

# Combining an ongoing panel with a new cross-sectional sample

Martin Spiess<sup>a</sup> and Ulrich Rendtel<sup>b</sup>

**Abstract.** In this paper, a weight is derived for the calculation of design based estimators of totals, means and proportions using the ongoing socio-economic panel (SOEP) and a new cross-sectional sample. In the first part of the paper, the selection schemes of the subsamples A, B, C and D of the ongoing panel and of the new sample (sample E) are described. Using some similarity properties and starting from a general formulation, an optimal weight in the sense of small variances of design-based estimators using both samples is derived. The merits of this approach as well as some disadvantages are discussed.

*Key words:* Design based inference; convex weighting estimator; complex surveys; panel survey; cross-sectional sample

## 1 Introduction

Although there exist suggestions on how to combine two (or more) samples to estimate certain parameters in finite populations (an early reference is Patterson, 1950), it is by no means straightforward to combine an ongoing panel and a new sample selected at distant points in time for efficient estimation, even if both samples are independent. Generally, one possible solution is to use the probabilities of an element to be selected for the first and second sample when calculating so-called  $\pi$ -estimators. Given both probabilities and the independence of the drawings, one can calculate the probability that a certain element is selected at least once and then use the inverse probability as the weight for that element.

---

<sup>a</sup>German Socio-Economic Panel Study, German Institute for Economic Research, Koenigin-Luise-Str. 5, 14195 Berlin, Germany

<sup>b</sup>Institute for Statistics and Mathematics, University of Frankfurt am Main, Germany

However, if an ongoing panel is to be combined with a new cross-sectional sample, the selection probabilities at time  $t$ , at which the new cross-sectional sample is selected, are needed for all elements of both samples. Unfortunately, the probabilities to be selected into the ongoing panel at time  $t$  of those elements observed in the cross-sectional sample only are unknown. Therefore, the probability of being selected in at least one sample cannot be derived for all sample elements.

Another approach is to follow the basic idea behind the construction of convex weighting estimators (e.g. Rendtel, 1999). According to this approach, an estimator can be constructed combining  $\pi$ -estimators calculated for both samples via convex weights, i.e. weights that sum up to unity. In the present paper, these weights are not estimated but are deduced using a model about an assumed selection process of the ongoing panel in 1998 and some assumptions concerning that selection. It follows that in deriving the (estimators of the) variances of the estimators of totals, means and proportions, the weights can be considered as constant, i.e. variance formulas are not further complicated as would be the case if the weights were estimators themselves. If, as a consequence of (partly) wrong assumptions, the proposed values of the weights are false, the corresponding design-based estimators are still unbiased, although their variances are larger.

This paper is organized as follows. In Section 2, the selection schemes of the subsamples of the ongoing panel as well as the selection scheme of the new sample are described. Section 3 gives the more theoretic part of the derivation of the weight, whereas Section 4 describes the part necessary to determine the value of the weight to be used with the German socio-economic panel and the new sample. A discussion can be found in Section 5.

## 2 Samples and selection schemes

The ongoing German Socio-Economic Panel (GSOEP) consists of several subsamples, selected from different subpopulations considered to be disjunct and starting at different points in time (see Pannenberg et al. 1998, or, Wagner et al., 1994).

The population from which subsample A was selected was defined to be the set of private households where the household head did not have the Turkish, Italian, Greek, (former) Yugoslavian or Spanish nationality. Subsample A was

selected in 1983/1984. The sampling design has two stages and two phases within the first stage. In the first phase of the first stage, the primary sampling units (units smaller than constituencies, i.e. ‘Stimmbezirke’; PSUs) were selected and in the second stage the secondary sampling units (households; SSUs) were selected. The scheme used to select the first-phase-first-stage sample of PSUs may be described as a systematic probability proportional-to-size without replacement scheme (systematic  $\pi$ ps-scheme, see e.g. Särndal et al. 1993, p. 96). However, since the sizes of the PSUs — given by the number of households belonging to the defined population — were unknown, they had to be estimated. This sample was then stratified according to several variables very similar to those variables used to sort the population elements (PSUs) for the first-phase-first-stage sample so as to mimic certain marginal distributions. Within each cell, again samples of PSUs according to a systematic  $\pi$ ps-scheme were selected. Given this second-phase-first-stage sample of PSUs, within each PSU the SSUs (households) were selected according to a scheme that may approximately be described by a circular systematic sampling scheme with random start (see, e.g., Särndal, 1993, p. 73 ff.). The number of private households in subsample A successfully interviewed in 1998, i.e. the number of observed and valid private households in subsample A in 1998, was  $n_{A,H} = 3345$ , covering a total of  $n_{A,P} = 6138$  successfully interviewed persons aged 16 and older. The number of children, i.e. persons aged 15 and younger, within these households in 1998 was  $n_{A,C} = 1609$ .

