

SOEPpapers
on Multidisciplinary Panel Data Research

SOEP – The German Socio-Economic Panel Study at DIW Berlin

556-2013

"Familien in Deutschland" – FiD
Enhancing Research on Families in Germany

Mathis Schröder, Rainer Siegers, C. Katharina Spieß

SOEPPapers on Multidisciplinary Panel Data Research at DIW Berlin

This series presents research findings based either directly on data from the German Socio-Economic Panel Study (SOEP) or using SOEP data as part of an internationally comparable data set (e.g. CNEF, ECHP, LIS, LWS, CHER/PACO). SOEP is a truly multidisciplinary household panel study covering a wide range of social and behavioral sciences: economics, sociology, psychology, survey methodology, econometrics and applied statistics, educational science, political science, public health, behavioral genetics, demography, geography, and sport science.

The decision to publish a submission in SOEPPapers is made by a board of editors chosen by the DIW Berlin to represent the wide range of disciplines covered by SOEP. There is no external referee process and papers are either accepted or rejected without revision. Papers appear in this series as works in progress and may also appear elsewhere. They often represent preliminary studies and are circulated to encourage discussion. Citation of such a paper should account for its provisional character. A revised version may be requested from the author directly.

Any opinions expressed in this series are those of the author(s) and not those of DIW Berlin. Research disseminated by DIW Berlin may include views on public policy issues, but the institute itself takes no institutional policy positions.

The SOEPPapers are available at
<http://www.diw.de/soeppapers>

Editors:

Jürgen **Schupp** (Sociology, Vice Dean DIW Graduate Center)
Gert G. **Wagner** (Social Sciences)

Conchita **D'Ambrosio** (Public Economics)
Denis **Gerstorff** (Psychology, DIW Research Director)
Elke **Holst** (Gender Studies, DIW Research Director)
Frauke **Kreuter** (Survey Methodology, DIW Research Professor)
Martin **Kroh** (Political Science and Survey Methodology)
Frieder R. **Lang** (Psychology, DIW Research Professor)
Henning **Lohmann** (Sociology, DIW Research Professor)
Jörg-Peter **Schräpler** (Survey Methodology, DIW Research Professor)
Thomas **Siedler** (Empirical Economics)
C. Katharina **Spieß** (Empirical Economics and Educational Science)

ISSN: 1864-6689 (online)

German Socio-Economic Panel Study (SOEP)
DIW Berlin
Mohrenstrasse 58
10117 Berlin, Germany

Contact: Uta Rahmann | soeppapers@diw.de

“Familien in Deutschland” – FiD

Enhancing Research on Families in Germany

Mathis Schröder^{*}, Rainer Siegers, and C. Katharina Spieß, DIW Berlin[♠]

May 2013

1. Introduction

There are various independent studies evaluating family policy measures in Germany. So far, a systematic evaluation considering the different goals inherent to these measures was missing. The evaluation of family policy measures on behalf of the Federal Ministry for Family Affairs, Senior Citizens, Women and Youth (BMFSFJ) and the Federal Ministry of Finance (BMF) is thus the first systematic overall study. In a feasibility study conducted for this overall evaluation in late 2008, one of the main conclusions was that “*Without additional data only a limited number of policies regarding families and children can be evaluated*” (authors’ translation, see Beninger et al. 2008). The available data sets were not sufficient for in-depth analyses, especially regarding specific family types which might be rare in the German population, but still important as targets for the ministries’ policies. Such families are especially single parents, large families with more than two children, low-income families, and families with very young children. The main studies to evaluate family policies existing at the time were the Socio-economic Panel (SOEP, see Wagner, Frick & Schupp, 2007), and the Panel Analysis of Intimate Relationships and Family Dynamics (*pairfam*, see Huinink et al., 2011). While the targeted groups are present in the SOEP (general population survey) and in *pairfam* (family survey), the actual case numbers for these families in these studies are far too low to provide sufficient statistical power for an evaluation of family policy measures.

This was the initiation for the data collection effort “Familien in Deutschland” (short FiD, for “Families in Germany”). This project started to collect data in 2010, with the focus on single parents, low income families, large families with three or more children, and families with particularly young children, namely those born between 2007 and 2010. FiD was initially financed by the two federal ministries in charge of the overall evaluation. The

^{*} Corresponding author, contact at mschroeder@diw.de, +49 30 89789-222.

[♠] The authors gratefully acknowledge funding from the “Bundesministerium für Familie, Senioren, Frauen und Jugend” (BMFSFJ) and the “Bundesministerium für Finanzen” (BMF). Thanks are especially in order for the BMFSFJ for financing data the collection in 2013 and thus providing the possibilities for a long-term longitudinal panel.

funding covered three waves of data collection, spanning the years 2010-2012, which were used in various studies for the overall evaluation.¹ A further wave of data collection in 2013 was funded by the Federal Ministry for Family affairs. As FiD collects longitudinal data very similar in content and structure to the SOEP data, the data collection will subsequently become part of the regular SOEP. However, it is already possible to jointly use the two data sets with sampling weights provided for this particular purpose. With the integration scheduled for 2014 (i.e. the FiD households will become part of the regular SOEP in the data collection of 2014), FiD will further strengthen the base of family research in the SOEP data.

