Using Analysis of Gini (ANoGi) for detecting whether two sub-samples represent the same universe:

The German Socio-Economic Panel Study (SOEP) experience

by

Joachim R. Frick [a], *Jan Goebel*[a], *Edna Schechtman*[b], *Gert G. Wagner*[a,c] *and Shlomo Yitzhaki*[d,*]

***** Corresponding author
[a] *German SOEP Study, DIW Berlin,* [b] *Ben Gurion University, Beer Sheva, Israel,* [c] *Berlin University of Technology Berlin and IZA Bonn, Germany* [d] *Dept. of Economics, Hebrew University, and Central Bureau of Statistics, Jerusalem, Israel* msruhama@mscc.huji.ac.il

Support from the German Israeli Foundation (GIF-Grant I-656-10.4/2000) is gratefully acknowledged.

Abstract

A particular shortcoming of panel surveys is potential bias arising from selective attrition. Based on data from the German Socio-Economic Panel Study (SOEP) we analyze potential artifacts (level, structure, inequality of income) by comparing results from two independently drawn panel sub-samples, started in 1984 and 2000, respectively. Both sub-samples carried on using the same set of follow-up rules. We apply ANOGI (ANalysis Of GIni) techniques, the equivalent of ANOVA (ANalysis Of VAriance) performed on the basis of the Gini coefficient. The decomposition followed is presented in Yitzhaki (1994). We rearrange, reinterpret and use the decomposition in the comparison of sub-populations from which the different sub-samples were drawn. Taking into account indicators for income, and for control purposes those for education and satisfaction as well, significant differences between these two sub-samples with respect to (income) inequality are found in the first year, which start to fade away in wave 2 and disappear in wave 3. We find credible indication for these differences to be driven by changes in response behavior of short term panel members rather than by attrition among members of the longer running sub-sample. Concluding from our empirical results and a discussion of survey methodology issues employed in the set-up of the considered sub-samples, we cannot reject the hypothesis that both represent the same universe.

Key Words: Panel Studies, Survey Research, Inequality Decomposition, Gini
JEL codes: C81, D31, D63

1 Introduction

Most population surveys assert the claim that they are representative of the underlying population universe. While this is, in reality, already a very ambitious goal on its own, *panel* surveys which follow households, families and individuals over time, have to cope with the problem of adequately covering any *changes* occurring in the underlying population since the time of the original sampling. For example, immigration may not be represented within an ongoing panel survey if recent immigrants found new households, which by definition did not have a positive sampling probability. Thus, new sub-samples which complement a panel might be required. Furthermore, due to panel attrition and the need to control for eventual selectivity within this process, long-running panel surveys especially may require to be complemented by additional sub-samples which serve two functions: firstly, such "refreshment" samples help to stabilize the number of observations and secondly, they provide a benchmark for the analysis of eventual selectivity due to panel attrition and changes in response behavior.

When analyzing economic well-being or income distribution issues on the basis of such micro-data, any undetected selectivity, for example given by a "middle income bias", creates a bias in estimates of income inequality measures.

The question to be answered in this paper can be presented in the following general way: given the existence of several independent samples (within one survey), do they represent the same population or universe? A common practice to answer this question is to look at the differences in various parameters among the populations with respect to the variable of interest. For example, one may wish to compare the moments of the distributions (means, variances, Ginis, medians, etc). The problem with a methodology such as this, is that there are only a few moments which are usually being tested and therefore, possible differences in other moments may not be detected. The methodology suggested in this paper is based on a decomposition of a measure of total variability to the contributions of sub-populations. Since the interest is in the inequality in economic well-being (as measured, for example, by income), the comparison between the sub-populations is done by decomposing the Gini of income of the overall population into the contributions of the sub-populations to the overall inequality. The

advantage of the decomposition is that it reveals a new parameter (called the overlapping index), which shows how intertwined the sub-populations are. Hence, unlike the comparison of moments that rely on each distribution separately, the overlapping index is based on the entwined observations of all distributions involved.

Intuitively, the methodology presented below can be referred to as ANOGI (ANalysis Of GIni) – the equivalent of ANOVA (ANalysis Of VAriance), performed with the Gini coefficient. The decomposition we follow is the one presented in Yitzhaki (1994). We rearrange and reinterpret that decomposition in order to use it in the comparison of the sub-populations from which the different samples were drawn (section 3).

In this paper we are *mainly* interested in the effect (possible bias) of attrition. For each year within a three year period (2000 to 2002) we have two sub-samples. Both of them are sub-samples from a chain of panel data, intended to represent the entire population of Germany. The major difference in the sub-samples is that the new one was started in the year 2000, while the main part of the old one was started as early as 1984 (see Table 2.1). For the period since the original sampling took place, both panel sub-samples carried on using the same set of straightforward follow-up rules. We are interested in seeing whether the sub-samples come from the same population (i.e. no effect of long-term attrition), or if attrition causes a bias.

As a central result, our analyses for the first year reveal a significant difference in income stratification and inequality between the sub-samples, while in the second and third years the two sub-samples overlap almost perfectly; i.e. the relevant substantive results converge rather quickly. When discussing reasons for the differences in the first year, we find indications that these are not to be attributed to attrition in the "old" sample, but rather to changes in the response behavior of the new sub-sample's members. In fact, there is some evidence that the answers of long-term panel respondents are of better quality than those of first-time respondents.

We start the paper with a detailed description of the underlying data and with a discussion of problems related to the representation of a population by means of different sub-samples (Section 2). In Section 3 we set up the ANOGI methodology, and Section 4 provides the estimators. In Section 5 we present results of inequality decomposition and Section 6 concludes.

2 Representation of a Population by Means of Different Samples

2.1 The German Socio Economic Panel Study (SOEP)

Established in 1984, the German Socio-Economic Panel Study (SOEP) is one of the main tools for social science and economic research for Germany, as well as for international comparisons (cf. Wagner et al. 1993, Haisken-DeNew and Frick 2003).[1]

In principle, the universe of the SOEP sample includes the entire resident population of Germany. As the SOEP started before the reunification of Germany occurred, the first sub-samples of the SOEP in 1984 were only conducted in West Germany. We summarize the various sub-samples in Table 2.1. A detailed description can be found in Appendix A.

Table 2.1. Description of the SOEP sub-samples.

Sample	Starting Year	Sample size[2] (no. of households in starting year)	Comments
A	1984	4,528	"West-German" sample
B	1984	1,393	Over-sampling of foreigners
C	1990	2,179	"East German" sample
D	1995	522	Immigrants since 1984
E	1998	1,067	Supplementary sample
F	2000	6,052	Supplementary Innovation sample
Cross-section 2000		13,258 (6,052)	All samples A-F (thereof F)
Cross-section 2001		11,947 (4,911)	All samples A-F (thereof F)
Cross-section 2002		11,468 (4,586)	All samples A-F (thereof F)

[1] The SOEP data are made available in user-friendly form ("scientific use file") to all independent researchers in the Federal Republic of Germany upon request and, in addition, worldwide to all independent research institutions. Analysis of the data is supported by an extensive internet and online service. Today the SOEP is also widely used by international organizations such as the OECD, especially for the analysis of the income distribution.

[2] Due to attrition on the one hand and the follow-up of newly founded households in case of split-offs on the other hand, the cross-sectional sample size in any year after the starting wave deviates from the initial sample size.

In 2000, the starting year of the supplementary innovation sub-sample F, all in all 24,586 adult individuals participated in the SOEP survey, which covered 13,258 households and included 6,659 children under 16 years of age.

2.2 Potential Artifacts Caused by Different Sub-samples

The realization of different independent samples which have the aim of representing the same population universe can cause problems due to different, mostly fieldwork related, reasons. Firstly, it is possible that they belong to different universes because the sample frames were not by intention, but in fact different with respect to specific sampling procedures. Secondly, initial response rates may differ across the various sub-samples. Thirdly, methodological problems occurring during the fieldwork can result in a couple of survey artifacts. All such specific problems could have occurred with the SOEP, especially because in the year 2000 the "old" sub-samples A through E were true panel sub-samples with more than one wave and in fact different with respect to the numbers of waves.

a) Different Sampling Procedures

There are two structural problems of sampling households in Germany, as in many other countries: firstly, the representation of foreigners/immigrants and secondly, households living in institutions. In the SOEP, the procedures of handling those sub-populations were changed over time: the sampling procedures of sub-samples A & B and F were slightly different for the first waves. Whereas in A and B Germans and foreigners are surveyed by different methods which allow a theoretically proper representation of the five immigrant groups which are represented by sub-sample B, this superior procedure was not possible for sub-sample F. Sub-samples C and D are samples with a special focus on East Germans and recent immigrants respectively, while E is a small sample which was drawn by basically the same procedure as sub-sample F. A more detailed description of these procedures is provided in Appendix B.

b) Response Rates

Given the massive confrontation with telephone surveys, ad-hoc interviews by marketing companies, etc., it is becoming increasingly problematic to motivate individuals to participate in population surveys. As such, the older sub-samples in SOEP clearly show higher initial response rates (e.g. sub-samples A and B with 61% and 68%, respectively) than newly introduced sub-samples such as sub-sample F with only 52% (see Table 2.2 below and Appendix C for more detailed information). This phenomenon also applies to longitudinal response rates over (two or) three waves; it does not matter whether one looks at the sub-sample specific first two or three waves or at the exact same time period (i.e. calendar years), e.g. 2000 through 2002.

c) Surveying Artifacts

While in the year 2000 sub-sample F is a "fresh" cross-section, sub-samples A to E consist of panel samples of varying duration. Therefore, sub-samples A to E and F could represent different populations, firstly, if it is not possible to correct for panel attrition and secondly, if response behavior changes over time ("panel effects"). SOEP data providers control for sub-samples A to E in an appropriate manner by means of weighting (cf. Rendtel 1995, Rendtel et al. 1995), but over the course of time, other panel effects certainly cannot be ruled out. First, conventional wisdom dictates that respondents can change their true behavior, which in turn produces some sort of bias in the result.