The population from which subsample B was selected in 1983/1984 was defined to be the set of private households where the household head had the Turkish, Italian, Greek, (former) Yugoslavian or Spanish nationality. In fact, subsample B consists of five samples selected from the above five disjunct sub-populations. Each of the five subsamples was selected in two stages, where the first-stage samples were selected according to a systematic  $\pi$ ps-scheme. The PSUs selected at the first stage were counties and metropolitan areas. The sizes of the PSUs were number of residents with the corresponding nationality. Given the first-stage samples of PSUs, within each PSU, addresses of persons aged 16 and older with a given nationality were selected according to a systematic sampling scheme with random start. The household selected in this manner was defined to be a sample element if the nationality of the household head was the same as the nationality of the selected person. The number of observed and valid

private households in subsample B in 1998 was  $n_{B,H} = 888$ , covering a total of  $n_{B,P} = 1949$  successfully interviewed persons aged 16 and older. The number of children within these households in 1998 was  $n_{B,C} = 659$ .

Subsample C, selected in 1990, was a sample of private households in the former East Germany. The selection followed a ‘two stage and two phases within the first stage’ design, similar to the selection scheme used for subsample A. In the first phase of the first stage, communities (PSUs) were selected according to a systematic  $\pi$ ps scheme with the sizes of the PSUs being the number of residents. The PSUs were then again stratified according to the variables used to sort the population elements so as to mimic certain marginal distributions. Within each cell, again samples of PSUs according to a systematic  $\pi$ ps-scheme were selected. Given this second-phase-first-stage sample of PSUs, within each PSU, the SSUs (households) were selected according to a scheme that may approximately be described by a circular systematic sampling scheme with random start. The number of observed and valid private households in subsample C in 1998 was  $n_{C,H} = 1867$ , covering a total of  $n_{C,P} = 3707$  successfully interviewed persons aged 16 and older. The number of children within these households in 1998 was  $n_{C,C} = 933$ .

Since the subject of the present paper is the construction of a weight to be used for the estimation of population parameters using subsamples A, B, C, D and the new sample, only those elements of D with a positive non-D-specific weight are considered ( $LHHRF > 0$ ,  $LPHRF > 0$ ; for details see Rendtel et al. 1997). The corresponding population can be defined as the set of private households with occupants who came to the former West Germany since 1984 but were not elements of the populations from which the samples A, B and C were selected. In fact, the part of D considered in this paper consists of two samples, selected in 1992/1994 ( $D_1$ ) and 1994/1995 ( $D_2$ ), respectively. As a result of several difficulties in selecting such a sample (for details, the reader is referred to Rendtel et al. 1997, or, Schulz et al., 1993), different selection schemes were used to select subsample D. In fact, one portion of sample D,  $D_1$  selected in 1992 and 1994, consists of two subsamples,  $D_{11}$  and  $D_{12}$ , say, each selected according to a different selection scheme. The selection scheme of the other part of D,  $D_2$  selected in 1994 and 1995, again differs from the selection schemes used to select the two subsamples  $D_{11}$  and  $D_{12}$ . However, the selection schemes of  $D_{11}$  and  $D_2$

are similar in that the selection of the first-stage sample is based on a systematic  $\pi$ ps-scheme. For both subsamples, the second-stage sample can approximately be described by a systematic sampling scheme with random start, where the valid sample elements are selected with a certain but unknown probability. Although this selection scheme has elements equal to the selection scheme used e.g. for the selection of subsample A, there are also some differences. For example, selected households in 1992, as part of the  $D_{11}$  second-stage sample, were asked whether they agreed with the storage of their addresses for future surveys. These addresses then were used for selecting sample  $D_{11}$ . Given addresses selected in the same way in 1994, quota sampling elements were used to select sample  $D_2$ . The other part of  $D_1$ ,  $D_{12}$ , was selected using telephone survey techniques, where phone numbers were randomly chosen in view of regional criteria ('InfraScope' system, see Infratest Sozialforschung, 1994, or Rendtel et al. 1997). As for samples  $D_{11}$  and  $D_2$ , the selected households were then asked whether or not they agreed with the storage of their addresses for future surveys. Those who agreed were then selected in 1994 into subsample  $D_{12}$ . The number of observed and valid private households in subsample D in 1998 was  $n_{D,H} = 255$ , covering a total of  $n_{D,P} = 528$  successfully interviewed persons aged 16 and older. The number of children within these households in 1998 was  $n_{D,C} = 235$ .