The following paper gives an overview of the FiD data – a quick look at the sample sizes and the number of conducted interviews in the first three waves follows in the next section. In section 3, the most important concepts on sampling, structure and weighting are presented, while section 4 offers some important measures of data quality, among them the longitudinal response rates. The last two sections briefly conclude the paper.

2. In a Nutshell: Quick Overview and Sample Sizes

FiD consists of three different samples – a “Screening Sample” drawn in 2010, a “Cohort Sample” drawn in 2010, as well as a “Screening Sample” drawn in 2011.² Overall, the interest is in four different groups of families: those with young children born between 2007 and 2010 (identified in the Cohort Sample), those who have low income, those who have three or more children, and those families with a single parent (all identified in the Screening Samples). The drawn households are repeatedly interviewed, such that FiD is a longitudinal study of households with the given characteristics (at the initial sampling time). The main focus of FiD is on the families and children – the parental questionnaires (filled out about their children) are about twice as long as the comparable questionnaires in the SOEP, and two questionnaires have been added entirely. In addition there are several questions designed to capture the challenges families face with regard to the return of mothers into the labour market – with respect to work place, work schedule, overtime, day care possibilities, etc. For the years of 2010 to 2012, Table 1 shows the sample sizes (households, individuals, children) captured by FiD.

[Table 1 about here]

¹ For an overview of studies, see the homepage of the BMFSFJ (accessed May 2013): <http://www.bmfsfj.de/BMFSFJ/familie,did=195944.html>.

² Even though the sample names are slightly misleading in their terminology (“Screening” describes a sampling procedure, “Cohort” a sample characteristic), these definitions are kept to be consistent to the documentation and the field reports by the survey agency.

3. Concepts

3.1. *Sampling*³

The goal for FiD was to provide data on four groups essential for family policies: single parents, low income families, large families with three or more children, and families with young children. With respect to the last group, sampling was relatively easy, because even though Germany does not have a central registry, local registries exist and provide sample draws based on certain characteristics, among them year of birth. Hence a sample of individuals born between 2007 and March 2010 was drawn in 160 sample points, which were stratified by state (“Bundesland”), administrative region (“Regierungsbezirk”) and a region’s population size (“BIK-Regionen”).

Due to a low expected response rate of households with a migration background, a decision to include more households of this group than representative for the total population was made early on. Migration households in the register sample were identified in two ways: on the one hand, the registries provide information on the nationality of a person in the register, which was taken to classify non-German households as migrant households. On the other hand, using only nationality would omit all cases who obtained the German nationality later or are immigrants in the second-generation with a German nationality. Thus a second way to identify households with a migration household was employing an onomastic procedure, which basically assigns a linguistic and regional origin to each address based on the person’s surname. Among the drawn cases, an oversampling of immigrant households was conducted, such that the percentage of migrants (identified via the register and onomastic procedure) in the sample was doubled for each sampling point. Then the actual face-to-face interviews were conducted. As the sample is characterized by children born in the cohorts from 2007 to 2010, this sample is also referred to as the “Cohort Sample”.

For single parents, low income families and families with three or more children there is no sampling frame in Germany. Even though single parents and large families could theoretically be identified by the local registries, data protection rules prohibit such a combination of individual records into household-specific information. Hence sampling had to be conducted using a screening process, for which the starting sample was provided by TNS Infratest Sozialforschung from omnibus studies conducted every month in a representatively drawn sample. Households participating in such an omnibus study are always asked whether they would respond to a future survey. These households pose the gross sample for the screening process for FiD, where households were asked in telephone interviews about

³ All details on the sampling procedure are available in the documentation by TNS Infratest (Jänsch et al., 2011).

their household composition (to identify single parents and large families) and their household income (to identify low income households). Households of the target population were identified and categorized according to the following criteria:⁴

- *Low income* if the household had a monthly income of less than
 - 2500 Euro, when composed of at least two adults and at least two children
 - 2000 Euro, when composed of at least two adults and one child
 - 1500 Euro, when composed of one adult and at least one child.
- *Single parent* if the household is composed of at most one adult and at least one child.
- *Large family* if the household includes three or more children.

Households meeting any of the three characteristics were asked if they were willing to participate in the FiD study. A positive answer to the participation question was then followed by an invitation to participate in the study and a visit by an interviewer (for details on the selection process and case numbers, see section 4.2). Due to the selection process, these cases are also referred to as the “Screening Sample 2010”.

In the initial phase of the project, it was uncertain whether enough single parent and large family households could be acquired through the screening process. Hence it was planned from the beginning to repeat the screening process in 2011 for an additional sample of single parent and large families. These cases are also referred to as the “Screening Sample 2011”.

3.2. Structure and Contents

In large parts, FiD resembles the SOEP. The same basic concept of questionnaire types is used, i.e.