A second important effect may be provided by an increasing familiarization of the respondents with the survey instrument, mostly a questionnaire, which minimizes errors in the answers and improves quality of the collected information. As such, due to this "learning effect", panel data yields a more realistic measure than data from a single cross-sectional sample of the very same population. The second effect especially takes place in the field of income surveying. We know that in the course of time the share of missing values on income variables (due to item non-response) declines, and that there is a special reason for different "answering styles".

A third problem can be created by a mix of interview modes, which is necessary, at least in Germany, to ensure that respondents remain willing to

participate over the long term[3]. Since 1998, "Computer Assisted Personal Interviewing" (CAPI) has been introduced gradually to complement the conventional "paper and pencil" questioning technique (which can be realized by interviewer-administered or self-administered interviews). Details are given in Appendix D.

Table 2.2: Cross-sectional and longitudinal response rates in SOEP by sub-sample

Initial (Cross-Sectional) Response Rate in Wave 1	
Sample A (1984)	61%
Sample B (1984)	68%
Sample C (1990)	70%
Sample D (1994/95)	>55%
Sample E (1998)	54%
Sample F (2000)	52%
Longitudinal Response Rate (balanced panel as a percent of starting wave's population)	
over 2 years (calendar years 2000-2001 and 2001-2002)	
Samples A-E	89-93%
Sample F	78-87%
over 3 years (calendar years 2000-2002)	
Samples A-E	80-86%
Sample F	<70%
over sample-specific first two waves (waves 1-2)	
Samples A-E	81-92%
Sample F	78%
over sample-specific first three waves (waves 1-3)	
Samples A-E	72-88%
Sample F	69%

Source: Authors' calculation from SOEP 2000-2002.

[3] See Schräpler/Wagner (2001) for details.

3 ANOGI: The methodology

Let y_i, $F_i(y)$, $f_i(y)$, μ_i, p_i represent the income, cumulative distribution, the density function, the expected value, and the share of sub-population *i* in the overall population, respectively. [4] Let $s_i = p_i\mu_i/\mu_u$ denote the share of group *i* in the overall income. The overall population is composed of the union of the sub-populations. That is: $Y_u = Y_1 \bigcup Y_2 \bigcup \ldots \bigcup Y_n$, where subscript *u* denotes the union of the populations from which all the sub-samples are drawn.[5]

Note that

$$F_u(y) = \sum_i p_i F_i(y) \qquad (1)$$

That is, the cumulative distribution (or ranks) of the overall population is the weighted average of the cumulative distributions of the sub-populations, weighted by the relative sizes of the populations.[6] The formula of the Gini used in this paper is (Lerman and Yitzhaki (1989)):

$$G = \frac{2\,\mathrm{cov}(y, F(y))}{\mu}\ , \qquad (2)$$

which is twice the covariance between the income y and the rank F(y) standardized by mean income μ.[7] The Gini of the entire population, G_u, can be decomposed as:

$$G_u = \sum_{i=1}^{n} s_i G_i O_i + G_b\ , \qquad (3)$$

where O_i is the overlapping index of subpopulation *i* with the entire population (explained below), and G_b is between group inequality. Equation (3) decomposes the Gini of the union into two related components: intra and inter-group components,

---

[4] In the sample, the cumulative distribution is estimated by the rank of the observation, normalized to be between zero and one.

[5] Note that Yu represents the entire population only if there is no attrition. Otherwise, it represents a biased population of the 'entire population', with the bias being a function of the patterns of the attritions.

[6] Actually, p_i is not a true population parameter. It is added here for generality and for handling samples of different sizes.

[7] Note that the relative version of Gini is used, which is as if one uses the coefficient of variation to perform ANOVA. Relative measure is chosen since it is the common parameter used in the income distribution literature.

connected in a way, which is relatively complicated. Note that while in ANOVA, the decomposition of the total variability is partitioned into inter and intra variances, in ANOGI we have inter and intra Gini's, but, in addition, there is an extra parameter, which is the overlapping index. We will return to this implication following the explanation of the individual components.

3.1 The overlapping parameter and its properties[8]

The parameter of overlapping is the one that distinguishes the decomposition of the Gini (ANOGI) from the decomposition of the variance (ANOVA), and is the main reason for expressing a preference for the methodology suggested in this paper over ANOVA.

Overlapping should be interpreted as the inverse of stratification. Stratification is a concept used by sociologists. We follow Lasswell's (1965, p.10) definition as: "In its general meaning, a stratum is a horizontal layer, usually thought of as between, above or below other such layers or strata. Stratification is the process of forming observable layers, or the state of being comprised of layers. Social stratification suggests a model in which the mass of society is constructed of layer upon layer of congealed population qualities."

According to Lasswell, perfect stratification occurs when the observations of each population (in our case sub-sample of the SOEP) are confined to a specific range of income, and the ranges of populations do not overlap. Stratification plays an important role in the theory of relative deprivation (Runciman, 1966), which argues that stratified societies can tolerate greater inequalities than non-stratified ones (Yitzhaki, 1982). In our case, this property plays an important role, because it tells us whether the different sub-samples represent different strata.

One can rarely find a perfect stratification, and an index describing the degree of stratification is called for. The index of overlapping is actually an index describing the extent to which the different populations are stratified. In this paper the goal is to find overlapping, i.e. non-stratification in the sense that two income distributions based on two independent panel sub-samples represent the same universe.[9]

---

[8] The proofs of all statements in this section are given in Yitzhaki (1994).

[9] An alternative use is to search for stratification. For example, Heller and Yitzhaki (2003) argue that a

Formally, overlapping of the overall population by sub-population *i* is defined as:

$$O_i = O_{ui} = \frac{\mathrm{cov}_i(y, F_u(y))}{\mathrm{cov}_i(y, F_i(y))} \ , \qquad (4)$$

where, for convenience, the index u is omitted and cov_i means that the covariance is according to distribution *i*, i.e.

$$\mathrm{cov}_i(y, F_u(y)) = \int (y - \mu_i)\,(F_u(y) - \overline{F}_{ui})\, f_i(y)dy \ , \qquad (5)$$

where $\overline{F}_{ui}$ is the expected rank of population *i* in the union (all observations of population *i* are assigned their union's rank and $\overline{F}_{ui}$ represents the expected value).[10],[11] The overlapping (4) can be further decomposed to identify the overlapping of subpopulation i with all other subpopulations that comprise the union. In other words, total overlapping of subpopulation *i*, O_i , is composed of overlapping of *i* with all sub-populations, including group *i* itself. This further decomposition of O_i is:

$$O_i = \sum_j p_j O_{ji} = p_i O_{ii} + \sum_{j \neq i} p_j O_{ji} = p_i + \sum_{j \neq i} p_j O_{ji} \qquad (6)$$

where $O_{ji} = \frac{\mathrm{cov}_i(y, F_j(y))}{\mathrm{cov}_i(y, F_i(y))}$ is the overlapping of group *j* by group *i*.

The properties of the overlapping index O_{ji} are the following:

(a) $O_{ji} \geq 0$. The index is equal to zero if no member of the *j* distribution lies in the range of distribution *i*. (i.e., group *i* is a perfect stratum).

(b) O_{ji} is an increasing function of the fraction of population j that is located in the range of population *i*.

---

perfect classification into groups is achieved if members of each group are similar among themselves (low intra-group variability), and different from others (stratified). This property of the decomposition of the Gini enables them to use overlapping as an indicator of the quality of classification of snails into groups, according to different observed variables.

[10] Ranking observations according to a different distribution is a rare concept in statistics. However, it is common in sports where each athlete is frequently ranked in his country and according to other scales (world, continent, gender, age group etc..).

[11] It is worth noting that the O_i is a kind of a Gini correlation. See Schechtman and Yitzhaki (1987,1999) for the properties of Gini correlations.

(c) For a given fraction of distribution j that is in the range of distribution *i*, the closer the observations belonging to *j* to the expected value of distribution *i,* the higher O_{ji}.

(d) If the distribution of group *j* is identical to the distribution of group *i*, then $O_{ji}=1$. Note that by definition $O_{ii}=1$. This result explains the second equality in (6). Using (6), it is easy to see that $O_i \geq p_i$ is a result to be borne in mind when comparing different overlapping indices of groups with different sizes.

(e) $O_{ji} \leq 2$. That is, O_{ji} is bounded from above by 2. This maximum value will be reached if all observations belonging to distribution *j* that are located in the range of *i,* are concentrated at the mean of distribution *i*. Note, however, that if distribution *i* is given then it may be that the upper limit is lower than 2 (see Schechtman, 2000). That is, if we confine distribution *i* to be of a specific type, such as normal, then it may be that the upper bound will be lower than 2, depending on the assumption of the distribution.

(f) In general, the higher the overlapping index O_{ji} the lower will O_{ij} be. That is, the more group *j* is included in the range of distribution *i*, the less distribution *i* is expected to be included in the range of *j*.