In 1998, a new sample was selected from the population of households given by the union of the (disjunct) subpopulations described above. The new sample, also denoted as subsample E, was selected independently from the ongoing panel (subsamples A through D). The selection scheme used for sample E essentially resembles the scheme also used in selecting subsample A. Again, the data are collected in two stages and two phases within the first stage, where the first- and second-phase samples are selected using the scheme also used for selecting subsample A. Although there are slight differences in the selection of the second-stage sample, mainly due to testing a new survey instrument (using a laptop for the personal interviews vs. paper-and-pencil personal interviews), the selection scheme is very similar to the one used to select the second-stage sample of subsample A. The number of observed and valid private households in subsample E in 1998 was  $n_{E,H} = 1066$ , covering a total of  $n_{E,P} = 1929$  successfully interviewed persons aged 16 and older. The number of children within these households in 1998 was  $n_{E,C} = 468$ .

For the rest of this paper, both the ongoing panel (subsamples A through D) and the new sample will be denoted as subsamples.

### 3 Integration of the two samples: Theory

By ‘integration of an ongoing panel and a new sample’, we mean the use of both subsamples to calculate design-based estimators of totals, means and proportions. To be more precise, we consider linear estimators of the form

$$\hat{\theta} = w \hat{\theta}_1 + (1 - w) \hat{\theta}_2, \quad 0 \leq w \leq 1$$

where  $\hat{\theta}_1$  is an estimator of  $\theta$ , the population parameter, using the ongoing panel in 1998,  $\hat{\theta}_2$  is an estimator of the same quantity using the first wave of the new sample, and  $w$  is a weight to be constructed. Both estimators,  $\hat{\theta}_1$  and  $\hat{\theta}_2$ , relate to a certain characteristic, e.g. monthly income or membership of a specific subpopulation. Note that the above estimator is a special case of the class of convex weighting estimators (e.g. Rendtel, 1999).

Clearly, if  $\hat{\theta}_1$  and  $\hat{\theta}_2$  are both unbiased estimators, then for every value  $0 \leq w \leq 1$ , the estimator  $\hat{\theta}$  is unbiased, too. Therefore, an additional criterion is needed to construct a weight  $w$  that is optimal in some sense.

The criterion used in the choice of  $w$  is the minimization of the variance of the estimator  $\hat{\theta}$ . Note, however, that  $w$  chosen as to minimize the variance of  $\hat{\theta}$  will also minimize the coefficient of variation and, if  $\hat{\theta}_1$  and  $\hat{\theta}_2$  are unbiased, the mean squared error.

Since both subsamples are independent draws, we have  $\text{Cov}(\hat{\theta}_1, \hat{\theta}_2) = 0$  and

$$\begin{aligned} \text{Var}(\hat{\theta}) &= w^2 \text{Var}(\hat{\theta}_1) + (1 - w)^2 \text{Var}(\hat{\theta}_2) \\ &= w^2 [\text{Var}(\hat{\theta}_1) + \text{Var}(\hat{\theta}_2)] - 2w \text{Var}(\hat{\theta}_2) + \text{Var}(\hat{\theta}_2) \\ &= w^2 \text{Var}(\hat{\theta}_1 + \hat{\theta}_2) - 2w \text{Var}(\hat{\theta}_2) + \text{Var}(\hat{\theta}_2). \end{aligned}$$

Straightforward calculation shows that given  $\text{Var}(\hat{\theta}_1 + \hat{\theta}_2) > 0$ ,  $\text{Var}(\hat{\theta})$  as a function of  $w$  has a unique minimum at

$$w = \frac{\text{Var}(\hat{\theta}_2)}{\text{Var}(\hat{\theta}_2) + \text{Var}(\hat{\theta}_1)} . \tag{1}$$

This is easily checked, since the first derivative of  $\text{Var}(\hat{\theta})$  with respect to  $w$  is given by  $2w \text{Var}(\hat{\theta}_1 + \hat{\theta}_2) - 2\text{Var}(\hat{\theta}_2)$  and the second derivative with respect to  $w$  is given by  $2\text{Var}(\hat{\theta}_1 + \hat{\theta}_2)$ .

The above result is not very useful for practical purposes, since the ratio of the variances are unknown and without further assumptions,  $w$  may vary freely over the whole range of possible values, depending on the estimator used and the characteristic considered. To emphasize the fact that  $w$  as well as quantities like  $\hat{\theta}$ ,  $\hat{\theta}_1$  and  $\hat{\theta}_2$  depend on the characteristic  $y$ , they will be denoted as  $w_y$ ,  $\hat{\theta}_y$ ,  $\hat{\theta}_{1,y}$  and  $\hat{\theta}_{2,y}$ , respectively, in what follows.

However, the range of admissible values  $w_y$  can be limited by using the known sample sizes and making the following assumptions:

- (A.1) If both subsamples would have been drawn at the same point in time, the design effects for the estimators  $\hat{\theta}_{1,y}$  and  $\hat{\theta}_{2,y}$  would have approximately been equal (for the definition of design effect, see the Appendix).
- (A.2) The design effect for  $\hat{\theta}_{1,y}$ , calculated using the 1998 data of the ongoing panel, is larger than for  $\hat{\theta}_{2,y}$ .