- there are household questionnaires for the household head (who is defined as the person most suited to answer financial questions in the household);
- each adult person (i.e. those turning 18 or older during the survey year) is asked to answer a personal questionnaire, which, in the first two years includes retrospective questions on childhood, education, and early work experiences;
- each person turning 17 during the survey year receives a “youth” questionnaire; and

⁴ To keep the screening process simple, the definition of “children” and “adults” had to be strictly on the basis of age. Children are thus all those individuals in the household who – at the beginning of the survey year – were at most 17 years old. Adults were then all others, i.e. those who at the beginning of the survey year were at least 18 years old. The actual family relationships were not relevant here, however, most of these households also are families with parent-child relationships. When determining the eligibility of households later, we kept the same definitions.

- for children in certain ages (namely those 0-1, 1-2, 2-3, 5-6, 7-8, or 9-10 years old), their parents are asked to fill out so-called parent questionnaires, which are slightly more elaborate than the similar mother-child questionnaires known in the SOEP.

For the most part, the contents of the FiD study are very similar to the SOEP, i.e. basic information on the household and each person is asked, including education, past and current labour market experiences, earnings and income, housing characteristics, health, some preferences and life satisfaction in general and for specific aspects. In addition, there is more focus on children and partnership: FiD includes a detailed partnership module, which retrospectively asks for marriages and partnerships lasting longer than six months.⁵ Compared to the SOEP, men and women are asked about their biological children in slightly more detail, including information about the partner's location and the marital status at the time of birth. Also, some aspects of child care at the work place are covered.

Completely new in FiD are questionnaires for the 1-2 year-olds, and the 9-10 year-olds, which previously did not exist in the SOEP (as of 2012, the SOEP added a questionnaire for the 9-10 year-olds, which is partly comparable to the FiD-version). Each of the questionnaires includes a module on child care, which, as the panel grows older, allows comparing child care decisions for one child over time. In these sections, parents are asked to specify the reasons for or against using day care, and in the case they use care they are asked about their satisfaction with it on different dimensions. Also covered are more detailed "outcome measures" to capture the skill development of children. Overall, the additional questions are designed to be comparable across the different parental questionnaires.

Table 2 provides a tabular overview of the main contents in FiD's Person Questionnaire and their comparability to the respective SOEP core questionnaires over all waves (2010-2013).⁶ As the table shows, there is a large overlap in the questionnaires, which makes the joint use of FiD and SOEP easy.

[Table 2 about here]

3.3. Interviewing

All personal interviews in FiD are conducted in a face-to-face mode using computer assisted personal interviews (CAPI). The only exceptions are the parent questionnaires, which can either be conducted on the interviewer's laptop or with pen and paper by the respondent

⁵ A similar module is now integrated in the new SOEP-Samples J and K in CAPI mode.

⁶ Note that this table can only provide an abstract overview. For details of the questionnaires, please refer to the documentations online, at <http://panel.gsoep.de/soepinfo2011/> (for the SOEP) and <http://www.diw.de/fid-soep> (for FiD).

herself. Future mode switches to pen and paper interviews (PAPI) are not feasible in later waves of FiD, because the questionnaire routing depends to some extent on the technical possibilities a computer offers. Using only CAPI interviews promises some benefits for data quality and lowers the amount of time necessary to test and verify the data. On the other hand, it limits the possibilities of obtaining data from reluctant respondents – the SOEP, for example, allows for interviews via mail (and thus PAPI mode) if at the end of the fieldwork period all other attempts have failed to convince a household to participate in the study. With the integration into the SOEP, future mode switches cannot be ruled out.

To thank the respondents for participation in the study, FiD could implement an incentive scheme that was especially targeted at families. € are paid for a completed household interview and the first personal interview. Each additional completed individual interview is rewarded with an additional €. If all questionnaires for eligible persons in the household are complete, there is an additional premium of € for each child in the household. In addition, as the sample consists of households with children, special panel care measures can be taken: each household receives a gift for Children's Day on November 20th, where, depending on the age of the child, balloons, washcloths, bibs, reflectors, pencils or similar small gifts are included with a thank you letter.

3.4. Representativity and weighting

The Cohort Sample drawn in 2010 is representative of the population of families in Germany with children born between January 2007 and March 2010. The sampling weights for this sample are constructed relatively easy, as for each household the sampling probability is known through the design of the survey. These design weights are first adjusted for the initial non-response due to refusal or inability to participate of eligible households. In a second step, they are then calibrated using a raking approach with the margins of the most important variables known for the German population from the Mikrozensus. In the waves after 2010, the sample can be regarded as being representative of the same population, although there is a small bias: some households may actually lose their “eligibility” of being a household with children born between 2007 and 2010, because the children may no longer be in the household. Similarly, the sample does not capture those households, who after 2010 include a child born in those years (e.g. through a moved-in partner with a child). However, these fluctuations in and out of the sample population are very small: of the 3,100 households participating in all three waves, only ten (0.3%) would have to be removed from the cohorts 2007-2010. This relationship will be stable at least for the initial years of the sample.

