect sample $s_B$ from $U_B$. Suppose we draw a subsample $s_{\text{fin}}B$ from $s_B$, using any particular sample design with (conditional) inclusion probabilities $\pi_b|s_A$ (i.e. $\pi_b|s_A$ is the probability that $b$ is chosen for the final indirect sample given the first stage direct sample $s_A$; note that for a given $b$ these probabilities can be different for different samples $s_B$ and even for the same sample $s_B$ resulting from different direct samples $s_A$). In the literature on direct sampling, this procedure is called two-phase sampling or double sampling; see e.g. Särndal et al. (1992, chapter 9).

Then, the following estimator is unbiased for the total $t_Y$ in $U_B$:

$$
\hat{t}_{Y,\text{IS,2phase}} := \sum_{b \in s_{\text{fin}}B} w_b^f y_b \quad \text{with} \quad w_b^f = \sum_{a \in s_A} \frac{\theta_{ab}}{\pi_a \cdot \pi_b|s_A \cdot \theta_b}.
$$

To see this, consider the following iterated expectations, where we first condition on the direct sample $s_A$ (denoting expectation over the first and second phase sampling by $E_1$ and $E_2$, respectively):

$$
E (\hat{t}_{Y,\text{IS,2phase}}) = E_1 \left( E_2 \left( \hat{t}_{Y,\text{IS,2phase}} | s_A \right) \right)
= E_1 \left( \sum_{a \in s_A} \sum_{b \in s_{\text{fin}}B} \frac{\theta_{ab}}{\pi_a \cdot \pi_b|s_A \cdot \theta_b} y_b | s_A \right)
= E_1 \left( \sum_{a \in s_A} \frac{1}{\pi_a} \cdot E_2 \left( \sum_{b \in s_{\text{fin}}B} \frac{\theta_{ab}}{\pi_b|s_A \cdot \theta_b} y_b | s_A \right) \right)
= E_1 \left( \sum_{a \in s_A} \sum_{b \in s_B} \frac{\theta_{ab}}{\pi_a \cdot \theta_b} y_b \right)
= E (\hat{t}_{Y,\text{IS}}) = t_Y
$$

Turning again to the situation that the elements of $U_B$ are actually clusters of individual units, we assume that we independently draw subsamples $s_b$ from every cluster $b$ of the final indirect sample $s_{B,\text{fin}}$. Let $\pi_i|b$ be the (conditional) inclusion probability of unit $i$ in cluster $b$. 

---

Section on Survey Research Methods – JSM 2010

2733
given that \( b \in s_B^{\text{fin}} \). Then \( y_b \) can be estimated by \( \hat{y}_b = \sum_{s_B} \frac{y_{bi}}{\pi_{ib}} \). We might call this a three-phase indirect sampling procedure: the first phase being the indirect sampling resulting in \( s_B \), the second phase being the subsampling resulting in \( s_B^{\text{fin}} \), the third phase being the subsampling from within the clusters in \( s_B^{\text{fin}} \).

Under this procedure, an unbiased estimator for \( t_Y \) may be defined as follows:

\[
\hat{t}_{Y,\text{IS},3\text{phase}} := \sum_{b \in s_B^{\text{fin}}} \sum_{i \in s_B} w_{bi}^t y_{bi} \quad \text{with} \quad w_{bi}^t = \sum_{a \in s_A} \frac{\theta_{ab}}{\pi_a \cdot \pi_{b|s_A} \cdot \theta_{+b}}.
\]

To prove the unbiasedness of \( \hat{t}_{Y,\text{IS},3\text{phase}} \), we calculate iterated expectations, where we first condition on the final indirect sample \( s_B^{\text{fin}} \) (denoting expectation over the first two sampling phases and the third phase by \( E_1 \) and \( E_3 \), respectively):

\[
E \left( \hat{t}_{Y,\text{IS},3\text{phase}} \right) = E_1 \left( E_3 \left( \hat{t}_{Y,\text{IS},3\text{phase}} \mid s_B^{\text{fin}} \right) \right)
\]

