



SONDERFORSCHUNGSBEREICH 504

Rationalitätskonzepte,
Entscheidungsverhalten und
ökonomische Modellierung

No. 07-08

**The German SAVE survey: documentation and
methodology**

Daniel Schunk*

February 2007

Financial support from the Deutsche Forschungsgemeinschaft, SFB 504, at the University of Mannheim, is gratefully acknowledged.

*Sonderforschungsbereich 504, email: dschunk@uni-mannheim.de



Universität Mannheim
L 13,15
68131 Mannheim

The German SAVE Survey 2001 - 2006. Documentation and Methodology

– February 2007 –

Daniel Schunk

Mannheim Research Institute for the Economics of Aging (MEA)

University of Mannheim

L 13, 17

68131 Mannheim

E-mail: save@mea.uni-mannheim.de

Technical discussion paper

The purpose of this document is to describe methodological details of the German SAVE survey and to provide users of SAVE with all necessary information for working with the publicly available SAVE dataset.

Contents:

1. Introduction
 2. Structure of the Questionnaire
 3. Survey Design and Methodology
 4. Item Nonresponse and Imputation
 5. Design of Weights
- Appendix 1: Questionnaire and Dataset
Appendix 2: Documentation of all Variables

Acknowledgements:

SAVE involves major efforts in survey design, field work, data preparation, and user support. Many researchers have been actively involved in these efforts, in particular Axel Börsch-Supan, Lothar Essig, Anette Reil-Held, and Joachim Winter. Numerous research assistants have spent days and nights on the SAVE dataset, I owe them many thanks for their tedious work.

Financial Support from the German Research Foundation (DFG) via the Sonderforschungsbereich 504 at the University of Mannheim is vital for the SAVE survey and is gratefully acknowledged.

1. Introduction

Understanding savings and investment behavior of German households is the main goal of the SAVE study. Germany is an interesting country to study savings behavior since nearly everyone – whether in the middle income bracket or richer - saves substantial amounts, even in old age.

This paper documents and describes the SAVE survey, the structure of the questionnaire, the survey design and the structure of all samples, as well as nonresponse and weighting. The purpose of this document is to provide potential users of SAVE with all necessary information for working with the SAVE data. The document draws from various sources, in particular Börsch-Supan and Essig (2005), Essig (2005), Heien and Kortmann (2003), and Heien and Kortmann (2005). It might be worth looking at these documents for further information.

The data situation for analyzing household financial behavior has been very limited in Germany. There has been no dataset available that records detailed data on both financial variables such as income, savings, and asset holdings, as well as sociological and psychological characteristics of households. The German Socio-Economic Panel (*German SOEP*) contains rich data on household behavior and records indicators of saving and asset choices, but it does not cover the quantitative composition of households' assets or any change in the amount of wealth in very much detail, though. The situation is similar in another representative survey (*Soll und Haben*). This study records detailed data on the composition of various financial assets, but it only has qualitative indicators and does not quantify asset holdings. Finally, the official budget and expenditure survey (*Einkommens- und Verbrauchsstichprobe, EVS*), conducted every five years by the Federal Statistical Office, has very detailed information on the amount and composition of income, expenditure, and wealth, but information on other household characteristics is very limited. Taking the Dutch CentER Panel and the U.S. Health and Retirement Study (HRS) as a basis, researchers of the University of Mannheim have cooperated with the Mannheim Center for Surveys, Methods and Analyses (ZUMA), NFO Infratest (Munich), Psychonomics (Cologne) and Sinus (Heidelberg) to produce a questionnaire on households' saving and asset choice; see Börsch-Supan and Essig (2005) for more detailed information.

With the data collected in the SAVE survey, we intend to have a basis for obtaining a better understanding of German households' saving behavior. Our present poor understanding of saving behavior has far-reaching consequences for economic policy. For instance, we do not understand well, to what extent saving must be encouraged so that enough savings are

available for financing the investment that forms the basis for long-term growth of our economy. Payments towards a saving scheme may either increase savings because of the higher return on the funds saved or – if the household has a specific target in mind – decrease them because the state bears a certain section of the costs.