The case is slightly different for the Screening Samples. Here, the sample drawn in 2010 is representative of the population of families in Germany, which are low-income families, large families or single parent families in 2010. While margins for this population in 2010 exist, the initial design weights have to be estimated, as the sampling probabilities are not known for the screening process. Compared to the Cohort Samples, this introduces some uncertainty, although the calibration can then be done in a similar way. The challenges increase in the following years: due to the fluctuations in and out of the three screening groups in every year, this sample does not represent the same three groups in the German population in 2011, or any of the following years. 44% of the households identified as low income move out of this group in at least one of the following years. In terms of their sample characteristics, things look much better for the other two groups: almost 80% of the single parent families remain in this state over the three years. The large family characteristic is even more stable: about 11% leave it within the first two years of the panel. Strictly speaking, the Screening-Samples should thus be seen as representative of the respective populations in 2010 and 2011, and then can be used to monitor the changes in these populations over time. However, as fluctuations at least in family compositions are not huge, the sampled groups remain close to the targeted groups at least in the first years.

Given these difficulties, constructing cross sectional sampling weights for FiD alone (i.e. the joint Cohort- and Screening-Samples without the SOEP-Samples) in 2011 is not trivial, but necessary as useful analyses also for future waves are possible only with these sampling weights as a basis. Because our approach to this problem is non-standard, we go into more detail here.

The starting point is the integration of the FiD households into the regular SOEP. This integration is achieved by treating the FiD cases as any other new sample in the SOEP, which would be integrated by including the old cases with their previous sampling weights (adjusted for attrition) and the new cases with their design weights (adjusted for the initial non-response). After an adjustment according to the number of observations in each group, the calibration follows a raking approach using margins from the general population (information from the “Mikrozensus”). The weights from this step are the building block for the cross-sectional weights in the FiD population in 2011, which consists of the following possible types of household characteristics:

- low income household in 2010 (from Screening 2010)
- large family household in 2010 or 2011 (Screening 2010 or 2011)
- single parent household in 2010 or 2011 (Screening 2010 or 2011)

- cohort household 2007-2010 in survey year 2010

In principle, the integrated weights for the joint FiD and SOEP samples provide the weights of this population. To calculate weights for the FiD population alone, the SOEP cases have to be removed, and their removal has to be adjusted for in the weights. However, the population described by the four characteristics above is not easily determined for all households available in 2011: For those households not present in 2010, the likelihood of being in one of the groups has to be estimated. This concerns new cases from 2011 (Screening 2011 from FiD as well as the new SOEP sample “J”) and cases in the SOEP, which did not participate in 2010 but returned in 2011. Of the overall 16,819 cases in the joint FiD-SOEP population, this concerns 4,226 households (Screening 2011: 915 households; Sample J: 3,136 households; temporary dropouts SOEP: 175 cases). For these cases it is assumed that if they have children in the four cohort years in 2011, they also had them in 2010. For the other three categories, the status for 2010 is unknown. Taking the cases which are available in both years, we estimate a logit model that predicts the likelihood of having the characteristics in 2010. We then use an out of sample prediction to categorize those cases only observed in 2011.

This procedure allows us to categorize all households in the joint FiD-SOEP sample into the four groups above, which also means that with the joint sampling weights we are able to produce the population equivalent of the above groups. To calculate the weights for the FiD population alone, the SOEP cases in this group need to be removed, and the FiD cases need to be scaled up to still remain at the same population total. While the simplest way would be to multiply each FiD weighting factor with the inverse of the fraction of SOEP cases in the total, this approach would ignore any systematic differences between the SOEP and the FiD cases – which, given the sampling design, are sure to exist. Instead we use an approach similar to a regular estimation of drop-out probabilities for an attrition analysis (see for example Kroh, 2011): By estimating the likelihood in the joint SOEP-FiD population of belonging to the FiD sample only, we obtain – by using the inverse of this probability – a household-specific factor that is multiplied with the integrated weights. The sum of these newly achieved factors leads to an estimate of the population similar to the one derived for the joint SOEP-FiD cases.

These cross-sectional weights for 2011 provide the starting point for the future cross-sectional weights as well. Those will not be calculated with the above method, because estimating the population characteristics in 2010 will become less precise with time. Instead, the staying probabilities can be used to arrive at the cross-sectional weights for 2012 and beyond, because no new samples will be integrated in FiD after 2011.