The first assumption can be justified by the similarities of the designs used to draw the subsamples (at least for subsample A through C and sample E; c.f. Infratest Burke Sozialforschung, 1998). The second assumption reflects the fact that the calculation of cross-sectional weights for wave two and later waves of the ongoing panel (for details see Rendtel, 1995) is based on models e.g. to compensate for attrition over time. The estimation of corresponding probabilities leads to a loss of precision of corresponding estimators relative to a strategy using the same design (including the same number of observed elements) and estimators, but without the need to compensate e.g. for panel attrition. Note that according to the follow-up strategies, it is possible that new members enter into the sample after wave one. However, the consequences with respect to the variances of corresponding estimators are ambiguous, and their number is small relative to the number of elements leaving the sample. Therefore, the second assumption still seems to be justified.

Both assumptions are formulated in terms of design effects and not in terms of variances. However, it can easily be shown (see the Appendix) that (1) can be

approximated by a function of design effects and sample sizes, i.e.

$$w_y \approx \frac{\frac{n_1 \text{deff}_{2,y}}{n_2 \text{deff}_{1,y}}}{1 + \frac{n_1 \text{deff}_{2,y}}{n_2 \text{deff}_{1,y}}} \quad (2)$$

(cf. Latouche et al., 1997), where  $n_1$  is the sample size of the ongoing panel,  $n_2$  is the sample size of the new sample,  $\text{deff}_{1,y}$  is the design effect of the estimator  $\hat{\theta}_{1,y}$  using the ongoing panel and  $\text{deff}_{2,y}$  is the design effect of the estimator  $\hat{\theta}_{2,y}$  using the new sample.

Before assumptions (A.1) and (A.2) as well as the approximation (2) can be fully utilized, some kind of model concerning the selection of the samples along with a few additional assumptions and definitions are necessary.

First, consider the ongoing panel, i.e. subsamples A–D. Although each subsample was selected from a different subpopulation, the elements of the ongoing panel are considered to be selected in 1998 from the same population as sample E. According to this model, the 1998 cross-sectional weights for the ongoing panel can be interpreted as the inverse estimated probabilities of the corresponding elements being observed in 1998. Assuming that the ongoing panel is selected in 1998, the scheme used to select this sample is not completely known.

On the other hand, three sequential phases can be identified leading to the sample observed in 1998. The selected elements at the start of the corresponding subsample of the panel including those who refused to respond can be considered as the first phase sample. The observed and valid elements at the start of the corresponding subsample of the panel are considered to be the second phase sample. Note that the population in 1998 is not the same as the union of the subpopulations from which the first waves of subsamples A–D were selected. However, those elements not in the population in 1998 but in the ongoing panel in the years before 1998 cannot be members of the ongoing panel in 1998, either. Due to the follow-up rules, elements that entered the population after the starting wave of the corresponding subsample have a probability exceeding zero of being selected into the ongoing panel in 1998. The effects of time, i.e. from the first wave until 1998, leading to attrition and new elements entering the sample are considered to be the third phase of the selection scheme.

More formally, let  $s$  be a specific sample, regarded as the outcome of a set-valued random variable  $S$ . The probability distribution of  $S$ , denoted by  $p(\cdot)$  or

$p$  for short, is called the sampling design. For simplicity, the term set-valued will be omitted for the rest of the paper. Now, consider the ongoing panel as being selected for the first time in 1998 from the same population as the new sample (sample E) by a sampling design which is only partly known. Furthermore, let  $S_d$  be the random variable ‘Sample selected by the sample design  $p$  in 1998’,  $S_r$  be the random variable ‘Observed sample at the first wave’ and  $S_t$  be the random variable modeling ‘time effects’. The random variables  $S_d$ ,  $S_r$  and  $S_t$  are of the same dimension, and the events  $S_d = s_d$ ,  $S_r = s_r$  and  $S_t = s_t$  are equivalent to the events  $I_d = i_d$ ,  $I_r = i_r$  and  $I_t = i_t$ , where  $I_d$ ,  $I_r$  and  $I_t$  are  $N \times 1$  random variables ( $N$  is the number of population elements) with elements equal to unity if the corresponding population element is selected into the sample of the corresponding phase and zero otherwise. Note that this is in fact not a new or unusual model but merely a reformulation of assumptions underlying the usual sequential procedures to calculate or estimate (longitudinal) weights in that it is assumed that the specific sample observed in 1998 is the outcome of a random variable which in turn is a function of  $S_d$ ,  $S_r$  and  $S_t$ .

Assuming that  $\hat{\theta}_{1,y}$  is unbiased for  $\theta_y$ ,

$$E(\hat{\theta}_{1,y}) = E_{S_d} E_{S_r} E(\hat{\theta}_{1,y} | S_d, S_r) = \theta_y \quad (3)$$

(see, e.g. Kendall and Stuart, 1976, p. 196) where  $E_S$  means taking the expectation over  $S$ . Since there is no variability of  $\hat{\theta}_{1,y}$  given  $S_t$ ,  $E(\hat{\theta}_{1,y} | S_d, S_r)$  is the expectation of  $\hat{\theta}_{1,y}$  over  $S_t$ .