Properties (a) to (f) show that O_{ji} is an index that measures the extent to which population *j* is included in the range of group *i*. Note that the indices O_{ji} and O_{ij} are not inter-related by a simple relationship. It is clear that the indices of overlapping are not independent.

3.2 Between group component G_b and its properties

As will be seen later, we are interested in two alternative parameters representing between-groups Gini. We start with the one appearing in Equation (3). The between group inequality G_b is defined in Yitzhaki and Lerman (1991) as:

$$G_b = \frac{2\,\mathrm{cov}\left(\overline{Y}, \overline{F}_u\right)}{\mu_u}. \qquad (7)$$

G_b is twice the covariance between the mean income of sub-populations and the sub-populations' mean ranks in the overall population, divided by overall expected income. That is, each sub-population is represented by its mean income,

and the mean rank of its members in the overall distribution. The term G_b equals zero if either the mean incomes or the mean ranks are equal for all sub-populations. In extreme cases, G_b can be negative, which occurs when the mean income is negatively correlated with mean rank.

One may argue that G_b is not really a Gini coefficient because it can be negative. An alternative between-groups Gini(G_{bp}) was defined by Pyatt (1976). (Mookherjee and Shorrocks (1982), Shorrocks (1984) and Silber (1989) also follow Pyatt). In this definition, the between-groups Gini is based on the covariance between mean income in each sub-population and its rank among the mean incomes of sub-populations. The difference between the two definitions is in the rank that is used to represent the group: under Pyatt's approach it is the rank of the mean income of the sub-population, while under Yitzhaki-Lerman it is the mean rank of all members. These two approaches yield the same ranking if complete stratification occurs in the population. It can be shown that:

$$G_b \leq G_{bp} \quad . \qquad (8)$$

The upper limit is reached and (8) holds as an equality, if the ranges of incomes that groups occupy do not overlap (i.e. perfect stratification).

Having explained the different components we now present a variation of decomposition (3) that will be used in this paper as:

$$G_u = \sum_{i=1}^{n} s_i G_i + \sum_{i=1}^{n} s_i G_i (O_i - 1) + G_{bp} + (G_b - G_{bp}) \quad . \qquad (9)$$

For the benefit of readers who are interested in a quick comparison with ANOVA, a summary table of ANOGI is shown below. The four components can be divided into two types: those which carry equivalent information to ANOVA (when using Gini instead of the variance as a measure of variability), and those with additional information.

Table 3.1: A Summary of ANOGI components in comparison to ANOVA

Component Identical to ANOVA	Formula	Range
Intra-Group	IG= $\sum_{i=1} s_i G_i$	$0 \le IG \le G_u$
Between-Group-Pyatt	BGp=G_{bp}	$0 \le BG_p \le G_u$
Additional Information		
Overlapping Effect on Intra-Group	IGO= $\sum_{i=1}^{n} s_i G_i (O_i - 1)$	
Overlapping Effect on Between-Group	BGO=G_b-G_{bp}	$-BG_p - IGO - IG \le BGO \le 0$

3.3 Summary of the decomposition components

3.3.1 Components which are identical to ANOVA:

Intra-Group component (IG): A weighted average of Groups' Ginis. It reaches the lower limit if all intra-group Ginis are equal to zero. It reaches the upper limit if all groups are identical (identical to MSE in ANOVA).

Between-Group component, based on Pyatt (BG_P*)*. It reaches the upper limit if all groups are concentrated at their means. It reaches the lower limit, zero, if the means of all groups are equal (identical to MSB in ANOVA). It measures between-group inequality, assuming a complete stratification.

3.3.2 Additional Components:

The effect of overlapping on intra-group component (IGO): This term "revises" the contribution of each subpopulation to intra-group variability, provided that inequality in the group is greater than zero. If the sub-population and the overall population are equally distributed, then there is no revision to its contribution (O_i=1). However, if a sub-population forms a strata in the population (O_i<1), then its contribution to the intra-group component is reduced, while its contribution to between-group is increased. On the other hand, if the scatter of the ranks of group members is larger than that of the population (O_i>1), the contribution of the group to intra group is increased, while its contribution to between-group is decreased.

The effect of overlapping on between-group component (BGO): The effect of overlapping on the between group component occurs only if the expected values of the subpopulations are not all equal. It is always non-positive, because overlapping

reduces the ability to distinguish between the groups. It reaches the upper limit (zero) if the ranges occupied by the different groups do not overlap. Note, however, that the combined effect of the between-group inequality and the impact of overlapping on it can be negative if the means of the groups are negatively correlated with the means of the ranks. This possibility occurs if, for example, the population is composed of two groups, with one group composed of a majority of poor people and a few very rich people, while the second group is composed of the middle class. In this case, the expected income of the first group is high (because of the few rich) while its expected rank is low (because of the majority of poor people), making the correlation negative.

Finally, an alternative and technical interpretation of equation (9) is as follows: the first term represents the variability of the variant within each group, the second term represents the variability of the expected values among groups, the third term represents the variability of the ranks in each group in the overall population, while the forth term represents the variability of the expected ranks.

In the empirical application we seek to find out whether all the intra-group Ginis are equal, and whether the second, third and fourth terms all converge to zero. We can interpret the terms in the following manner:

G_{bp} =0 implies that all expected values are equal, $(G_b - G_{bp})$ =0 implies that the expected ranks of the sub-populations in the overall population are equal, while $\sum_{i=1}^{n} s_i G_i (O_i - 1)$ =0 implies that each sub-population perfectly overlaps with the entire population. Comparison of the Gini's insures that variability is the same.

Clearly, we are using terms that are connected. However, each parameter adds insight, and there is no redundancy or double counting because the sum of all of them adds up to the overall Gini, and one can produce examples where one term is equal to zero and the others are not. The advantage of ANOGI over ANOVA is that the decomposition of Gini adds a new parameter to the existing inter and intra terms, namely the overlapping index. Hence, not only are the equivalents of first and second moments examined, but the extent of population intertwining is also considered.

4 Estimation and Testing

The decomposition (9) involves four parameters, which need to be estimated from the data: G_i, O_i, G_b and G_{bp}.

The estimation technique used here is based on U-statistics. For each parameter, a kernel of the proper degree is found and then, a U-statistic is constructed. The advantage of dealing with U-statistics is that they are unbiased estimators and their limiting distribution is normal, under regularity conditions (see, for example, Randles and Wolfe (1979), and Hoeffding (1948)). Also, the jackknife method for variance estimation works well for U-statistics (see Shao and Tu (1995), Arvesen (1969) and Schechtman and Wang (2004)). Since the estimation procedures were already detailed elsewhere, we chose to provide the estimators here and refer the reader to the relevant literature for details.

a) Estimation of G_i

Let $Y_1,\ldots,Y_n$ be a random sample from subgroup i, with a distribution function $F_i(y)$, then a U-statistic for estimating G_i, which is an unbiased estimator, is given by

$$\hat{G}_i = \frac{2}{n_i(n_i-1)} \sum_{i<j} | y_i - y_j |$$

where n_i is the sample size coming from sub-population *i* (see Schechtman and Yitzhaki, 1987 for details).

b) Estimation of O_i

Recall that the numerator of O_i is a covariance, which can be expressed as a function of three means, as shown below. The denominator is simply the Gini of sub-population *i*. Therefore, we represent O_i as

$$O_i = O_{ui} = \frac{\mathrm{cov}_i(y, F_u(y))}{\mathrm{cov}_i(y, F_i(y))} = \frac{E_i(yFu(y)) - E_i(y)E_i(Fu(y))}{G_i} = \frac{\theta_1 - \theta_2\theta_3}{G_i}.$$

Each mean is estimated by a U-statistic, and hence, the estimator of Oi is a function of four (dependent) U-statistics.

$$\hat{\theta}_1 = \frac{1}{n_i n_u} \sum_{j=1}^{n_i} y_j (\#\, y's \le y_j)$$

$$\hat{\theta}_2 = \bar{y}$$

and

$\hat{\theta}_3 = \frac{1}{n_i n_u} \sum_{j=1}^{ni} (\#\, y's \le y_j)$, where n_u is the size of the entire population.

Combining the pieces together, the estimator of O_i, based on four dependent U-statistics, is

$$O_i = \frac{\hat{\theta}_1 - \hat{\theta}_2 * \hat{\theta}_3}{\hat{G}_i}$$

Details are given in Schechtman, 2000.

c) Estimation of G_b

The parameter G_b is defined as :

$$G_b = \frac{2\,\mathrm{cov}(\overline{Y}, \overline{F}_u)}{\mu_u}$$

Where $\overline{F}_u$ is the vector of average ranks of the members of the *n* sub-populations, ranked within the entire population. The denominator of G_b can easily be estimated by the sample mean. The numerator can be written as a function of three expectations:

$$\mathrm{cov}(\overline{Y}, \overline{F_u}) = E(\overline{Y}\overline{F}_u) - E(\overline{Y})E(\overline{F}_u).$$

The estimators of $E(\overline{Y}\overline{F}_u)$ and of $E(\overline{F}_u)$ involve the sample version of $\overline{F}_u$. Let $\overline{F}_{u_t}$ be the *t*-th component of $\overline{F}_u$, then

$\hat{\overline{F}}_{u_t} = \dfrac{\sum(\#\, y \le y_i)}{n_u n_t}$ where the summation is over $y_i \in$ sub-population t, t=1,…,n.