3.5. Datasets

The datasets provided in FiD resemble very closely the respective data in the SOEP. Hence FiD also reproduces the general structure users know from the SOEP. There are the basic datafiles such as *ppfad* and *hpfad*, with which the user can monitor the development of each person and household through the panel life, along with some generated information. The *\$pbrutto* and *\$hbrutto* files provide similarly important information about the interviewing process for each wave. *hbrutt10_fid* and *hbrutt11_fid* contain the gross sample with which the survey started, i.e. the Cohort and Screening Gross Sample in 2010 and the Screening Gross Sample in 2011.⁷ Identical to the SOEP, FiD distributes original data files, i.e. those which contain the unaltered data from the questionnaires directly (except for answers to open questions). In resemblance to the SOEP, these files have wave identifiers – „f10“ for FiD in survey year 2010, „f11“ for survey year 2011, and so on. Along with the original data files, FiD also produces the main generated data files – such as *\$pgen*, *\$hgen* – from the SOEP. In addition, spell files – like *artkalen* or *pbiospe* – are also available. A large part of the distribution is “biographical” data, which – due to the very nature of FiD – is sometimes more extensive than in the SOEP. For example, the *bioage* files, which contain detailed information from the parent questionnaires about their children, contain many more variables than their counterparts in the SOEP. However, FiD and SOEP use the same naming conventions in this case to make the joint use of the data easier. Some datasets known from the SOEP are not provided, mainly because the respective information has not been asked (yet). Table 3 shows a list of all datasets available in the FiD distribution 3.0.

[Table 3 about here]

4. Survey Quality

4.1. Initial Response Rates

Due to the different sampling procedures, the initial non-response in FiD has to be evaluated separated by the three samples. The Cohort Samples are drawn from the register based on the children’s birth years in the households. The gross sample was thus known in advance, which makes the calculation of the response rates easier. In the following, we split the samples by cohort year (2007-2010) and also look separately at those cases that had been identified as

⁷ Note that for the screening samples, only those households agreeing to participate are included in these datasets.

migrant households through the registry and via the onomastic procedure (see section 3.1). Overall, 5,200 eligible cases were drawn from the registers, of which 46% were identified to have a migration background. A total of 2,074 households (40% of the gross sample) participated in the study. Table 4 shows the response rates according to cohorts and migration background.

[Table 4 about here]

German households have a significantly higher likelihood to participate in the study than those with a migration background. While there are almost no differences among the German cohorts, the rate of participation for migrants in the cohort of 2010 is significantly higher (on the 1% level) than the participation in the other three cohorts. Overall, the participation rate is rather high compared to the SOEP's new sample "J" in 2011, although this only holds for the German households: participation was at 33% overall; 34% for German households and 26% for those with a migration background (determined only via the onomastic procedure).

The calculation of response rates is not as straightforward for the Screening Samples. Here, a five-step procedure was followed: from the omnibus studies conducted by TNS Infratest Sozialforschung, a starting sample of households eligible for screening was constructed. For all of these households, contact attempts were made by telephone to screen whether they belong to the target population of low-income households, large families, or single parents. If a household could be categorized in one (or more) of the three groups, they were asked whether they would participate in the main study. In each of these steps, a failure is possible, which reduces the sample accordingly. Table 5 shows the different steps, the initial gross samples and the finally achieved net sample for both Screening-Samples.

[Table 5 about here]

The table shows that there is a large group of households for which it is unclear whether they belong to the target group or not (50% in 2010, 43% in 2011). Because of this, the actual response rate cannot be finally calculated. On the other hand, if a household is known to belong to the target group, the participation rate is rather high, with more than 2/3 of the sample participating in the actual study. Readers may notice the large difference in belonging to the target group in 2010, where 22% were not in the target group, compared to the 42% not

in the target group in 2011. This may be the result from not including low-income households in 2011. However, given a household belongs to the target group, participation rates are almost identical.

4.2. Retention Rates

One of the most crucial aspects of a panel dataset is the longitudinal stability as measured by the fraction of households that remains from the previous year. Table 6 shows the retention rates for all samples, where also the initial drawing characteristics are shown (i.e. cohort year and screening characteristics). The initial sample is reduced by those cases that turned out to be ineligible after the survey data were evaluated with respect to the characteristics of the target population. This slightly reduces the Screening Samples, mainly because households turn out not to be in the low income population. Neither new households (i.e. split-offs from old households) nor households with a temporary drop-out are considered in this table. Accordingly, the total number of interviewed households is slightly higher in 2011 and 2012 than shown in this table.

[Table 6 about here]

Due to the sampling procedure used for the Screening Samples the higher retention rates in these groups are not surprising. For the Cohort Sample it is encouraging to observe an increase in the retention rates from 2011 (79%) to 2012 (83%), given that this sample is extremely valuable due to the sampling design. It is also interesting to see that for the first wave retention rates, there are differences between the screening groups: e.g. being a low income households is associated with low retention in general; in fact, the lowest overall retention rate is observed among those households with low income in combination with having at least 3 children (77%).⁸ In 2012, these differences among the groups are not visible any more – all groups are at retention rates around 90%.

Table 7 now follows with a view on the individual longitudinal stability. As was the case for households, we do not consider individuals who have dropped out temporarily – however, movers are considered in this case as long as they belong to the initial sample. For readability reasons, the distinction between the different screening groups is not kept up here.

⁸ The group of low-income, single parent families with three or more children is too small to say anything valid about their retention rate.

[Table 7 about here]

As would be expected, the individual retention rates are slightly lower than those for the households. Again, the Cohort Sample shows a lower retention rate by about five percentage points in both years. Still the individual retention increases for both samples from 2010, such that it seems likely that the retention in the future will remain high.