The variance of the estimator  $\hat{\theta}_{1,y}$  for  $\theta_y$  in 1998 can then be written as

$$\begin{aligned} \text{Var}(\hat{\theta}_{1,y}) &= E_{S_d} E_{S_r} \text{Var}(\hat{\theta}_{1,y} | S_d, S_r) + E_{S_d} \text{Var}_{S_r} E(\hat{\theta}_{1,y} | S_d, S_r) + \\ &\quad \text{Var}_{S_d} E_{S_r} E(\hat{\theta}_{1,y} | S_d, S_r), \end{aligned} \quad (4)$$

where the the variance operators are defined corresponding to the expectation operators above. This result follows easily from the repeated application of the well-known result (e.g. Kendall and Stuart, 1976, p. 196) that the variance of a random variable can be expressed as the sum of the expected value of conditional variances and the variance of conditional expectations. To see this, note that given the second phase sample, the variance of the corresponding estimator can be written as  $\text{Var}(\hat{\theta}_{1,y} | S_d, S_r)$ . Given the first phase sample, the variance of the corresponding estimator is given by

$$\text{Var}(\hat{\theta}_{1,y} | S_d) = \text{Var}_{S_r} E(\hat{\theta}_{1,y} | S_d, S_r) + E_{S_r} \text{Var}(\hat{\theta}_{1,y} | S_d, S_r).$$

Since

$$\text{Var}(\hat{\theta}_{1,y}) = \text{Var}_{S_d} \text{E}(\hat{\theta}_{1,y}|S_d) + \text{E}_{S_d} \text{Var}(\hat{\theta}_{1,y}|S_d)$$

and  $\text{E}(\hat{\theta}_{1,y}|S_d) = \text{E}_{S_r} \text{E}(\hat{\theta}_{1,y}|S_d, S_r)$ , result (4) follows.

Since  $\text{E}(\hat{\theta}_{1,y}|S_d, S_r)$  is the conditional expectation of the estimator over the third phase samples given  $S_d$  and  $S_r$ , the portion of variance due to the ‘time effects’ is given by

$$\gamma_y^2 \equiv \text{E}_{S_d} \text{E}_{S_r} \text{Var}(\hat{\theta}_{1,y}|S_d, S_r). \quad (5)$$

The portion of the variance due to ‘extended design stage effects’, i.e. group, design and nonresponse effects at the stage of drawing the sample, is given by

$$\sigma_{1,y}^2 \equiv \text{E}_{S_d} \text{Var}_{S_r} \text{E}(\hat{\theta}_{1,y}|S_d, S_r) + \text{Var}_{S_d} \text{E}_{S_r} \text{E}(\hat{\theta}_{1,y}|S_d, S_r). \quad (6)$$

A similar decomposition can be made for the variance of  $\hat{\theta}_{2,y}$ , however, without the need to model ‘time effects’. That is, the variance of  $\hat{\theta}_{2,y}$  is a function of the ‘extended design effects’ only.

In what follows, define  $\sigma_{2,y}^2 \equiv \text{Var}(\hat{\theta}_{2,y})$  and let  $\text{Var}_{1,SI}(\hat{\theta}_{1,y,SI})$  be the variance of the estimator  $\hat{\theta}_{1,y,SI}$  under the simple random sampling without replacement ( $SI$ ) scheme given the sample size of the ongoing panel, and let  $\hat{\theta}_{1,y,SI}$  be the expression for  $\hat{\theta}_{1,y}$  under this design. Correspondingly, let  $\text{Var}_{2,SI}(\hat{\theta}_{2,y,SI})$  be the variance of the estimator  $\hat{\theta}_{2,y,SI}$  under the  $SI$  scheme given the sample size of subsample E and  $\hat{\theta}_{2,y,SI}$  be the expression for  $\hat{\theta}_{2,y}$  under this design. For the ease of presentation, let  $\text{Var}_{1,SI} \equiv \text{Var}_{1,SI}(\hat{\theta}_{1,y,SI})$  and  $\text{Var}_{2,SI} \equiv \text{Var}_{2,SI}(\hat{\theta}_{2,y,SI})$ .