Then, $E(\overline{Y}\,\overline{F}_u)$ is estimated by $\dfrac{1}{n}\sum_{t=1}^{n}\overline{Y}_t\hat{\overline{F}}_{u_t}$.

d) Estimation of G_{bp}

The parameter G_{bp} is actually a Gini of the vector of means. Therefore, its estimator is basically the same as $\hat{G}_t$, after replacing Y_i by $\overline{Y}_i$.

As mentioned above, the estimators are U-statistics or functions of several U-statistics. Therefore, inference can be made, using the fact that their limiting distributions are approximately normal, under regularity conditions. The only missing link here is a way to estimate the variances, which are difficult to obtain analytically. We therefore estimated the variances using the jackknife method. The method, which can be best described as "delete one at a time" can be generally explained as follows: given a sample $X_1,\ldots X_n$ of size n, and a sample statistic g(X), whose variance needs to be estimated, follow the two steps:

1. Calculate n values $g_i(X)$, i=1,…n, where $g_i(X)$ is g(X), computed for the original sample, after deleting X_i (i.e. based on (n-1) observations)
2. Use the n values $g_i(X)$ to estimate the variance of g(X) by

$$\frac{n-1}{n}\sum(g_i(X)-\overline{g}_{\cdot}(X))^2$$

where $\overline{g}_{\cdot}(X)$ is the average of $g_1(X),\ldots,g_n(X)$. (For details see, for example, Shao and Tu, 1995).

The case of jackknifing a two-sample statistic is a bit more complicated and we will not go into details here. The interested reader can find the details in Arvesen (1969) and Schechtman and Wang (2004).

5 Results of ANOGI comparing different samples

This section provides empirical results of the decomposition of the Gini by different SOEP sub-samples ("old" sub-samples A through E versus "new" sub-sample F) for two different income variables. Related to the theoretical considerations in Sections 3 and 4 we would expect the following results from the empirical application if both sub-samples represent the exact same population or universe ("=" means no significant difference):

- Mean income $\mu_{AE} = \mu_F$
- Mean rank: $F_{AE} = F_F = 0.5$
- Gini coefficient: $G_{AE} = G_F$
- Overlapping Index: $O_{AE} = O_F = 1$
- Between Group inequality: $G_b = 0$

Any *significant* deviation from these results would have to be interpreted as an indication that the two sub-samples do *not* represent the exact same population.

In order to analyze whether our results on the income distribution are driven by selective attrition or by changing answering behavior of respondents (see Section 2.2 above), we complement this analysis by using another objective variable, namely years of education[12], and by a subjective variable, namely life satisfaction. Education is a non-complex concept which is not very difficult for respondents to report. However, there is some evidence that answers to questions on satisfaction vary in quality during the time span of a panel, as over the course of the first (three) waves respondents learn to deal with this complex concept better (cf. Landua, 1991).

We analyze two income concepts: on the one hand, annual post-government household income (i.e., post-tax post-transfer[13]) which is a generated variable based on an explicit aggregation of various income components (labor income, capital income, private and public transfers such as pensions, child allowances, social

[12] This information is derived from various variables on formal qualification levels for schooling and vocational training.

[13] The process of deriving annual income figures in the SOEP and the tax-simulation procedures are described in Butrica (1997) and Schwarze (1995).

assistance, etc.) across household members, and on the other hand the monthly disposable household income ("screener") asked in the household questionnaire[14]. Missing data due to item non-response in the components of annual income are imputed by means of longitudinal and various cross-sectional techniques. Details on the imputation procedure are given in Appendix E.

In stark contrast to the annual income figures, the monthly screener variable is not imputed in case of missing data; the share of item non-response here ranges between 5% and 10%. The question of the "income screener" itself appears to be a rather simple one[15]. However, it is not easy to give a proper answer because the respondent, in most cases the household head, must calculate the net income from different income sources and, in case of larger households, across several household members.

The variable "Years of education" is analyzed only for the prime age population (aged 25-55). As is the case for income, this variable is also an objective variable, describing an important social and demographic dimension. However, asking for educational attainment appears to be not as complex and sensitive as asking for income.

Finally, we also make use of the subjective measure "satisfaction with life in general", which becomes an increasingly important indicator in socio-economic analyses (as a proxy for utility) as well as in psychological research. An important advantage of this concept is that in the answers of the respondents there are almost no item non-responses. However, there are "panel effects", as we believe for answers on income, in the sense that respondents learn to handle the question more sensitively over the course of time.

---

[14] Both are adjusted – without having an impact on the methodological research question - for different household needs by the modified OECD equivalence scale. This scale is used to assign the appropriate weight to each household member in the sample. This scale gives the first adult a weight of 1.0, additional adults (over 14 years of age) a weight of 0.5, and children (up to age 14) a weight of 0.3.

[15] The original question reads: "If you take a look at the total income from all members of the household: how high is the monthly household income today?

Please state the net monthly income, which means after deductions for taxes and social security. Please include regular income such as pensions, housing allowance, child allowance, grants for higher education support payments, etc. If you do not know the exact amount, please estimate the amount per month.".

In Tables 5.1-5.3 we present the results of income distribution analyses and the Gini decomposition (ANOGI) for annual income in the two SOEP sub-samples "A-E" and the new sub-sample "F". In general, statistically significant differences in the Gini coefficient can be found in 2000, however disappear thereafter (values in parentheses give standard errors according to jackknife estimators). These differences are in need of an explanation given that the two sub-samples are intended to represent the (same) population of individuals living in households in Germany[16]. The long running sub-samples "A-E" are in fact a conglomerate of five different population subgroups (see Section 2) which have partly been added in order to cope with changes in the German population caused by reunification in 1990 and by ongoing immigration while sub-sample "F" represents just *one* big enlargement sub-sample drawn in 2000.[17]

Tables 5.1-5.3 show that the average income in sub-sample F is lower in 2000, adapts to the level of sub-samples A-E in 2001 and is almost identical in 2002. The mean rank for sub-samples A-E in the overall distribution (normalized between 0 and 1) decreases from each period to the next (0.514 to 0.5). Accordingly, the mean rank for sub-sample F increases from 0.483 to 0.5. The group-specific Gini coefficients are only significantly different for the first wave of sub-sample F. Inequality between groups is extremely low in all three years: in 2000 it starts at 0.22% of overall inequality and disappears completely in 2002.

The overlapping information shows that the identification of these two sub-samples as distinct "groups" in terms of their position in the income distribution is only given in 2000. If the overlap component is larger than one, the distribution has

---

[16] Institutionalized households are included in all empirical analyses presented here. Sensitivity analyses focusing on the impact of this sub-population on income distribution measures show the expected result: income of such non-private households is below average and inequality decreases when excluding these households from the analysis. Nevertheless, the substantive finding in Table 5.1 concerning the significant deviation of the Gini-coefficients for the two SOEP sub-samples persists: 0.264 for sub-samples A-E vs. 0.279 for sub-sample F instead of 0.265 vs. 0.281, respectively. Further details about coverage of institutionalized households in the SOEP are given in Appendix B.

[17] It should be noted that households consisting solely of adult respondents who recently immigrated to Germany (i.e. after 1998) had a positive sampling probability in the new sub-sample F. However, this was not the case in the one which already existed (A-E), where the most recent sub-sample was drawn in 1998. However, in our data this phenomenon appears to be of minor relevance, given that there are only six such households in sub-sample F.

the highest relative densities at the tails of the other group-specific distribution, which is the case for sub-sample F in 2000.

Table 5.1: ANOGI for Sub-Samples A-E and F (Year 2000)

(1) Group	(2) Frequency (Pi)	(3) Income share (Si)	(4) Mean Income (μi)	(5) Mean rank (Fio)	(6) Gini (Gi)	(7) Overlapping component (Oi)
Sample A-E	0.55	0.56	33601	0.51	0.265 (.00284)	0.983 (.00216)
Sample F	0.45	0.44	32285	0.48	0.281 (.00261)	1.018 (.00220)
Total	1.00	1.00	33010	0.50	0.272 (.00214)	-
Between group G_b			0.0006 (.00022)	00.22%		
Within group			0.2716	99.78%		
Between Group / max. Between Group (G_b/G_{bp})			0.0607			
G_b - G_{bp} (see equation 9)			-0.0093			

Source: Authors' calculation from SOEP 2000. Standard errors are given in brackets.

Table 5.2: ANOGI for Sub-sample A-E and F (Year 2001)

(1) Group	(2) Frequency (Pi)	(3) Income share (Si)	(4) Mean Income (μi)	(5) Mean rank (Fio)	(6) Gini (Gi)	(7) Overlapping component (Oi)
Sample A-E	0.55	0.56	34332	0.51	0.266 (.00375)	0.997 (.0023)
Sample F	0.45	0.44	33461	0.49	0.267 (.00210)	1.003 (.0023)
Total	1.00	1.00	33941	0.50	0.266 (.00223)	-
Between group			0.0002 (.00013)	00.07%		
Within group			0.2660	99.93%		
Between Group / max. Between Group			0.0310			
G_b - G_{bp} (see equation 9)			-0.0062			

Source: Authors' calculation from SOEP 2001. Standard errors are given in brackets.