4.3. *Partial Unit Non-Response (PUNR)*

Similar to the SOEP, FiD uses the concept of “household eligibility”: any member of an eligible household is part of the gross sample and needs to be interviewed if she turns 18 or older during the survey year. Hence the completeness within the household is important, i.e. whether all interviews that were supposed to be done were actually conducted. Because of the large number of questionnaires, also about children living in the household, FiD used incentives also targeted at the completion of households (see section 0). In Table 8, we show the fractions of complete households split up by sample. The last two lines in Table 8 show the fractions of complete households in terms of adult questionnaires and children questionnaires.

[Table 8 about here]

The numbers in Table 8 show a positive development similar to many longitudinal studies. While the partial unit non-response is rather high in the beginning, it is generally reduced in the following waves. The rate of completed parent questionnaires is extremely high, with over 99% of eligible questionnaires completed in 2011 and 2012. This is even more remarkable considering that FiD does not allow for a mode switch after the field work to retrieve respondents who would otherwise not participate (see also section 0).

4.4. *Item Non-response*

During an interview, the respondents have the possibility to refuse an answer. The fewer missing values there are in an interview the better the quality of that interview, because the analysing researcher does not have to deal with large amounts of missing data. Because the total number of questions depends on the individual interview and the filtering (i.e. not all persons are exposed to all questions even within the same questionnaire), it is useful to report the percentages of missing answers instead of the actual number of missing values. To some

extent, the percentages also allow the comparison across different questionnaires and years. Since the distribution of missing answers is heavily skewed (to the right), we show the 90th percentile of missing answers in Table 9, rather than the mean (which overall is at around 1.2%) or the median (which is slightly above zero).

[Table 9 about here]

Overall, Table 9 confirms the rather high data quality in FiD: overall, 90% of the interviews in FiD contain less than 3% of missing answers. There is also a slight decrease from 2010 to 2012, which is likely due to respondents becoming used to the questions. The numbers are slightly higher for the parent questionnaires. One possible reason for this finding may be that these questionnaires are conducted mainly in PAPI mode, which is also more likely to be filled out in self completion rather than with the help of the interviewer. Without this help, respondents may find it harder to answer all questions and thus may truly not know the answer. The very high percentage in the gap questionnaires for those with a temporary absence in FiD (10% of all cases have 20% or more missing answers) may be the result of different factors: a) it is also a PAPI questionnaire, b) the period asked about is more than a year ago, thus there may be some recall error, and c) these respondents may be also the most reluctant, and hence might more easily refuse an answer. In addition, the number of questions asked is rather small (15 questions on average), such that a missing answer has a larger impact.

5. Data Access

The data from “Familien in Deutschland” are currently (May 2013) accessible for the scientific community similar to the SOEP data. Interested researchers can apply for the data usage at the SOEP-group by filling out a two-page form, including a short description of the research proposal. Following an evaluation of the application, a contract between the researcher and the SOEP needs to be signed before the data are made available to the new user, currently via one-time downloads. For further information on the application process (and all possible changes to it) please consider the information at www.diw.de/fid-soep.

6. Summary and Outlook

The FiD data collection effort has become a success story for the SOEP group at DIW Berlin. An entirely new and relatively large sample was drawn and proved to be of similar

longitudinal stability as the regular SOEP samples. With respect to different quality indicators (e.g. item non-response, partial unit non-response) the FiD data adhere to the high standards set by the SOEP. Several new and extended questionnaires were implemented within a very short period of time, which provide new and improved data on children. As such, FiD allows for more in-depth analyses of families and children.

In this sense it is only fitting that after FiD was financed by the ministries for three years and a fourth wave was commissioned by the BMFSFJ to continue a full data collection in 2013, the FiD-samples will be integrated into the main SOEP for the data collection of 2014. The data distribution of 2015 will then contain for the first time the complete set of SOEP and FiD cases, with identical variable names and datasets also for those years where FiD was collected in parallel to the SOEP.

References

- Beninger, D., H. Bonin, M. Clauss, H. Hofmann, J. Horstschräer, S. Munz, C. K. Spieß, M. Weding und K. Wrohlich (2008): Machbarkeitsstudie zur stufenweisen Gesamtevaluation des Gesamtableaus ehe- und familienbezogener Leistungen in Deutschland. Studie im Auftrag der Prognos AG, 2008.
- Huinink, J., J. Brüderl, B. Nauck, S. Walper, L. Castiglioni & M. Feldhaus (2011): Panel Analysis of Intimate Relationships and Family Dynamics (*pairfam*): Conceptual framework and design. *Zeitschrift für Familienforschung*, Jg 23., Heft 1, S. 77-100.
- Jänsch, A., S. Huber, N. A. Siegel, S. Stimmel, & D. Geue (2011): „Familien in Deutschland“ (FiD) 2010 Methodenbericht: Anlage und Ergebnisse der FiD-Stichproben. TNS Infratest, München. (Available at www.diw.de/fid-soep, “FiD-Dokumentation”.)
- Kroh, M. (2011). Documentation of Sample Sizes and Panel Attrition in the German Socio Economic Panel (SOEP) (1984 until 2010). DIW Data Documentation 59.
- Wagner, G. G., J. R. Frick & J. Schupp (2007). The German Socio-Economic Panel Study (SOEP) - Scope, Evolution and Enhancements. *Schmollers Jahrbuch*, Jg. 127, Heft 1, S. 139-169.