\[
= E_1 \left( \sum_{a \in s_A} \sum_{b \in s_B} \sum_{i \in s_B} \frac{\theta_{ab}}{\pi_a \cdot \pi_{b|s_A} \cdot \theta_{+b}} y_{bi} \mid s_B^{\text{fin}} \right)
\]

\[
= E_1 \left( \sum_{a \in s_A} \sum_{b \in s_B^{\text{fin}}} \frac{\theta_{ab}}{\pi_a \cdot \pi_{b|s_A} \cdot \theta_{+b}} \cdot E_3 \left( \sum_{i \in s_B} y_{bi} \mid s_B^{\text{fin}} \right) \right)
\]

\[
= E_1 \left( \sum_{a \in s_A} \sum_{b \in s_B^{\text{fin}}} \frac{\theta_{ab}}{\pi_a \cdot \pi_{b|s_A} \cdot \theta_{+b}} \cdot y_b \right)
\]

\[
= E \left( \hat{t}_{Y,\text{IS},2\text{phase}} \right) = t_Y.
\]

### 3.2 Three Stage Indirect Sampling

In this section we consider a slightly different sampling procedure that also allows for an unbiased estimation. With three phase indirect sampling, as described in the previous section, there is no guarantee that every unit in the first stage direct sample \( s_A \) is linked to a unit in the final indirect sample \( s_B^{\text{fin}} \), although this might be preferable in some applications (see section 4 for an example). In the following, we present an alternative procedure with the desired property.

Again, consider first the simpler case that \( U_B \) consists of final sampling units. The direct sampling procedure results in the direct sample \( s_A \) of \( U_A \). Consider now for any \( a \in U_A \) the set \( \Omega_a \) of all units in \( U_B \) that are linked to \( a \). (Note that \( U_B \) is the union of all \( \Omega_a \) \( a \in U_A \), but apart from the special case of cluster sampling, the \( \Omega_a \) need not be pairwise disjoint.) The idea is now to independently draw a subsample \( \Omega_a^{\text{sub}} \) from \( \Omega_a \) for every \( a \in s_A \). For any \( b \in \Omega_a \), let \( \pi_{b|\Omega_a^{\text{sub}} \Omega_a} \) be the (conditional) probability to be in the subsample \( \Omega_a^{\text{sub}} \), given that \( a \) is in the sample \( s_A \). The final indirect sample \( s_B^{\text{fin}} \) then consists of the union of all \( \Omega_a^{\text{sub}} \). Because the subsampling is done independently, some units \( b \in U_B \) might appear in different \( \Omega_a^{\text{sub}} \); this has to be considered when constructing an estimator.

The following estimator is unbiased for the total \( t_Y \) in \( U_B \):

\[
\hat{t}_{Y,\text{IS},2\text{phase}} := \sum_{b \in s_B^{\text{fin}}} w_{bs}^t y_{b}, \quad \text{with} \quad w_{bs}^t = \sum_{a \in s_A} \frac{\theta_{ab} \cdot 1(b \in \Omega_a^{\text{sub}})}{\pi_a \cdot \pi_{b|\Omega_a^{\text{sub}} \Omega_a} \cdot \theta_{+b}},
\]
where \(1(b \in \Omega_{a \text{sub}})\) is equal to 1, if \(b \in \Omega_{a \text{sub}}\), and 0 otherwise.

To prove the unbiasedness, we condition on the first stage direct sample \(s_A\) (again denoting expectation over the first and second stage sampling by \(E_1\) and \(E_2\), respectively):

\[
E(\hat{t}_{Y, IS, 2\text{stage}}) = E_1\left(E_2\left(\hat{t}_{Y, IS, 2\text{stage}}|s_A\right)\right)
\]

\[
= E_1\left(E_2\left(\sum_{a \in s_A} \sum_{b \in \Omega_{a \text{sub}}} \frac{\theta_{ab} \cdot 1(b \in \Omega_{a \text{sub}})}{\pi_a \cdot \pi_{b \in \Omega_{a \text{sub}}} \cdot \theta_{+b}} y_b|s_A\right)\right)
\]

\[
= E_1\left(E_2\left(\sum_{a \in s_A} \frac{1}{\pi_a} \cdot \sum_{b \in \Omega_{a \text{sub}}} \frac{\theta_{ab}}{\frac{\pi_{b \in \Omega_{a \text{sub}}} \cdot \theta_{+b}}{y_b}|s_A}\right)ight)
\]

\[
= E_1\left(\sum_{a \in s_A} \frac{1}{\pi_a} \cdot \sum_{b \in \Omega_{a \text{sub}}} \theta_{ab} \cdot \frac{y_b}{\theta_{+b}}\right)
\]

\[
= E(\hat{t}_{Y, IS}) = t_Y.
\]