A particular case in point is retirement saving and its role in pension reform. In fact, we do not have a reliable empirical basis on which to assess the success of the recent German pension reform named after the then labor secretary Walter Riester in creating additional saving. Will such saving exactly compensate for the reductions in pay-as-you-go pensions? Or will substitution be less than perfect? Will the new retirement saving simply displace other saving, i.e. will the increase in savings made in life insurances and pension funds coincide with a reduction of saving e.g. in homeownership and real estate?

So far, we do not have good answers to these questions, and one purpose of the SAVE panel is to shed light on them during an important transition period when the new multipillar pension system in Germany will slowly replace the monolithic pay-as-you-go pension system, in which 85% of retirement income was the state-provided pension.

Household savings decisions are the results of a very complex decision-making process. In order to better understand this process, a huge amount of data is needed, including information on household socio-demographic, financial, and psychological characteristics, in particular information on household preferences and future expectations. Containing a representative sample of German households and a wide range of characteristics from those fields, makes the SAVE data unique and particularly appropriate to help obtaining answers to questions in the field of savings behavior and public policy.

2. Structure of the Questionnaire

The questionnaire has been designed in such a way that the interview should not exceed 45 minutes and consists of six parts (see table 1).

The first, relatively short part explains the purpose of the study and describes the precautions that have been taken with respect to confidentiality and data protection. Part 2 lasts about 15 minutes and contains questions on the socio-economic structure of the household, including age, education and labor-force participation of the respondent and his or her spouse. Beginning in 2005, this part also contains questions about the health situation of the household. Part 3 of the questionnaire contains qualitative and simple quantitative questions on saving behavior and on how households deal with income and assets, including hypothetical choice tasks and questions on savings motives; questions on financial decision processes, rules of thumb, and attitudes towards consumption and money are also included. Part 4 is the critical part of the questionnaire. It contains a comprehensive financial review of the household and therefore the most sensitive questions in financial items such as income from various sources and holdings of various assets. Apart from financial assets, the questions also cover private and company pensions, ownership of property, business assets, and debt. Part 5 contains questions about psychological and social variables, e.g. the social environment, expectations about income, the economic situation, health, life expectancy, and general attitudes to life. The interview ends with open-ended questions about the interview situation and the question whether the respondent would be willing to participate in a similar survey in the future (part 6).

Table 1: Basic structure of the questionnaire of the SAVE Survey.

<i>Part A:</i>	Introduction, determining which person will be surveyed in the household
<i>Part B:</i>	Basic socio-economic data of the household; health questions (since 2005)
<i>Part C:</i>	Qualitative questions concerning saving behavior, income, and wealth
<i>Part D:</i>	Quantitative questions concerning income and wealth
<i>Part E:</i>	Psychological and social determinants of saving behavior
<i>Part F:</i>	Conclusion: Interview-situation

The SAVE questionnaire has undergone slight changes from year to year. There exists an Excel-spreadsheet that documents which variable was asked in which year. As well, the

questionnaires are available for each year. Please send an e-mail to save@mea.uni-mannheim.de for further information.