Table 1: Sample Sizes and Conducted Interviews by Questionnaire, 2010-2012

	2010	2011	2012
Interviews			
Household Questionnaire	4,574	4,529	4,186
Person Questionnaire (17+ year-olds)	7,807	7,648	7,165
Youth Questionnaire (16-17 year-olds)	190	262	293
Parent Questionnaire 1 (0-1 year-olds)	1,321	207	212
Parent Questionnaire 2 (1-2 year-olds)	787	644	568
Parent Questionnaire 3 (2-3 year-olds)	871	740	555
Parent Questionnaire 4 (5-6 year-olds)	473	486	424
Parent Questionnaire 5 ^a (7-8 year-olds)	425	527	501
Parent Questionnaire 6 ^a (9-10 year-olds)	404	510	475
Gap (Luecke) Questionnaire ^b			227
Totals			
Persons	17,002	17,129	15,850
Adults 17+	8,301	8,052	7,630
Youth (16-17)	190	262	293
Children (0-16)	8,511	8,815	7,927

Source: FiDv3.0

Note that these are wave specific net samples, for the longitudinal samples see Table 6 in section 4.2. The numbers in this table are restricted to those households with a completed household interview. There are a few households where only a person interview was conducted (these data are available in the regular distribution).

^a Number of cases with at least one interview. Parent Questionnaires 5 and 6 are answered by both mother and father if applicable, such that two observations exist for many children in these age groups. The actual number of interviews is thus larger than the sample sizes given here.

^b The Gap Questionnaire is listed here in "2012", as it was filled out in this year. However, information was gathered for the previous year (2011).

Table 2: Topics in FiD (Person Questionnaire) over the Years and SOEP-Comparability

Topic	FiD Contents				SOEP
	2010	2011	2012	2013	
<i>Contemporaneous^a</i>					
Labour market (Employment, unemployment, tenure, overtime, commute, job mobility, classification, firm size, gross/net wages)	X	X	X	X	X
Income (current and last year's income)	X	X	X	X	X
Event History Calendar (last year and life)	X	X	X	X	X
Health (current health status, sleep, disability, doctor visits, hospital stays, work absence)	X	X	X	X	X
Health Insurance	X	X ^c	X ^c	X ^c	X
Political participation	X	-	-	-	X
Satisfaction with life and specific areas	X	X	X	X	X
Networks		X	X ^c		X
Wealth			X		X
BIG 5 (Personality Traits)			X		X ^d
Effort Reward Imbalance			X		X ^e
Firm specific family support	X	X	X	X	-
Attitudes towards family life	X	X ^f	X	-	-
Fertility decisions	X	X ^f	X	-	-
<i>Retrospective^b</i>					
Education	X	X	X	X	X
Fertility	X	X	X	X	X
Migration Background	X	X	X	X	X
Partnership History	X	X	X	X	X
Job History		X	X	X	X
Parental Background		X	X	X	X

Notes

The SOEP questionnaire was the basis for the FiD contents. For this reason, the "x" in the SOEP column indicates in general that the respective content is available for the same year(s) in FiD and SOEP.

^a Contemporaneous questions are those which are asked on a yearly basis

^b Retrospective questions are only asked once per respondent. FiD uses a two-part biography module, where the first four topics are asked first.

^c Module is asked only partly.

^d The BIG 5 module was asked in 2009 and 2013 in the SOEP.

^e Effort Reward Imbalance was asked in 2011 in the SOEP.

^f Only asked for those new to FiD.

Table 3: FiD Datasets Similar to the SOEP Distributions

Wave specific data	Data across all waves
\$p, \$h,\$kind,\$hbrutto, \$pbrutto, \$hgen, \$pgen, \$lela, \$pkal, \$eltern1, \$eltern2, \$eltern3, \$eltern4, \$eltern5, \$eltern6, \$paradata, hbrutt10_fid, hbrutt11_fid, \$luecke	ppfad, hpfad, bioage01,bioage02, bioage03, bioage06, bioage08p1/bioage08p2, bioage10p1/bioage10p2, bioage1, biomars, biocouply, biobirth, artkalen, pbiospe, biojob, phrf/phrf_soep, hhrf/hhrf_soep

Table 4: Initial Response Rates in the Cohort Sample (Survey Year 2010)

	Cohort				Total
	2007	2008	2009	2010	
Total	37.65	40.04	36.90	45.98	39.88
German	50.34	50.40	49.44	51.90	50.49
Migration Background	23.33	27.10	22.95	37.82	27.10

Source: unweighted data, FiDv3.0.