Finally, we turn again to the situation where the elements of \(U_B\) are actually clusters of individual elements. Suppose that we independently draw subsamples \(s_b\) from every cluster \(b\) of the final indirect sample \(s_{B \text{fin}}\). Let \(\pi_{i|b}\) be the (conditional) inclusion probability of unit \(i\) in cluster \(b\), given that \(b \in s_{B \text{fin}}\). Then \(y_b\) can be estimated by \(\hat{y}_b = \sum_{s_b} y_{i|b} \pi_{i|b}\). We call the complete procedure three stage (as opposed to three phase) indirect estimation, since subsampling from any \(\Omega_a\) does not depend on the remainder of the first stage sample \(s_A\), which would be typical for multi-phase sampling.

Under this three stage procedure, an unbiased estimator for \(t_Y\) may be defined as follows:

\[
\hat{t}_{Y, IS, 3\text{stage}} := \sum_{b \in s_{B \text{fin}}} \sum_{i \in s_b} w_{bi} y_{bi} \quad \text{with} \quad w_{bi} = \sum_{a \in s_A} \frac{\theta_{ab} \cdot 1(b \in \Omega_{a \text{sub}})}{\pi_a \cdot \pi_{b \in \Omega_{a \text{sub}}} \cdot \pi_{i|b} \cdot \theta_{+b}},
\]

We prove the unbiasedness of \(\hat{t}_{Y, IS, 3\text{stage}}\) using iterated expectations again, first conditioning on the final indirect sample \(s_{B \text{fin}}\) (denoting expectation over the first two sampling stages and the third stage by \(E_{12}\) and \(E_3\), respectively):

\[
E(\hat{t}_{Y, IS, 3\text{stage}}) = E_{12}\left(E_3\left(\hat{t}_{Y, IS, 3\text{stage}}|s_{B \text{fin}}\right)\right)
\]

\[
= E_{12}\left(E_3\left(\sum_{a \in s_A} \sum_{b \in \Omega_{a \text{sub}}} \sum_{i \in s_b} \frac{\theta_{ab} \cdot 1(b \in \Omega_{a \text{sub}})}{\pi_a \cdot \pi_{b \in \Omega_{a \text{sub}}} \cdot \pi_{i|b} \cdot \theta_{+b}} y_{bi}|s_{B \text{fin}}\right)\right)
\]

\[
= E_{12}\left(\sum_{a \in s_A} \sum_{b \in s_{B \text{fin}}} \theta_{ab} \cdot 1(b \in \Omega_{a \text{sub}}) \cdot E_3\left(\sum_{i \in s_b} \frac{y_{bi}}{\pi_{i|b} \cdot \theta_{+b}}|s_{B \text{fin}}\right)\right)
\]

\[
= E_{12}\left(\sum_{a \in s_A} \sum_{b \in s_{B \text{fin}}} \theta_{ab} \cdot 1(b \in \Omega_{a \text{sub}}) \cdot \frac{y_b}{\pi_a \cdot \pi_{b \in \Omega_{a \text{sub}}} \cdot \theta_{+b}}\right)
\]

\[
= E(\hat{t}_{Y, IS, 2\text{stage}}) = t_Y.
\]
4. Application: Sampling of Kindergarten Children for NEPS

The German National Educational Panel Study (NEPS) is a new educational survey with a complex design. Six different samples representing different age cohorts of the population in Germany are drawn and then followed over time. In figure 1, the evolution of the different samples over time is shown.