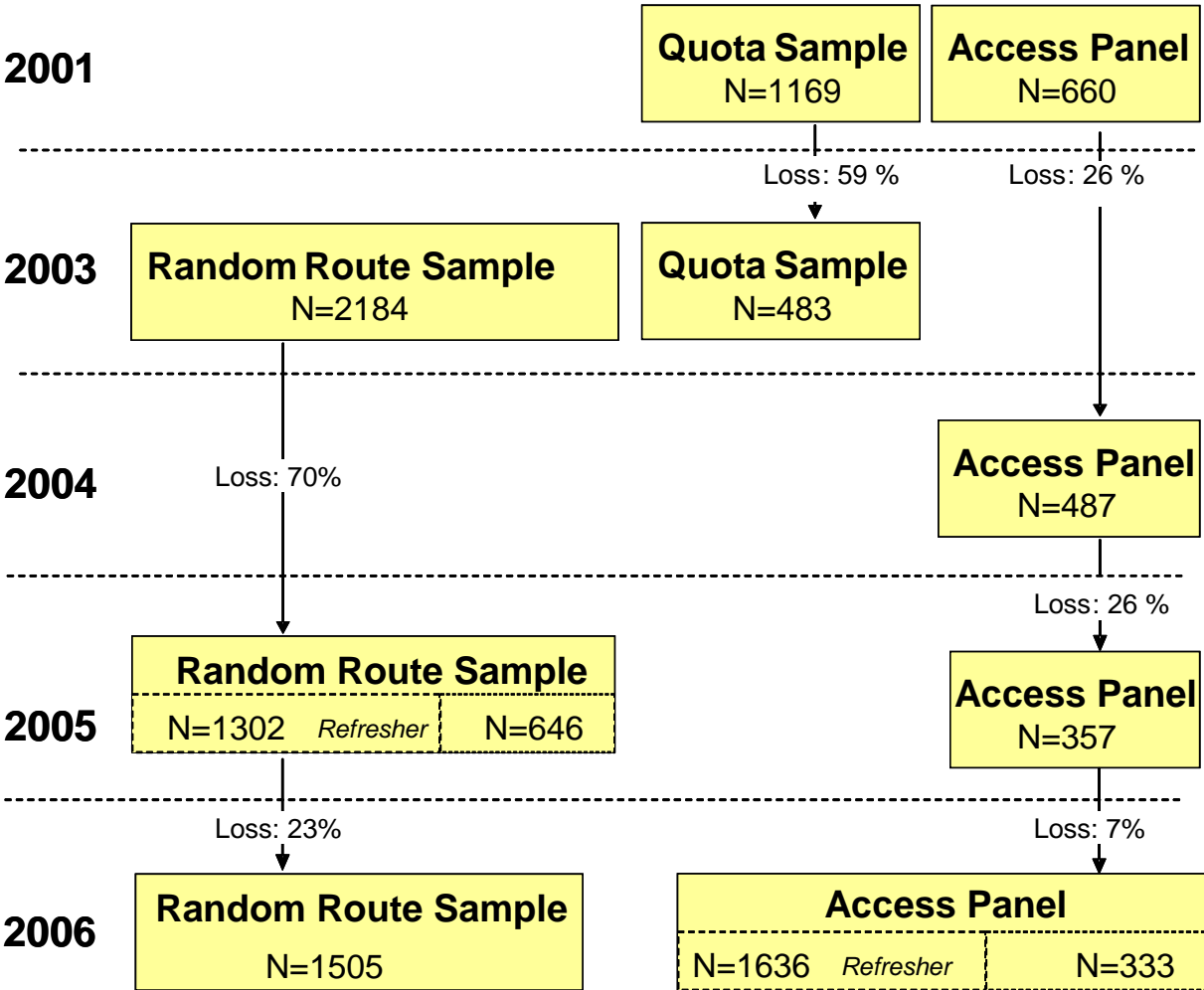
3. Survey Design

3.1 General Design of the SAVE Survey

Figure 1 presents the different waves of the SAVE survey. As one can see, SAVE consists of several different subsamples, which will be described in a later section.

As the questionnaire of the SAVE Survey includes topics that are unusual in surveys such as detailed questions about personal income and financial wealth, both respondents and interviewers have been specially prepared for the questioning. In particular, because of the sensitivity of the topic, we were interested in learning about interviewer modes and question modes. Therefore, the first SAVE wave (2001) additionally included an experimental component. This first wave was used to learn about interview methodology in the particular case of the sensitive financial questions of the SAVE study.

Figure 1: Waves of the SAVE survey



In all waves, a letter describing the aim of the study and the protection of privacy was handed out by the interviewers. The detailed questions about personal income and wealth were also non-standard and demanding for the interviewers. Interviewers for each wave were selected by TNS Infratest. TNS Infratest underwent an intensive interviewer selection and additional training and motivation. Additionally, from 2005 on, a short motivation video (featuring Prof. Börsch-Supan, Ph.D. and explaining the aim and the importance of the survey, as well as interview methodology) was sent to the interviewers so that they could watch it on their laptops before the questioning. Various incentives were handed out to the participants.