The numbers in this table show the percentage of households in each group with at least one interview based on the number of households in the gross sample. The gross sample was adjusted for ineligible households (e.g. due to death, move out of the country, or failure to work a complete sample point) and then amounts to N=5,200 cases.

Table 5: Initial Response Rates in the Screening Samples (Survey Year 2010 and 2011)

	2010			2011		
	N	%	% Target	N	%	% Target
Total Gross Sample	13,653	100		8,400	100	
No contact	3,257	23.9		1,939	23.1	
Contact, but no cooperation	3,740	27.4		1,634	19.5	
Not in target group	2,955	21.6		3,515	41.8	
In target group, refused to participate	440	3.2	11.9	158	1.9	12.0
In target group, willing to participate	3,261	23.9	88.1	1,154	13.7	88.0
Net sample (participating households)	2,500	18.3	67.5	924	11.0	70.4

Source: unweighted data, FiDv3.0.

The numbers in this table show the totals and the respective fractions during the screening process for 2010 and 2011. The “N” column shows the total, the “%” column shows the percentage of the total in each step. The “% Target” column depicts the percentages of those cases which were identified to be eligible for the study. Note that “Net sample” is a subset of the category “In target group, willing to participate”.

Table 6: Household Longitudinal Observations and Retention Rates

	2010 N	2011 N	2012 N	2011 Retention Rate	2012 Retention Rate
Total (Screening 2010 + Cohort)	4,337	3,579	3,100	0.83	0.87
Total (Screening 2010, 2011, Cohort)		4,494	3,895		0.87
Screening 2010	2,263	1,938	1,734	0.86	0.89
Low income (LI)	636	535	474	0.84	0.89
Single parents (SP)	444	395	350	0.89	0.89
Large families (LF)	367	335	303	0.91	0.90
LI+SP	494	415	373	0.84	0.90
LI+LF	225	174	154	0.77	0.89
SP+LF	77	71	67	0.92	0.94
LI+SP+LF	20	13	13	0.65	1.00
Cohort	2,074	1,641	1,366	0.79	0.83
2007	515	404	340	0.78	0.84
2008	535	418	357	0.78	0.85
2009	503	404	324	0.80	0.80
2010	521	415	345	0.80	0.83
Screening 2011		915	795		0.87
Single parents (SP)		408	346		0.85
Large families (LF)		466	412		0.88
SP+LF		41	37		0.90

Source: unweighted data, FiDv3.0

This table only considers households that were eligible in their first wave. Households in the Screening Samples, which were found not to belong to any of the three groups, are excluded in this table, because they were not interviewed in the following waves.

Table 7: Individual Longitudinal Observations and Retention Rates

	2010 N	2011 N	2012 N	2011 Retention Rate (%)	2012 Retention Rate (%)
Total (Screening 2010 + Cohort)	7,501	6,046	5,160	0.81	0.85
Total (Screening 2010, 2011 + Cohort)		7,533	6,424		0.85
Screening 2010	3,731	3,112	2,722	0.83	0.87
Cohort	3,770	2,934	2,438	0.78	0.83
Screening 2011		1,487	1,264		0.85

Source: unweighted data, FiDv3.0

This table only considers individuals in households that were eligible in their first wave. Individuals from households in the Screening Samples, which were found not to belong to any of the three groups, are excluded in this table, because they were not interviewed in the following waves..

Table 8: Household Completion Rates by Sample and Questionnaire Type in %

	2010 %	2011 %	2012 %
Total	87.8	91.9	93.2
Screening 2010	89.7	93.1	92.8
Cohorts	85.6	90.3	94.2
Screening 2011		92.5	92.4
Completed all Adult Questionnaires	90.3	92.7	94.0
Completed all Parent Questionnaires	97.1	99.4	99.4

Source: unweighted data, FiDv3.0

This table considers all households ever interviewed. Shown are the percentages of completed questionnaires in a household based on the number of all eligible questionnaires in that household.

Table 9: Item Non-Response (90th Percentile) by Questionnaire Type and Survey Year

	% missing		
	2010	2011	2012
Household Questionnaire	2.91	2.68	2.68
Person Questionnaire	2.42	2.34	2.77
Youth Questionnaire	1.51	2.90	3.18
Parent Questionnaire 1	3.57	7.14	3.23
Parent Questionnaire 2	5.36	4.90	2.94
Parent Questionnaire 3	5.22	4.72	3.88
Parent Questionnaire 4	4.40	4.04	3.19
Parent Questionnaire 5	4.59	3.70	2.80
Parent Questionnaire 6	6.45	7.28	4.27
Gap Questionnaire			20.00
Total	3.23	2.94	2.88

Source: unweighted data, FiDv3.0

This table considers all interviews ever conducted. The numbers show the 90th percentile of the missing value distribution, i.e. 90% of respondents answering the respective questionnaire have at most the shown percentage of missing answers (due to a refusal, inability to answer or implausible answers). The numbers take into account that respondents are asked different sets of questions depending on the filtering of the questionnaire.