The design effects can now be written as

$$\text{deff}_{1,y} = \frac{\sigma_{1,y}^2 + \gamma_y^2}{\text{Var}_{1,SI}} \quad (7)$$

and

$$\text{deff}_{2,y} = \frac{\sigma_{2,y}^2}{\text{Var}_{2,SI}}. \quad (8)$$

From the assumptions (A.1), (A.2) and the assumed selection model, it follows that

$$\frac{\sigma_{1,y}^2}{\text{Var}_{1,SI}} \approx \frac{\sigma_{2,y}^2}{\text{Var}_{2,SI}} \quad (9)$$

and  $\text{deff}_{1,y} > \text{deff}_{2,y}$ . Using this result, we get

$$\frac{\text{deff}_{2,y}}{\text{deff}_{1,y}} = 1 - \frac{\gamma_y^2}{\sigma_{1,y}^2 + \gamma_y^2} \quad (10)$$

and

$$w_y \approx \frac{\frac{n_1}{n_2} \left( 1 - \frac{\gamma_y^2}{\sigma_{1,y}^2 + \gamma_y^2} \right)}{1 + \frac{n_1}{n_2} \left( 1 - \frac{\gamma_y^2}{\sigma_{1,y}^2 + \gamma_y^2} \right)}, \quad (11)$$

where  $n_1$  and  $n_2$  are the sizes of the observed parts of the samples, since the unobserved parts are assumed to be correctly accounted for in the models used to compensate e.g. for nonresponse or panel attrition. Note that  $w_y$  attains its approximate maximum at

$$w_{y,\max} \approx \frac{n_1}{n_2 + n_1}. \quad (12)$$

This value is achieved for  $\gamma_y^2 \ll \sigma_{1,y}^2$ . Although the approximation of the optimal weight  $w_y$  still depends on the characteristic and estimator used, the weight of the corresponding term may be restricted to a plausible range, thereby hopefully restricting  $w_y$  to a small range.

## 4 Integration of the two samples: Application to the GSOEP

The number of private households observed 1998 in the ongoing panel is  $n_{H,1} = 6354$  and in the new sample (sample E) is  $n_{H,2} = 1066$ . A further assumption considered to be plausible is that the portion of the variance due to modeling the dynamic effects of the panel over time relative to the sum of this portion and the portion of the variance due to the variance between the subpopulations, the sample design and the nonresponse at the start of the panel is smaller than .5 but larger than .2, i.e.

$$0.2 \leq \frac{\gamma_y^2}{\sigma_{1,y}^2 + \gamma_y^2} \leq 0.5.$$

Together with the known sample sizes, this assumption leads to the approximate range of the optimal values for  $w_{H,y}$ , where  $w_{H,y}$  is a weight to be used for

estimators on the household level,

$$\frac{5.96 \times 0.5}{1 + 5.96 \times 0.5} \leq w_{H,y} \leq \frac{5.96 \times 0.8}{1 + 5.96 \times 0.8}$$

$$0.75 \leq w_{H,y} \leq 0.83.$$

Note that the upper bound of  $w_{H,y}$  for the given sample sizes is 0.856.

The above range is assumed to be the range of optimal values of  $w_{H,y}$  for all estimators of the form given in Section 3 and all characteristics. Since it seems impossible to find plausible assumptions that allow for further reduction of the range of  $w_{H,y}$  and it also seems not plausible that exactly one value of  $w_{H,y}$  is optimal in the sense defined in Section 3 for each and every characteristic and estimator, the above range of admissible values of  $w_{H,y}$  seems to be a narrow range. However, for the integration of both subsamples as defined in Section 3, a single value is needed for  $w_{H,y}$ . The value chosen for the GSOEP is  $w_{H,y} = 0.80$ .

As a consequence of the above arguments, one value for  $w_{H,y}$  chosen from the range of values for  $w_{H,y}$  cannot be expected to be optimal for all characteristics and estimators. On the other hand, it is hoped that for most applications, the value chosen leads to near-optimal properties of the estimators in the sense of small variances. It should be noted that the estimators as defined in Section 3 remain unbiased regardless of the value of  $w_{H,y}$ .

The integration of the two subsamples with respect to households is realized by multiplying the cross-sectional weights (HHRF) of the ongoing panel (sample A—D) by 0.8 and the cross-sectional weights of the new sample (sample E) by 0.2 from 1998 on. If, however, one is interested in the ongoing panel or the new sample only, beginning with 1998, the weights have to be multiplied by  $1/0.8$  and  $1/0.2$ , respectively. On the other hand, using both subsamples if another weight, say  $w_{H,y,\text{new}}$ , seems to be preferable for a specific analysis, then the corresponding weights have to be multiplied by  $w_{H,y,\text{new}}/0.8$  and  $(1 - w_{H,y,\text{new}})/0.2$ , respectively.

The number of adults (16 years and older) living in the private households observed in 1998 in the ongoing panel is  $n_{P,1} = 12322$  and in the new sample is  $n_{P,2} = 1929$ . The number of children (15 years and younger) living in the private households observed in 1998 in the ongoing panel is  $n_{C,1} = 3436$  and in the new sample is  $n_{C,2} = 468$ . By the same arguments used to derive an interval for the weight to be used for analyses on the household level, intervals for weights to be used for estimators on the individual level are derived. Since

the intervals for adults and children are both very close to the interval derived for the weight to be used on the household level, the same weight that is used on the household level is also used for adults and children. The integration of the two subsamples with respect to adults and children is realized by multiplying the cross-sectional weights (PHRF) of the ongoing panel (sample A—D) by 0.8 and the cross-sectional weights (PHRF) of the new sample (sample E) by 0.2 from 1998 on. Of course, the same arguments corresponding to the weight for the household level also apply to the weight for the level of individuals.