A detailed description of sampling scheme, questioning modes, and incentives is presented in the following subsections. In the data, the affiliation of an observation to a certain subsample is encoded by the variable “*wave*” (table 2).

Table 2: Encoding of the different waves in the SAVE data.

Subsample	Value of “wave”
2001 Access Panel	0
2001 Quota Sample	1
2003 Quota Sample	2
2003 Random Route Sample	3
2004 Access Panel	4
2005 Random Route Sample	5
2005 Access Panel	6
2006 Random Route Sample	7
2006 Access Panel	8

3.2 SAVE 2001

The surveys took place in early summer 2001. In this year, the fieldwork for the personal interviews took place between May 29 and June 26, 2001, whereas the fieldwork for the Access Panel took place between June 29 and July 24, 2001.

Experimental design of the SAVE 2001 survey

The first four versions were computer aided personal interviews (CAPI); they were carried out by NFO Infratest, Munich. In contrast, the fifth version was a conventional paper questionnaire ("paper and pencil", P&P). The CAPI interviews were carried out using quota samples whereas conventional P&P questionnaires were given to a so-called Access Panel

operated by the company TPI (Test Panel Institute, Wetzlar). The only difference in the four versions of the CAPI interview is in the critical part 4 of the questionnaire. In versions 1 and 2, all questions were administered by CAPI in the presence of the interviewer. The difference between these versions is that the questions on asset holdings were presented using an open-ended format with follow-up brackets (range cards) in version 1 and with 'forced' brackets in version 2. In the data, the affiliation to a certain version is indicated by the variable "*version*". Because many of these questions relate to intensely personal matters of income and wealth, there is another modification in versions 3 and 4. In these two versions, part 4 was not part of the personal CAPI interview, but left as a paper-and-pencil questionnaire by the interviewer (this mode is termed "P&P drop off" in the sequel). In version 3, the interviewer came back personally to collect the drop-off questionnaire; in version 4, the questionnaire had to be returned by mail using a pre-paid envelope. If this was not done within a specified number of days, the respondent was reminded by telephone several times. This helped increase response rates for the drop-off questionnaire, but nevertheless, they were significantly lower in version 4 than in version 3 (90.5% vs. 98.0%).

Both the CAPI (quota sample) and the P&P (TPI Access Panel) segments were targeted at households with head of the household aged between 18 and 69 years. For the CAPI versions, the quota performance targets were related to the dimension gender (male respondent ratio of 75 percent) and age (a distribution in age classes under 25, 25-34, 35-50 and 50-70 years) according to the current official population statistics (and, in particular, the 2000 micro census).

For the TPI interviewees, the quota targets were also based on the 2000 micro census and either related to the dimensions gender (male respondent ratio of 75 percent) and age (a distribution in age classes 18-29: 13%; 30-39: 24%; 40-49: 22%; 50-59: 21%; 60-69: 20%), and, additionally, whether the respondent is a wage earner or a salaried employee, and the size of the household.

The findings from the experimental manipulation of question format were used to investigate the impact of different survey modes on response behavior (see Essig and Winter, 2003). The next waves benefited from the methodological findings of the 2001 wave and were conducted in summer 2003.

Table 3: Sampling scheme, questioning modes, and incentives in SAVE 2001.

	Version 1	Version 2	Version 3	Version 4	Version 5
Sampling Scheme	Quota	Quota	Quota	Quota	Access Panel
Mode: Parts 1, 2, 3, 5, 6	CAPI	CAPI	CAPI	CAPI	P&P (mail back)
Mode: Part 4 (sensitive items)	CAPI	CAPI	P&P drop-off (pick up)	P&P drop-off (mail back)	P&P (mail back)
Incentives	NO	NO	NO	NO	Present*/hh
Nonresponse due to “no time”/“don’t want”	n/a	n/a	n/a (Return P&P: 98%)	n/a (Return P&P: 90.5%)	n/a
Overall response rate	n/a	n/a	n/a	n/a	n/a
Net number of households	295	304	294	276	660