Table 5.3: ANOGI for Sub-sample A-E and F (Year 2002)

(1) Group	(2) Frequency (Pi)	(3) Income share (Si)	(4) Mean Income (μi)	(5) Mean rank (Fio)	(6) Gini (Gi)	(7) Overlapping component (Oi)
Sample A-E	0.55	0.55	35220	0.50	0.282 (.00305)	1.0004 (.00254)
Sample F	0.45	0.45	35427	0.50	0.284 (.00256)	0.9997 (.00263)
Total	1.00	1.00	35313	0.50	0.283 (.00228)	-
Between group			0.0000 (.00002)	000.00%		
Within group			0.2827	100.00%		
Between Group / max. Between Group			0.0014			
G_b - G_{bp} (see equation 9)			-0.0015			

Source: Authors' calculation from SOEP 2002. Standard errors are given in brackets.

Possible reasons for the significant differences in the first year could be panel attrition, the imputation models used to adjust for item non-response, or respondent behavior effects.

- If attrition was the cause for the distinctiveness of the two sub-samples then those who dropped out from the survey were systematically different from those persons who were willing to further participate, something already notable in the second wave.
- The explanation via the imputation of missing values could be a relevant issue if the assumption of missing at random (MAR[18]) does not hold or the imputation model (or parts of it) was not correctly specified.
- Last, but not least, response behavior may cause the significant differences in the results for 2000, if there were changes in the behavior due to learning effects in using and answering a complex questionnaire or/and by an improved personal relationship between respondent and interviewer[19] which enhanced confidence.

In order to differentiate between the different causes we introduce two amendments to the further analysis. Firstly, an income concept which is not influenced by any imputation strategy in case of missing information (i.e. the monthly income "screener"), and secondly, a balanced panel design which considers only those observations which were part of the survey for three consecutive years, i.e. 2000 to 2002. If panel attrition caused the differences in 2000 then these differences should disappear when using a 3-year balanced panel, in contrast to the cross-sectional population, especially in wave one. If the imputation procedure was causing the differences for 2000, then vanishing significant results if using a non-imputed income concept could be an indication in this direction.

Table 5.4 shows the comparison of the results for the annual income and the monthly screener income. By definition, the share of missing values for the annual income is zero whereas the share of imputed values for the screener is zero. Note that

[18] See for a detailed description of missing data within surveys e.g. Little and Rubin (2002) or Schafer (1997).

[19] In principle, each year the same interviewer consults the very same interviewees in SOEP.

the trend for the sub-samples A-E is more or less stable for the share of missings as well as for the share of imputed values. The trend for sub-sample F for the two income concepts is also rectified, but different to sub-samples A-E. Item non-response in the monthly income, as well as the share of imputed values in the annual income are higher in the first year and converge to the level of sub-samples A-E. The small increase from wave 2 to wave 3, i.e. from 2001 to 2002, may be linked to the introduction of the Euro on 1st January 2002, which complicated answering due to a lack of familiarization to the "new" currency.

Table 5.4: Comparison of annual and monthly income (Cross-sectional)

	Annual income			**Monthly income (Screener)**		
			Mean (in DM)			
	A-E	**F**	**Total**	**A-E**		**Total**
2000	33,601	32,285	33,009	2,543	2,486	2,517
2001	34,332	33,461	33,941	2,594	2,527	2,565
2002	35,208	35,403	35,296	2,691	2,635	2,666
			Gini * 100			
Year	**A-E**	**F**	**Total**	**A-E**	**F**	**Total**
2000	26.48	28.07	27.22	24.58	25.92	25.19
2001	26.57	26.67	26.62	24.59	25.13	24.84
2002	28.16	28.39	28.72	25.53	26.43	25.94
			Overlapping component (O_i)			
	A-E	**F**	**Total**	**A-E**	**F**	**Total**
2000	0.983	1.018	-	0.987	1.016	-
2001	0.997	1.003	-	0.993	1.008	-
2002	1.0004	0.9997	-	0.994	1.007	-
			% Missing			
	A-E	**F**	**Total**	**A-E**	**F**	**Total**
2000	-	-	-	6.58	9.78	8.02
2001	-	-	-	5.10	8.20	6.49
2002	-	-	-	5.91	8.86	7.24
			% at least one missing income component			
	A-E	**F**	**Total**	**A-E**	**F**	**Total**
2000	20.42	27.68	23.69	-	-	-
2001	18.73	24.29	21.23	-	-	-
2002	20.55	23.91	22.06	-	-	-

Source: Authors' calculation from SOEP 2000-2002.

Unsurprisingly, the Gini coefficients for the two income concepts are different in terms of magnitude[20], but they are identical in terms of trends and changing patterns. The increase in the Gini for 2002 appears to be very distinct. However, this fits the development of increasing income inequality in Germany since the second half of the 1990s.

Table 5.5: Comparison of annual and monthly income (3-year balanced panel design)

	Annual income			Monthly income (Screener)		
			Mean (in DM)			
	A-E	**F**	**Total**	**A-E**	**F**	**Total**
2000	33,557	31,844	32,783	2,538	2,472	2,509
2001	34,409	33,340	33,926	2,590	2,529	2,563
2002	35,642	35,668	35,654	2,702	2,653	2,680
			Gini * 100			
Year	**A-E**	**F**	**Total**	**A-E**	**F**	**Total**
2000	26.19	27.62	26.87	24.40	25.36	24.84
2001	26.22	26.47	26.35	24.36	24.97	24.64
2002	27.67	28.38	27.99	25.58	26.37	25.94
			Overlapping component (O_i)			
	A-E	**F**	**Total**	**A-E**	**F**	**Total**
2000	0.9827	1.0171	-	0.9896	1.0123	-
2001	0.9953	1.0045	-	0.9922	1.0090	-
2002	0.9976	1.0030	-	0.9966	1.0041	-
			% Missing			
	A-E	**F**	**Total**	**A-E**	**F**	**Total**
2000	-	-	-	5.62	10.02	7.61
2001	-	-	-	5.28	7.84	6.44
2002	-	-	-	5.94	8.11	6.92
			% at least one missing income component			
	A-E	**F**	**Total**	**A-E**	**F**	**Total**
2000	19.93	26.97	23.11	-	-	-
2001	18.03	23.42	20.47	-	-	-
2002	20.18	23.51	21.69	-	-	-

Source: Authors' calculation from SOEP 2000-2002.

[20] Note that the annual income concept used here clearly differs from that one of monthly income which is observed from regular (normally monthly) income flows. Following the recommendations of the Canberra Group (2001), our measure of annual income explicitly considers capital income, irregular cash income components like Christmas bonuses or gratifications, as well as a major non-cash income component, namely imputed rent from owner occupied housing. Due to the rather unequal distribution of these income components, inequality for annual income is higher than for monthly income.

Performing the same analysis for the *balanced* panel should control for panel attrition and provide an estimate for the degree of selectivity (see Table 5.5). The effect observed in the cross-sectional analysis is also present in the longitudinal population. We see differences for the first year with overlapping indices significantly different from one and a rather quick convergence in the results over time. This holds not only for Gini, mean and overlapping index, but also, slightly less distinct, for the share of item non-response and the mass of imputed income.

In conclusion, the comparison of Tables 5.4 and 5.5 indicates strong evidence that neither panel attrition nor imputation of item-non-response cause the differences between the results for the first wave of sub-sample F and the longer running sub-samples A-E. If learning and confidence building are important within empirical surveys, this phenomenon should be found in other variables as well. However, if learning effects are not relevant, results should remain stable over time.

- According to the literature, it is well known that the "response styles" for *satisfaction* questions change over time (see e.g. Schräpler, 2001). Based on SOEP-data, Landua (1991) has shown that respondents of questions about satisfaction change their answering behavior over the first four years. Within the first years the respondents tend to overstate their satisfaction more often than in later waves by ticking the highest two categories on an eleven point scale running from zero (“completely dissatisfied”) to ten (“completely satisfied”). On the basis of this finding we may expect, with respect to life satisfaction, differences for all three years, but with a declining trend.
- On the other hand, the results for *educational attainment*, being objective and not an intimate information to ask for, should not be different between the two sub-samples, not even in the first wave.

In order to differentiate attrition from learning effects, all analyses are carried out for the cross-sectional population, as well as for the balanced panel design. In fact, the analysis of educational attainment shows no significant differences between the two sub-samples for all three years in cross-sectional and longitudinal design (see left and right panel of Table 5.6). Especially in the starting year, as hypothesized, means and overlapping indices are similar. As can be expected (at least for the

balanced panel population), we find that means increase in both sub-samples, and inequality does follow the same trend in both samples as well (though it is not *a priori* clear whether to expect an increase or a decrease in inequality of educational attainment).

Table 5.6: Comparison of educational attainment (years of education)

	Cross-sectional design			3-years balanced panel design		
			Mean (in years)			
	A-E	**F**	**Total**	**A-E**	**F**	**Total**
2000	12.51	12.50	12.50	12.51	12.51	12.51
2001	12.54	12.61	12.57	12.53	12.58	12.56
2002	12.54	12.60	12.56	12.56	12.60	12.57
			Gini * 100			
Years	**A-E**	**F**	**Total**	**A-E**	**F**	**Total**
2000	10.51	9.97	10.27	10.39	9.75	10.11
2001	10.54	9.92	10.26	10.44	9.84	10.18
2002	10.46	9.92	10.22	10.52	9.87	10.24
			Overlapping component (O_i)			
	A-E	**F**	**Total**	**A-E**	**F**	**Total**
2000	1.0007	1.0012	-	1.0029	0.9983	-
2001	1.0077	0.9930	-	1.0081	0.9917	-
2002	1.0054	0.9958	-	1.0086	0.9909	-

Source: Authors' calculation from SOEP 2000-2002.