## 5 Discussion

The derivation of the weights proposed in Section 3 and 4 rests upon several assumptions. One assumption is that if both subsamples were drawn at the same point in time, the design effects for the two estimators were approximately equal (see assumption (A.1) in Section 3). Although, as noted in Section 3, this seems to be justified by the similarities of the selection schemes, the fraction of the corresponding ‘design’ effects may depart from unity to some extent. In this case, the decomposition of the variance in Section 3 is not valid. However, this does not necessarily mean that the assumption of a ‘residual’ proportion to be lower or equal than 0.5 but greater or equal than 0.2 (Section 4) is wrong, too.

As mentioned in Section 4, it is unplausible that ‘one-value-fits-all’ solutions like the values determined in Section 4 exist, since the optimal weights — in the sense of minimal variances — are functions of a given estimator and characteristic. The same problem, of course, would have to be faced if the values of the convex weights were estimated. Furthermore, in the latter case, deriving the variance of corresponding estimators of totals, means or proportions, the stochastic nature of the estimated weights would have to be taken into account generally leading to larger variances.

As it is not possible to derive the inclusion probabilities for the joint sample, i.e. the ongoing panel and the new sample, the merit of the approach chosen in this paper to derive a single weight for all elements (households or individuals) of one sample is that if an analysis is done using some specific estimator and characteristic and the analyst believes the weight not to be optimal, he or she may easily choose another weight as described in Section 4. This is also true if

only the ongoing panel or only the new sample is used. Also, if new (exogenous) information is available, the weights can easily be adapted if necessary.

# APPENDIX

## Design effect

For details concerning the concepts used in this section, see e.g. Särndal, (1993). Let  $s$  be a specific sample and  $Pr(S = s) = p(s)$  be the probability of selecting  $s$  under a given sample selection scheme. The function  $p(s)$  or  $p$  for short, is called the sampling design. For a complex statistic or estimator,  $\hat{\theta}$ , the design effect for  $\hat{\theta}$  under a specific design  $p$  is

$$\text{deff}_p(\hat{\theta}) = \frac{\text{Var}_p(\hat{\theta})}{\text{Var}_{SI}(\hat{\theta}_{SI})},$$

where  $\text{Var}_{SI}(\hat{\theta}_{SI})$  is the variance of the estimator  $\hat{\theta}_{SI}$  under the simple random sampling without replacement ( $SI$ ) scheme, and  $\hat{\theta}_{SI}$  is the expression for  $\hat{\theta}$  under this design. For a fair comparison, both designs should have the same (expected) sample size. For example, if  $\hat{\theta}_{SI}$  is the  $\pi$ -estimator of the population total of a characteristic  $y$  under the  $SI$  design, then

$$\text{Var}_{SI}(\hat{\theta}_{SI}) = N^2 \frac{1-f}{n} S_{y,U}^2,$$

where  $N$  is the number of elements in the population,  $n$  is the number of elements in the sample,  $f = n/N$  and

$$S_{y,U}^2 = \frac{1}{N-1} \sum_U (y_k - \bar{y}_U)^2.$$

In the last expression,  $\sum_U$  means summation over all population elements  $y_k$ ,  $k = 1, \dots, N$  and  $\bar{y}_U$  is the population mean. If the design effect exceeds unity, then precision is lost by using the design  $p$  and the estimator  $\hat{\theta}$  rather than using the  $SI$  design and the estimator  $\hat{\theta}_{SI}$ . If the design effect is less than unity, precision is gained relative to the use of the  $SI$  design and the  $\hat{\theta}_{SI}$  estimator.

## $w$ as a function of design effects

Let  $n_1$  be the sample size of the ongoing panel,  $n_2$  the sample size of the new sample,  $f_1 = n_1/N$  and  $f_2 = n_2/N$ . Then

$$w_y = \frac{\text{Var}(\hat{\theta}_2)}{\text{Var}(\hat{\theta}_2) + \text{Var}(\hat{\theta}_1)}$$

can be written as

$$w_y = \frac{\frac{\text{Var}(\hat{\theta}_2)}{N^2 S_{y,U}^2}}{\frac{\text{Var}(\hat{\theta}_2)}{N^2 S_{y,U}^2} + \frac{\text{Var}(\hat{\theta}_1)}{N^2 S_{y,U}^2}}.$$