Notes:

- *Data for 2001 quota sample and access panel were not recorded*
- *No information on reasons for nonresponse available for Access Panel*

3.3 2003/2004

The SAVE 2003 wave consisted of two major samples. The first one consisted of the households which already participated in the SAVE 2001 CAPI sample. The second one was a newly added “refreshment” random sample. Interview modes for the two subsamples were identical. They were CAPI interviews except for part 4 (drop-off with mail-back / collection by the interviewer), see table 4.

3.3.1 Panel CAPI sample

One of the major interests of the SAVE study is to analyze behavioral and financial changes over time. Therefore, we tried to re-contact the interviewees from the 2001 personal interviews (N=1169) again in 2003. 72% (= 840 households) were available as gross sample in 2003. After different stages of losses (moved away/died, refused, no time, not available) and rejecting some incomplete interviews, only 483 completed interviews were available. The fieldwork for the 2001 CAPI sample in 2003 took place between June 2 and July 18, 2003.

3.3.2 Random Route sample

The data universe for the SAVE 2003 random sample were all German speaking households in Germany with the households' head being eighteen years and older. Interviewees were selected from a multiply stratified multistage random sample. All communities were segmented into stratifications by regional criteria. Stratification criteria were states (Bundesländer), districts, and community types. For further sampling details, see Heien and Kortmann (2003).

The fieldwork for the Random Route sample began on May 26 and ended on July 14.

3.3.3 Access Panel

The Access Panel was re-contacted in 2004 only. As figure 1 shows, 487 households participated in the Access Panel.

Table 4: Sampling scheme, questioning modes, and incentives in SAVE 2003/2004.

-2003-

	CAPI-AR 2001	CAPI-SR 2003
	Version 3	Version 3
Sampling Scheme	Quota	Random Route
Mode: Parts 1, 2, 3, 5, 6	CAPI	CAPI
Mode: Part 4 (sensitive items)	P&P drop-off (pick up)	P&P drop-off (pick up)
Incentives	NO	NO
Nonresponse due to “no time”/“don’t want”	25.3%	36.7%
Overall response rate	63.4%	47.1%
Net number of households	483	2184

-2004-

	TPI Access 2001
	Version 5
Sampling Scheme	Access Panel
Mode: Parts 1, 2, 3, 5, 6	P&P (mail back)
Mode: Part 4 (sensitive items)	P&P (mail back)
Incentives	Present [*] /hh
Nonresponse due to “no time”/“don’t want”	n/a
Overall response rate	85.4%
Net number of households	487

Notes:

- No information on reasons for nonresponse available for Access Panel

3.3.4 Survey Participation

Table 5: Participation in SAVE 2003

	2003			
	CAPI-SR (new part.)		CAPI- AR (old part.)	
CAPI-Interviews SAVE I			1169	100%
refusal to be interviewed again			329	28,10%
Gross number for SAVE II	4772	100%	840	100%
wrong adress, tp does not exist	-	-	14	1.7%
tp deceased	-	-	8	1.0%
tp moved away	-	-	36	4.3%
no person of the target group at home	31	0.6%	-	-
total neutral losses	31	0.6%	58	6.9%
remaining adresses	4.741	100%	782	100%
nobody at home	455	9.6%	11	1.4%
tp not at home	151	3.2%	28	3.6%
tp out of town / in holidays	32	0,70%	27	3.5%
tp ill / not able to answer	54	1.1%	8	1.0%
Lack of time	559	11.8%	58	7.4%
refusal / other reason	1180	24.9%	140	17.9%
language difficulties	79	1.7%	-	-
other losses	-	-	14	1.8%
total losses	2510	52.9%	286	36.6%
realized interviews	2231	47.1%	496	63.4%
not able to analyze	47	1.0%	13	2.6%
analyzed interviews	2184	46.1%	483	61.8%

3.4 SAVE 2005

3.4.1 Access-Panel sample

In 2005, there have been three different groups of participants. The members of the Access-Panel were interviewed for the third time between Mai 4 and Mai 31; 360 interviews were realised (some more than the estimated 320). Information about unit nonresponse can be found in table 7. Unfortunately, detailed information about the reasons of the losses is not available due to the written form of the questionings.