As shown in figure 1, a sample of children aged 4 that attend a kindergarten is drawn in 2010. In 2011, the sampled children will be aged 5 and still attend kindergarten. In 2012, they will be aged 6 and (at least most of them) attend 1st grade in primary school. See Blossfeld et al. (2009) or [www.uni-bamberg.de/en/neps/](http://www.uni-bamberg.de/en/neps/) for more information on the survey.

Children in Germany are not obliged to go to a kindergarten, but roughly 95% of all children aged 4 attend some kind of kindergarten or pre-school. Unlike in some other countries, kindergartens in Germany are completely separated from primary schools. Unfortunately, there is no complete listing of kindergartens in Germany available for sample selection. On the other hand, a complete sampling frame for primary schools is available. Also, despite the spatial separation, kindergartens might be seen as “linked” to primary schools, since every child that eventually leaves a kindergarten joins a particular primary school. Thus, indirect sampling is possible in this application.
Using the notation from the previous sections, let \( U_A \) be the population of primary schools, and let \( U_B \) be the population of kindergartens. There are two rather obvious ways to define a link function on \( U_A \times U_B \):

(i) \( \theta_{ab} = 1 \) if there was at least one child moving from kindergarten \( b \) to school \( a \) in a particular reference period; otherwise \( \theta_{ab} = 0 \);

(ii) \( \theta_{ab} \) = number of children having moved from kindergarten \( b \) to school \( a \) in a particular reference period.

Both definitions result in unbiased estimators. In the general context of indirect sampling, the decision which definition of link function to choose depends on the variance of the respective estimators (which in practical applications have to be estimated by a simulation study) or practical considerations concerning how easy it is to get the values \( \theta_{ab} \) and \( \theta_{+b} \) from the sampled units.

In our application, the first step is to draw a sample of primary schools \( s_A \). Then, every school \( a \in s_A \) is asked to provide \( \theta_{ab} \) for every \( b \in \Omega_a \). In case of link function (i) this means providing the set of kindergartens that sent children to school \( a \) in some reference period (e.g. last year). In case of link function (ii) this means providing for every child that joined school \( a \) as a first grader in the reference period the name of the kindergarten the child was sent from. Pre-tests for the survey have shown that primary schools are usually able to come up with both kinds of information from their files.

Since the number of kindergartens that a primary school is linked to can be quite different, for budget reasons a decision was made not to survey the complete indirect sample \( s_B \) but to use some kind of subsampling. \( s_{fin}^B \) will be drawn by the procedure described in section [3.2]. The reason for this is that the complete direct sample of primary schools \( s_A \) will be used to get a sample of first graders in 2012, and it is desired to then find in every school \( a \in s_A \) at least some children that were in the kindergarten sample of 2010.

In every kindergarten \( b \in s_{fin}^B \), we then ask for the value of \( \theta_{+b} \). In case of link function (i) this means providing the number of schools that all those children joined who left kindergarten \( b \) during the reference period. In case of link function (ii) this means providing the number of children that left kindergarten \( b \) during the reference period and joined some primary school. Pre-tests have shown that the latter information can be given by the kindergartens much more reliably. They know quite well how many children left, but they usually do not know exactly which primary schools (or how many of them) these children joined. For this reason, link function (ii) will be used for the sampling in NEPS.

Finally, again for budget reasons and because the sizes of kindergartens (in terms of number of children aged 4) vary considerably, in every kindergarten \( b \in s_{fin}^B \) a subsample \( s_b \) of children is drawn. Thus, the sample of kindergarten children in NEPS will be drawn following the three stage sampling procedure described in section [3.2].

5. Conclusion

Indirect sampling proved to be a feasible way to draw a sample of kindergarten children for the German National Educational Panel Study (NEPS). For budget reasons, three stage indirect sampling will be used; we have shown that this procedure allows unbiased estimation of population totals. Since NEPS is a voluntary survey, nonresponse will inevitably occur on every stage of the sampling procedure, i.e. among the primary schools in the first stage direct sample, among the kindergartens in the final indirect sample, and among the sampled children. Methods for dealing with nonresponse in the context of indirect surveys are described in Lavallée (2007) and Xu and Lavallée (2009) and will be used for nonresponse adjustment of NEPS.
REFERENCES