Multiplication by  $((1 - f_1)/n_1)/((1 - f_1)/n_1)$  and  $((1 - f_2)/n_2)/((1 - f_2)/n_2)$ , respectively, leads to

$$\begin{aligned} w_y &= \frac{\frac{1-f_2}{n_2} \frac{\text{Var}(\hat{\theta}_2)}{N^2 \left(\frac{1-f_2}{n_2}\right) S_{y,U}^2}}{\frac{1-f_2}{n_2} \frac{\text{Var}(\hat{\theta}_2)}{N^2 \left(\frac{1-f_2}{n_2}\right) S_{y,U}^2} + \frac{1-f_1}{n_1} \frac{\text{Var}(\hat{\theta}_1)}{N^2 \left(\frac{1-f_1}{n_1}\right) S_{y,U}^2}} \\ &= \frac{\frac{1-f_2}{n_2} \text{deff}_2}{\frac{1-f_2}{n_2} \text{deff}_1 + \frac{1-f_1}{n_1} \text{deff}_1}, \end{aligned}$$

where  $\text{deff}_1 \equiv \text{deff}_{p_1}(\hat{\theta})$  is the design effect for  $\hat{\theta}$  under the design that should have been used to draw the ongoing panel in 1998 if this would have been the start of the ongoing panel ( $p_1$ ), and  $\text{deff}_2 \equiv \text{deff}_{p_2}(\hat{\theta})$  is the design effect for  $\hat{\theta}$  under the design that was used to draw the new sample in 1998 ( $p_2$ ).

Since  $(1 - f)/N = 1/n - 1/N$ , for both designs we have

$$\begin{aligned} w_y &\approx \frac{\frac{1}{n_2} \text{deff}_2}{\frac{1}{n_2} \text{deff}_2 + \frac{1}{n_1} \text{deff}_1} \\ &= \frac{\frac{n_1}{n_2} \frac{\text{deff}_2}{\text{deff}_1}}{1 + \frac{n_1}{n_2} \frac{\text{deff}_2}{\text{deff}_1}}, \end{aligned}$$

where we have ignored the terms  $N^{-1}\text{deff}_1$  and  $N^{-1}\text{deff}_2$ . Note that this result holds not only for  $\pi$ -estimators of totals but also for  $\pi$  estimators of means and proportions.

## References

- Infratest Sozialforschung (1994). *SOEP '94, Zuwanderer-Befragung (Teilstichprobe D 1), Methodenbericht*. München: Infratest Sozialforschung.
- Infratest Burke Sozialforschung (1998). *SOEP '98, Erstbefragung der Stichprobe E, Methodenbericht*. München: Infratest Burke Sozialforschung.
- Kendall, M. & Stuart, A. (1976). *The Advanced Theory of Statistics, Vol. 3, 3rd ed.*. London: Griffin.
- Latouche, M., Michaud, S., Tremblay, J., Merkouris, T. & Renaud, M. (1997). Integration of a Cross-Sectional and a Longitudinal Survey to Produce Income Estimates. Mimeo, Statistics Canada.
- Pannenberg, M., Pischner, R., Rendtel, U. & Wagner, G.G. (1998). Sampling and Weighting. In: J.P. Haisken-De New & J.R. Frick, *Desktop Companion to the German Socio-Economic Panel Study (GSOEP), Version 2.2 (chap. 4)*. Berlin: German Institute for Economic Research.
- Patterson, H.D. (1950). Sampling on successive occasions with partial replacement of units. *Journal of the Royal Statistical Society B*, 12, 241–255.
- Rendtel, U. (1995). *Lebenslagen im Wandel: Panellausfälle und Panelrepräsentativität*. Frankfurt: Campus.
- Rendtel, U. (1999). *The Application of the Convex Weighting Estimator to Household Panel Surveys*. Mimeo, Frankfurt.
- Rendtel, U., Pannenberg, M. & Daschke, S. (1997). Die Gewichtung der Zuwanderer-Stichprobe des Sozio-oekonomischen Panels (SOEP). *Vierteljahreshefte zur Wirtschaftsforschung*, 66, 2, 271–286.
- Särndal, C.-E., Swensson, B. & Wretman, J. (1993). *Model assisted survey sampling*. New York: Springer.
- Schulz, E., Rendtel, U., Schupp, J. & Wagner, G. (1993). *Das Zuwanderer-Problem in Wiederholungsbefragungen am Beispiel des Sozio-oekonomischen*

*Panels (SOEP)*. Deutsches Institut für Wirtschaftsforschung, Diskussionspapier Nr. 71.

Wagner, G., Schupp, J. & Rendtel, U. (1994). Das Sozio-ökonomische Panel – Methoden der Datenproduktion und -aufbereitung im Lngsschnitt. In R. Hauser, N. Ott & G. Wagner (Hrsg.), *Mikroanalytische Grundlagen der Gesellschaftspolitik – Band 2, Erhebungsverfahren, Analysemethoden und Mikrosimulation* (pp. 70–112). Berlin: Akademie Verlag.