As expected, the results for life satisfaction draw a very different picture (see Table 5.7). The Gini indices are more unequal and interestingly the differences in the results between cross-sectional and longitudinal population are larger within sub-samples A-E than within sub-sample F[21].

With respect to mean as well as the marginal distribution of the variable "life satisfaction" (see Table F-1 in the appendix) our results clearly confirm the finding by Landua (1991) which states that first time users of such a scale tend to tick the highest categories more often. In 2000, more than twice as many respondents in sub-sample F than in sub-samples A-E indicated that they are "completely satisfied": 9.7% and 4.1%, respectively. This gap decreases most remarkably to only 2.1% (3.2% among sub-samples A-E vs. 5.3% in sub-sample F). Again this picture does not change when moving from a purely cross-sectional to a longitudinal design,

[21] In order to run the Gini calculation properly on positive values only, we transformed the original eleven point scale in the following way: values 1 through 10 have been multiplied by 10 and the value 0 was coded into 0.1.

which we interpret as an indication for learning effects among the short panel members.

Table 5.7: Comparison of life satisfaction

	Cross-sectional design			3-years balanced panel design		
			Mean			
	A-E	F	Total	A-E	F	Total
2000	6.89	7.28	7.07	6.91	7.33	7.10
2001	6.93	7.26	7.08	6.96	7.28	7.10
2002	6.77	7.07	6.91	6.74	7.06	6.88
			Gini * 100			
Years	A-E	F	Total	A-E	F	Total
2000	15.17	14.55	14.94	14.74	14.15	14.54
2001	15.24	13.89	14.68	14.88	13.72	14.41
2002	15.66	14.39	15.13	15.79	14.38	15.21
			Overlapping component (O_i)			
	A-E	F	Total	A-E	F	Total
2000	0.9806	1.0064	-	0.9791	1.0048	-
2001	0.9976	0.9930	-	0.9971	0.9934	-
2002	0.9989	0.9929	-	1.0025	0.9882	-

Source: Authors' calculation from SOEP 2000-2002.

6 Concluding summary

The main aim of this paper is to study the "representativeness" of different sub-samples of the German SOEP on the field of income distribution. This issue was chosen for analysis because an unbiased measurement of household incomes is (a) a real challenge for survey research (cf. Canberra Group 2001) and (b) the analysis of income distribution and mobility is one of the main tasks of household panel surveys such as the SOEP.

However, it appears that the inclusion of a new (independently drawn) representative sub-sample into an existing, longer-running panel survey may yield slightly deviating results, which may be caused by panel attrition or by differences in the answering behavior of respondents.

The methodology used in this paper is based on the analysis of Gini (ANoGi) which differs from the analysis of variance because it includes an additional term which reflects the overlapping between the distributions of the different sub-samples. This is the first time that this methodology is empirically applied and we believe that the paper demonstrates its usefulness.

Concluding from our empirical results and a discussion of survey methodology issues employed in the set-up of the considered sub-samples, we cannot reject the hypothesis that both represent the same universe. Recapitulating from our analyses on objective and subjective indicators for income, education and satisfaction within a cross-sectional, as well as a longitudinal framework, we conclude that there is convincing evidence within the SOEP for changing respondent behavior due to learning effects with respect to the applied instruments and questioning. However, we find the convergence process, in which empirical results based on a new sub-sample approach those of a longer running sub-sample, to be of different lengths for the various indicators under investigation. This may be driven by the different degree of complexity of the underlying constructs, especially in case of "satisfaction".

With respect to the originally motivating question on income inequality, we would especially reject the hypothesis that due to panel attrition, a new sub-sample after two waves is as selective as a longer running panel. Instead of arguing that results from a cross-section survey (as is the first wave of any panel study) yield more reliable estimates than those stemming from a panel which may be affected by attrition, we would like to reverse this argument and state that a reliable measurement of complex issues, such as the construction of an annual income measure or a satisfaction measure, clearly profits from repeated surveying as is the case in panel studies.

References:

Arvesen, J.N. (1969). Jackknifing U-statistics. Annals of Mathematical Statistics, 40, 2076-2100.

Butrica, B. A. 1997: Imputation methods for filling in missing values in the PSID-GSOEP Equivalent File 1980-1994, Cross-National Studies in Aging. Program Project Paper, Center for Policy Research, The Maxwell School. Syracuse, NY: Syracuse University

Canberra Group (2001), Expert Group on Household Income Statistics: Final Report and Recommendations, Ottawa.

Couper, M.P. and Nicholls II, W.L. (1998), "The History and Development of Computer Assisted Survey Information Collection." In M.P. Couper, R.P. Baker, J. Bethlehem, C.Z.F. Clark, J. Martin, W.L. Nicholls, and J. O'Reilly (eds.), Computer Assisted Survey Information Collection. New York: Wiley.

Dagum, C. (1980). "Inequality Measures Between Income Distributions With Applications," Econometrica, 48, 7,1791-1803.

Dagum, C. (1985). "Analysis of Income Distribution and Inequality by Education and Sex in Canada," Advances in Econometrics, 4, 167-227.

De Leeuw, E. D. (2002). The effect of computer assisted interviewing on data quality: A review of the evidence [CD-Rom]. In: J. Blasius, J. Hox, E. de Leeuw & P. Schmidt (Eds.), Social science methodology in the new millennium. Opladen, FRG: Leske + Budrich

Frick, J.R. and M. M. Grabka (2003): Missing Income Data in the German SOEP: Incidence, Imputation and its Impact on the Income distribution. DIW-Discussions Papers No. 376, October 2003, Berlin: DIW Berlin.

Fuchs, M.; Couper, M. and S. E. Hansen (2000): Technology Effects: Interview Duration in CAPI and Paper and Pencil Surveys. In: Ferligoj, Anuska; Mrvar, Andrej (Hrsg.): Developments in Survey Methodology. Ljubljana, Slovenia, S. 149-166.

Grabka, M. M. and J. R. Frick (2003), Imputation of Item-Non-Response on Income Questions in the SOEP 1984–2002, DIW Research Notes No. 29, DIW Berlin.

Haisken-DeNew, J. P. and J. R. Frick (2003): Desktop Companion to the German Socio-Economic Panel Study (GSOEP), Version 7.0 – Update to Wave 19, DIW Berlin.

Hoeffding,W. (1948). A Class of Statistics With Asymptotic Normal Distrubution. Annals of Mathematical Statistics,19,293-325.

Landua, D. (1991): "An Attempt to Classify Satisfaction Changes: Methodological and Content Aspects of a Longitudinal Problem", Social Indicators Research, 26, 221-241.

Lasswell, T. E. (1965). Class and Stratum, Houghton Mifflin Company, Boston, Massachusetts.

Laurie, H. (2000) 'From PAPI to CAPI: consequences for data quality on the British Household Panel Survey', paper presented at the Fifth International Conference on Logic and Methodology, University of Cologne, October 3 - 6, 2000.

Laurie, H. and Moon, N. (1997) 'Converting to CAPI in a Longitudinal Panel Survey' BHPS Working Paper No 97-11, ESRC Research Centre on Micro-social Change, University of Essex.

Lerman, R. and S. Yitzhaki (1984). “A Note on the Calculation and Interpretation of the Gini Index,” Economics Letters, 15, 363-68.

Little, R.J.A. & Rubin, D.B. (2002). Statistical Analysis with Missing Data (2nd ed.). New York: John Wiley & Sons.

Little, R.J.A. and Su, H.-L. (1989): Item Non-Response in Panel Surveys. In: Kasprzyk, D., Duncan, G., Kalton, G. and Singh, M. P. (eds.): Panel Surveys. John Wiley, New York: 400-425.

Mookherjee, D. and A. F. Shorrocks (1982). "A Decomposition Analysis of the Trend in U. K. Income Inequality," Economic Journal, 886-902.

Pyatt, G. (1976). "On the Interpretation and Disaggregation of Gini Coefficient," Economic Journal, 86, 243-255.

Randles, R.H., and Wolfe, D.A. (1979). Introduction to the Theory of Nonparametric Statistics. John Wiley and Sons, New York.

Rendtel, U. (1995): Panelausfälle und Panelrepräsentativität. Campus Verlag, Frankfurt/Main - New York.

Rendtel, U., Wagner, G. and Frick, J. (1995): Eine Strategie zur Kontrolle von Längsschnittgewichtungen in Panelerhebungen - Das Beispiel des Sozio-Oekonomischen Panels (SOEP). Allgemeines Statistisches Archiv, 79(3), 252-277.

Shao, J. and Tu, D. (1995). The Jackknife and Bootstrap. Springer-Verlag, New York.

Schafer, J.L. (1997). Analysis of Incomplete Multivariate Data. London: Chapman & Hall.

Schechtman, E. (2000) Stratification: Measuring and Inference, Mimeo, Dept. of Industrial Engineering and Management, Ben-Gurion University, Israel.

Schechtman, E. and Yitzhaki, S. (1987). A Measure of Association Based on Gini’s Mean Difference. Commun. Statist.-Theor. Meth.,16(1),207-231.

Schechtman, E., and Yitzhaki, S. (1999). On the Proper Bounds of the Gini Correlation. Economics letters, 63, 133-138

Schechtman, E., and Wang, S. (2004). Jackknifing Two-Sample Statistics. Journal of Statistical Planning and Inference, 119,2, 329-340.