3.4.2 Random Route sample

The interviews with the participants of the CAPI-SR group (second questioning) were conducted in two phases. At first, all panel members received a P&P questionnaire, parallel to the interviews of the Access Panel. 381 interviews could be conducted this way. After an address investigation CAPI-interviewers were sent to the panel members who had not sent back the P&P questionnaire. This period lasted from June 20 until August 19, achieving 308 more interviews. Altogether, this is a quota of 54.1% (see table 6).

The third group (SR-BUS, new participants) were interviewed parallel to the second phase of the CAPI-SR group, from June 20 until August 19. Information about unit nonresponse is presented in table 6 as well.

Table 6: Sampling scheme, questioning modes, and incentives in SAVE 2005.

	CAPI-SR 2003		CAPI BUS 2005	TPI Access 2001
	Version 5	Version 3	Version 3	Version 5
Sampling Scheme	Random Route	Random Route	Random Route (BUS)	Access Panel
Mode: Parts 1, 2, 3, 5, 6	P&P (mail back)	CAPI	CAPI	P&P (mail back)
Mode: Part 4 (sensitive items)	P&P (mail back)	P&P drop-off (pick up)	P&P drop-off (pick up)	P&P (mail back)
Incentives	15 €/hh	15 €/hh	15 €/hh	Present*/hh
Nonresponse due to “no time“/“don’t want“	24.5%		40.0%	n/a
Overall response rate	57.9%		40.1%	86.9%
Net number of households	368	278	1302	357

Notes:

- * Present: „Small objects of daily use“, e.g. pen etc.; value: usually between 3€ and 5€.
- No information on reasons for nonresponse available for Access Panel

Table 7: SAVE 2005

2005						
Access-Panel (3rd questioning) CAPI-SR (2nd questioning) SR-BUS (new part.)						
Gross number I					4500	100%
written refusal of participation in the panel					821	18.2%
Gross number II	411	100%	1372	100%	3679	100%
adress not findable			147	10.7%	69	1.9%
adress not used			-	-	132	3.6%
Other neutral losses			31	2.2%	180	4.9%
Total neutral losses			178	13.0%	381	10.4%
remaining adresses			1.194	100%	3298	100%
refusal between phase I and II		<i>detailed information is not available for the Access Panel</i>	52	4.4%	-	-
nobody at home			60	5.0%	277	8.4%
tp not at home			30	2.5%	142	4.3%
tp out of town / in holidays			27	2.3%	125	3.8%
tp ill / not able to answer			31	2.6%	102	3.1%
lack of time			79	6.6%	234	7.1%
refusal / other reason			214	17.9%	1084	32.9%
language difficulties			10	0.8%	13	0.4%
Total losses			503	42.1%	1977	59.9%
realized interviews	360	87.6%	691	57.9%	1321	40.1%
not able to analyze	3	0.7%	45	6.5%	19	1.4%
analyzed interviews	357	86.9%	646	54.1%	1302	39.5%

3.5 SAVE 2006

3.5.1 Random Route sample

The Random Route sample consists of two sub-samples. 552 members (of 646 in 2005) from the CAPI-SR 2003 agreed to participate a third time, whereas 1.171 members (of 1.302 in 2005) of the CAPI-BUS 2005 signaled to take part in a second interview.

488 (1.043) interviews were conducted with the participants from the CAPI-SR 2003 (CAPI-BUS 2005).

These interviews were realized in two phases. The P&P questionnaire lasted from March 9 until April 21. Those households who did not answer the P&P questionnaire were contacted by an interviewer. Those CAPI-interviews were conducted between April 24 and July 7.

Detailed information about unit nonresponse is presented in tables 8 and 9.