Schwarze, J. (1995), Simulating German Income and Social Security Tax Payments Using the GSOEP, Syracuse University, Syracuse, NY: Cross-National Studies in Ageing Project Paper No. 19.

Schräpler J.-P. (2002): Respondent Behavior in Panel Studies - A Case Study for Income-Nonresponse by means of the German Socio-Economic Panel (GSOEP). In DIW-Discussion Paper, No. 299.

Schräpler, J.-P, J. Schupp and Gert G. Wagner (2004): Changing from PAPI to CAPI: A longitudinal study dealing with Mode-Effects in the German Socio-Economic Panel (SOEP) using an Experimental Design. DIW Berlin: mimeo.

Schräpler, J.-P. and G. G. Wagner (2001): Das Verhalten von Interviewern - Darstellung und ausgewählte Analysen am Beispiel des "Interviewerpanels" des Sozio-ökonomischen Panels, Allgemeines Statistisches Archiv, 85, 45-66.

Schupp, J. and G. G. Wagner (2003), Maintenance of and innovation in long-term panel studies: The case of the German Socio-Economic Panel (GSOEP), Allgemeines Statistisches Archiv, 86 (2), 163-176.

Shorrocks, A. F. (1984). "Inequality Decomposition by Population Subgroups," Econometrica, 52 (6), 1369-1385.

Shorrocks, A. F. (1982). "On The Distance between Income Distributions," Econometrica, 50 (5), 1337-9.

Silber, J. (1989). "Factor Components, Population Subgroups, and the Computation of Gini index of Inequality," Review of Economics and Statistics, 71 (2), 107-115.

Spieß, M. and J. Goebel (2003): Evaluation of the ECHP imputation rules, in CHINTEX Working Paper No. 17.

Spieß, M. and M. Pannenberg (2003), Sample Sizes and Panel Attrition in the German Socio-Economic Panel (GSOEP) (1984 until 2002), in DIW Research Notes No.28.

Spiess, M. and U. Rendtel (2000) Combining an ongoing panel with a new cross-sectional sample, in DIW Discussion Paper No. 198.

Wagner, G., R.V. Burkhauser, and F. Behringer (1993): "The English Language Public Use File of the German Socio-Economic Panel Study", Journal of Human Resources, 28 (2), 429-433.

Yitzhaki, S. (1994). “Economic Distance and Overlapping of Distributions, “Journal of Econometrics, 61, 147-159.

Yitzhaki, S. and R. Lerman (1991). "Income Stratification and Income Inequality," Review of Income and Wealth, 37(3), 313-329.

Appendix A. The German SOEP – details.

The main random sub-sample "A" included around 4,500 households. In order to allow separate analyses of the five groups of labor migrants most strongly represented in the Federal Republic of Germany at that time, they were over-sampled in the study with a total of 1,400 households in a disproportional random sample approach. This random sub-sample "B" was itself subdivided into 5 sub-groups.[22]

In order to observe the massive social and economic changes in East Germany, along with their respective impacts, the first wave of the East German sub-sample was collected in June 1990, *before* the currency, economic, and social union in Germany occurred on July 1st. This sample, sub-sample C, covered about 2,200 households.

Since the start of SOEP in 1984, Western Europe (and especially Germany) has experienced immigration on a large scale, which cannot be covered by any ongoing longitudinal survey. In order to correct for this bias, an explicit supplement for immigrants was necessary. For this reason, sub-sample "D" was collected in 1994/95 for about 500 households of immigrants who had arrived since 1984.

In 1998 a "supplementary random sample" has been started as a test. This sub-sample fulfilled a number of aims: (1) stabilization of the number of observations in the SOEP for cross-sectional and longitudinal data, (2) allowing for analysis of "panel effects" and (3) allowing for analysis of representativeness.

It was proven that a supplementary sample such as this could be integrated in a user-friendly manner into the ongoing "old sub-samples" (see Spiess / Rendtel 2000 for solving the problem of setting up an integrated weighting scheme). Thus the methodological basis was established for significantly increasing the sample size, which would boost the value of the study for policy analysis by allowing the changes for relatively small groups of the population to be analyzed on the basis of sufficiently large numbers of cases. An enlargement such as this took place in the year 2000. The first wave of sub-sample "F" consists of 10,890 adult respondents and 2,993 children who live in 6,052 households.

---

[22] Sub-sample "A" therefore also includes households headed by a foreigner not belonging to the nationalities covered by sub-sample "B", albeit on a negligible scale (e.g., Dutch, Swiss).

Appendix B. SOEP sampling procedures – details.

In Germany, as in many other countries, sampling of foreigners is a problem. Although a *major improvement* of random route samples which are conducted in Germany was implemented with sub-sample F, a difference to sub-samples A and B remains. Because the local polling registers (*Wählerverzeichnisse*) are the basis for drawing sample points, sample points with different shares of non-German citizens are not drawn by probabilities which mirror their correct weight for the population living within German territory (*Wohnbevölkerung*). Thus all standard random route samples (according to the so-called ADM standard procedure) underestimate the share of foreigners in Germany. In order to reduce the impact of this shortcoming, the field-work organization *Infratest* introduced an over-representation of foreigners in the random walk of SOEP sub-sample F. The number of addresses to be collected during a random walk was doubled, but within the so-called "excess addresses" only households with foreigners were selected for interviews. Through this procedure the share of foreigners in the sample mirrors the true share in the underlying population quite well. However, the structure of the foreigners is eventually biased, because sample points with a high share of foreigners have a downward biased probability of being included in the sample. In principle, this bias can be corrected for by weighting procedures (to be applied).

It is a common problem for population surveys around the world to adequately cover households living in institutions. Unfortunately, it was not possible to include institutionalized households in the first waves of the SOEP in a representative manner. However, by following respondents after a residential move, a panel takes into consideration those who left private households for institutionalized households, whereby over the course of time, the institutionalized population is included in the SOEP. However, on the other hand, any new sub-sample starts with this problem, which may produce an artificial difference between old and new sub-samples. Nevertheless, the population in the starting wave of sub-sample F (year 2000) in fact includes 47 institutionalized households (approx. 0.8% of all household interviews in this sample) as compared to 85 institutionalized households in sub-samples A through E (approx. 1.2%).

One may conclude from the discussion of these various sampling procedures applied to SOEP that the respective universe or population to be represented by the different sub-samples (A through E vs. F) differs only marginally.

Appendix C. SOEP response rates - details.

Table C.1: Wave and sample sizes in the SOEP (cross-sectional)

	Sample A-E								Sample F		SOEP
	Soep-West A+B		Soep-East C		Migrants D1+D2		Innovation E		Refreshment F		
Year	wave	Obs.	wave	Obs.	wave	Obs.	wave	Obs.	wave	Obs.	Total
1984	1	12245		-		-		-		-	12245
1985	2	11090		-		-		-		-	11090
1986	3	10646		-		-		-		-	10646
1987	4	10516		-		-		-		-	10516
1988	6	9710		-		-		-		-	9710
1990	7	9519	1	4453		-		-		-	13972
1991	8	9467	2	4202		-		-		-	13669
1992	9	9305	3	4092		-		-		-	13397
1993	10	9206	4	3973		-		-		-	13179
1994	11	9001	5	3945	1	471		-		-	13417
1995	12	8798	6	3892	2	1078		-		-	13768
1996	13	8606	7	3882	3	1023		-		-	13511
1997	14	8467	8	3844	4	972		-		-	13283
1998	15	8145	9	3730	5	885	1	1910		-	14670
1999	16	7909	10	3709	6	838	2	1629		-	14085
2000	17	7623	11	3687	7	837	3	1549	1	10890	24586
2001	18	7424	12	3576	8	789	4	1464	2	9098	22351
2002	19	7175	13	3466	9	780	5	1373	3	8427	21221

Source: SOEPinfo, URL: http://panel.gsoep.de/soepinfo2002/info/persons.html

Table C.2: Development of sample sizes: Two-wave balanced panel design

Year	Sample A-E								Sample F		SOEP
	Soep-West A+B		Soep-East C		Migrants D1+D2		Innovation E		Refreshment F		
	Obs.	% left	Obs.	% left	Obs.	% left	Obs.	% left	Obs.	% left	Total
'84-85	10563	86.3	-	-	-	-	-	-	-	-	10563
'85-86	9941	89.6	-	-	-	-	-	-	-	-	9941
'86-87	9859	92.6	-	-	-	-	-	-	-	-	9859
'87-88	9551	90.8	-	-	-	-	-	-	-	-	9551
'88-89	9190	91.7	-	-	-	-	-	-	-	-	9190
'89-90	9001	92.7	-	-	-	-	-	-	-	-	9001
'90-91	8946	94.0	4033	90.6	-	-	-	-	-	-	12979
'91-92	8845	93.4	3804	90.5	-	-	-	-	-	-	12649
'92-93	8705	93.6	3745	91.5	-	-	-	-	-	-	12450
'93-94	8540	92.8	3708	93.3	-	-	-	-	-	-	12248
'94-95	8387	93.2	3698	93.7	435	92.4	-	-	-	-	12520
'95-96	8199	93.2	3673	94.4	979	90.8	-	-	-	-	12851
'96-97	8023	93.2	3655	94.2	908	88.8	-	-	-	-	12586
'97-98	7779	91.9	3564	92.7	837	86.1	-	-	-	-	12180
'98-99	7506	92.2	3527	94.6	786	88.8	1554	81.4	-	-	13395
'99-00	7294	92.2	3500	94.4	783	93.4	1458	88.3	-	-	13035
'00-01	7060	92.6	3416	92.6	746	89.1	1394	90.0	8617	78.4	21233
'01-02	6839	92.1	3300	92.3	717	90.9	1300	88.8	7957	87.5	20113

Source: SOEPinfo, URL: http://panel.gsoep.de/soepinfo2002/info/persons.html

Table C.3: Development of sample sizes: Three-wave balanced panel design

Year	Sample A-E								Sample F		SOEP
	Soep-West A+B		Soep-East C		Migrants D1+D2		Innovation E		Refreshment F		
	Obs.	% left	Obs.	% left	Obs.	% left	Obs.	% left	Obs.	% left	Total
'84-86	9485	77.5	-	-	-	-	-	-	-	-	9485
'85-87	9256	83.5	-	-	-	-	-	-	-	-	9256
'86-88	9027	84.8	-	-	-	-	-	-	-	-	9027
'87-89	8774	83.4	-	-	-	-	-	-	-	-	8774
'88-90	8556	85.4	-	-	-	-	-	-	-	-	8556
'89-91	8490	87.4	-	-	-	-	-	-	-	-	8490
'90-92	8402	88.3	3657	82.1	-	-	-	-	-	-	12059
'91-93	8307	87.7	3489	83.0	-	-	-	-	-	-	11796
'92-94	8119	87.3	3502	85.6	-	-	-	-	-	-	11621
'93-95	7994	86.8	3482	87.6	-	-	-	-	-	-	11476
'94-96	7843	87.1	3502	88.7	413	87.7	-	-	-	-	11758
'95-97	7683	87.3	3474	89.3	869	80.6	-	-	-	-	12026
'96-98	7395	85.9	3401	87.6	787	76.9	-	-	-	-	11583
'97-99	7206	85.1	3379	87.9	750	77.2	-	-	-	-	11335
'98-00	6970	85.6	3331	89.3	734	82.9	1391	72.0	-	-	12426
'99-01	6783	85.8	3253	87.8	698	83.3	1318	80.9	-	-	12052
'00-02	6545	85.9	3174	86.1	681	81.4	1242	80.2	7560	69.4	19202

Source: SOEPinfo, URL: http://panel.gsoep.de/soepinfo2002/info/persons.html

Appendix D. SOEP interview modes.

"Computer Assisted Personal Interviewing" (CAPI) was introduced in the SOEP within a controlled mode experiment with sub-sample E in 1998 and after a successful testing phase, this survey method was also introduced for the first time in the existing SOEP-sub-samples A through D in 2000 (see table D.1)[23]. In Sample F, this methodology was used from the very first wave (year 2000) yielding an *overall* share of CAPI-interviews of 28% in 2002.

Table D.1: Interview Mode in the SOEP 1999/2002

Interview Mode	1999	2002
Oral Interview/Interviewer	43%	29%
Self-Completed w/o. Interviewer	29%	25%
Written (Snail-Mail)	14%	10%
Self-Completed w. Interviewer	5%	3%
CAPI	5%	28%
Part Oral / Part Self-Completed	4%	4%
Phone	0.07%	0%
Proxy	0.04%	0.04%

Source: Authors calculation from SOEP 1999-2002

Methodological research focusing on sustainable mode or technology effects[24] caused by the introduction of CAPI show a mixed picture. While in the case of the BHPS no significant effects were found[25], for the SOEP Schräpler et al (2004) find some indications that CAPI increases the probability of item-non-response on income questions, but also that it reduces the probability of unit-non-response in the subsequent wave. Thus, the variation in the use of CAPI across the SOEP-sub-samples A through D, E and F yields a potential for survey artifacts.

[23] Information about the mode of the interview are stored together with the survey data and thus interview artifacts can be analyzed by any researcher using SOEP data.

[24] For an overview on this line of discussion see Couper/Nicholls (1998), de Leeuw (2002) and Fuchs et al (2000).

[25] In 1999, CAPI was – finally (Laurie/Moon 1997) - introduced in the British Household Panel Study (BHPS). In contrast to the SOEP, this change of technology was done completely and at once for the whole sample. First results show no mode effects (Laurie 2000).

Appendix E. SOEP Imputation procedures.

Recent studies provide evidence that using only cross-sectional data for imputation of missing data in panel surveys is inferior to using longitudinal data (see Spiess & Goebel 2003, Frick & Grabka 2003). Thus, the imputation of item-non-response related missing income data in the SOEP follows a two step procedure: the general principle is to employ the "row and column imputation technique" as developed by Little and Su (1989) whenever longitudinal income data is available and to apply specific cross-sectional imputation techniques otherwise.

The Little & Su method takes advantage of information on the very same individual over time by combining row (unit) and column (period/trend) information. In principle, the imputed value is the result of a combination of *row effect, column effect and a residual effect.* The column effects are calculated for each year of data and are given by $c_j = (j * Y_j) / \Sigma Y_k$, where *Yj* is the sample mean income for year *j*. The row effects, $r_i = m_i^{-1} * \Sigma (Y_{ij} / c_j)$, are computed for each sample member where Y_{ij} is the income for individual *i* in year *j* and m_i is the number of recorded months with receipt of a given income component over the last year. Sorting cases by r_i and matching the incomplete case *i* with information from the nearest complete case, say *l*, yields the imputed value $i = [r_i] * [c_j] * [Y_{lj} / (r_l * c_j)]$. The first two terms estimate the predicted mean, and the last term is the stochastic component of the imputation from the matched case. Overall, the corresponding bias in variance appears to be somewhat less severe.

However, given that the empirical implementation of Little & Su fails in all those cases where a given income component is not observed in any other wave of data, purely cross-sectional imputation techniques have to be used which are based on data observed from other units (individuals or household, respectively) in the very same wave. See Grabka & Frick (2003) for a complete overview of the techniques applied for the various SOEP income variables. In general, the following cross-sectional imputation techniques are applied:

- *Institutional or external information* is used to logically impute missing amounts of those income components which are perfectly related to otherwise observed

information, e.g. child benefit which is fixed per child or support from the nursing care insurance which is fixed to the observed needs.

- *Median Substitution* is applied for income components which are of minor relevance with respect to the income level (e.g. military service pay, maternity benefit) or in terms of the number of affected observations ($n < 10$). *Median Substitution for Subgroups* is performed for e.g. housing benefit for owner occupiers by household size.
- *Median Share Substitution* takes place if two income variables are clearly linked to each other, e.g. the median share of the monthly labor earnings and the Christmas bonus in the private sector in Germany is about 35%. Following from this, any observation with item-non-response on Christmas bonuses in the private sector is assigned an imputed value given by the individually observed labor income times the (median) share of 35%. This procedure appears to adequately ensure a more realistic variation of the imputed income values than single median substitution methods would do.
- Finally, in case of more complex income constructs such as individual labor income *regression-based imputation* is applied, using a Mincer-type wage regression models.

Appendix F. Distribution of Life satisfaction – details.

Table F-1: Comparison of life satisfaction distributions (in %)

	Cross-sectional design			3-years balanced panel design		
2000	**A-E**	**F**	**Total**	**A-E**	**F**	**Total**
0=low	0.5	0.5	0.5	0.5	0.5	0.5
1	0.4	0.3	0.4	0.4	0.3	0.4
2	1.6	1.0	1.3	1.2	0.9	1.0
3	2.3	2.1	2.2	2.4	1.8	2.1
4	3.9	2.7	3.4	3.7	2.6	3.2
5	13.6	11.6	12.7	13.5	11.4	12.5
6	11.3	9.3	10.4	11.4	9.0	10.3
7	22.3	18.7	20.7	23.1	18.4	21.0
8	29.8	30.7	30.2	29.9	31.6	30.7
9	10.1	13.4	11.6	9.8	13.9	11.6
10=high	4.2	9.7	6.7	4.1	9.7	6.6
2001	**A-E**	**F**	**Total**	**A-E**	**F**	**Total**
0=low	0.5	0.4	0.5	0.4	0.3	0.4
1	0.6	0.2	0.4	0.4	0.3	0.3
2	1.4	0.9	1.2	1.3	1.1	1.2
3	2.5	2.0	2.3	2.4	1.8	2.1
4	4.0	2.4	3.3	4.1	2.2	3.3
5	12.4	11.2	11.8	12.6	10.9	11.8
6	11.1	9.4	10.4	11.0	9.2	10.2
7	22.4	19.9	21.3	22.5	20.5	21.6
8	29.6	32.6	31.0	30.0	32.4	31.1
9	10.8	13.6	12.0	10.6	14.0	12.1
10=high	4.7	7.4	5.9	4.7	7.4	5.9
2002	**A-E**	**F**	**Total**	**A-E**	**F**	**Total**
0=low	0.5	0.4	0.5	0.5	0.4	0.5
1	0.5	0.5	0.5	0.4	0.5	0.4
2	1.6	1.0	1.3	1.8	1.0	1.4
3	2.9	2.4	2.7	3.0	2.4	2.7
4	4.6	3.1	4.0	4.7	3.1	4.0
5	13.8	11.7	12.9	14.1	11.5	12.9
6	11.8	11.2	11.5	12.1	11.1	11.7
7	23.3	21.9	22.7	23.5	22.4	23.0
8	28.1	30.6	29.2	27.2	30.9	28.9
9	9.5	11.5	10.4	9.3	11.4	10.2
10=high	3.2	5.7	4.3	3.2	5.3	4.2

Source: Authors' calculation from SOEP 2000-2002.