Table 8: SAVE 2006

	2006			
	Access-Panel		CAPI-SR and CAPI-BUS	
Gross number for SAVE IV	1974	100%	1723	100%
wrong adress, tp does not exist			8	0.5%
tp deceased			4	0.2%
other neutral losses			20	1.2%
total neutral losses			32	1.9%
remaining adresses			1691	100%
		<i>detailed information is not available for the Access Panel</i>		
refusal between phase I and II			9	0.5%
nobody at home			24	1.4%
tp not at home			7	0.4%
tp out of town / in holidays			8	0.5%
tp ill / not able to answer			12	0.7%
lack of time			13	0.8%
refusal / other reason			46	2.7%
language difficulties			1	0.1%
other losses			40	2.4%
total losses			160	9.5%
realized interviews	1974	100%	1531	90.5%
not able to analyze	5	0.3%	26	1.7%
analyzed interviews	1969	99.7%	1505	89.0%
appointed interviewers	n/a		183	

3.5.2 Access Panel

Parallel to the P&P questionnaire phase of the random route sample the members of the Access Panel were interviewed from March 9 until April 21.

In 2006, 333 members of the Access-Panel 2001 remained (compared to 357 in 2005).

Additionally, TNS Infratest TPI drew a new sample of 2.500 persons. Unlike the Access-Panel 2001, the sampling from the Access-Panel 2006 was done respecting an upper age limit of 79 years.

Before participating, all households were informed about the long term character of the survey and the resulting requirements. Finally, 1.636 interviews (65.4% of the gross sample of 2.500) were conducted for the sub-sample Access-Panel 2006.

Table 9: Sampling scheme, questioning modes, and incentives in SAVE 2006.

	CAPI-SR 2003		CAPI BUS 2005	TPI Access 2001	TPI Access 2006
	Version 5	Version 3	Version 3	Version 5	Version 5
Sampling Scheme	Random Route	Random Route	Random Route (BUS)	Access Panel	Access Panel
Mode: Parts 1, 2, 3, 5, 6	P&P (mail back)	CAPI	CAPI	P&P (mail back)	P&P (mail back)
Mode: Part 4 (sensitive items)	P&P (mail back)	P&P drop-off (pick up)	P&P drop-off (pick up)	P&P (mail back)	P&P (mail back)
Incentives	20 €/hh	20 €/hh	20 €/hh	Present*/hh	Present*/hh
Nonresponse due to “no time”/“don’t want”	3.5%			n/a	n/a
Overall response rate	89.0%			98.8%	65.4%
Net number of households	373	111	1.021	333	1.636

Notes:

- * Present: „Small objects of daily use“, e.g. pen etc.; value: usually between 3€ and 5€.
- No information on reasons for nonresponse available for Access Panel

4. Item Nonresponse and Imputation

However, various interdependent factors that can only be controlled to a limited extent, such as privacy concerns, respondent uncertainty, cognitive burden of the questions, and survey context, lead to unit nonresponse and item nonresponse. Unit nonresponse is the lack of any information on a given observation and as such is the strongest type of refusal. Unit nonresponse rates have already been described above. The phenomenon that only a subset of the information is missing, e.g. only the response to the question on household income, is referred to as item nonresponse.

The general phenomenon of item nonresponse to questions in household surveys has been analyzed by various authors, beginning with the work by Ferber (1966); see also Schnell (1997) and Beatty and Herrmann (2002) for reviews. Recent examples for Germany, focusing on income, saving, and asset choice, are Biewen (2001), Riphahn and Serfling (2005), and Schräpler (2003), who work with data from the German Socio-Economic Panel (GSOEP). Finally, Essig and Winter (2003) describe and analyze nonresponse patterns to financial questions in the first wave of the German SAVE study. They exploit that this first wave has included a controlled experiment specifically designed to analyze the effects of interview mode and question format on answering behavior.

For the large majority of variables in SAVE, item nonresponse is not a problem. For example, there is hardly any nonresponse to detailed questions about socio-demographic conditions of the household, to questions about households' expectations and about indicators of household economic behavior. Mainly due to privacy concerns and cognitive burden, though, there are significantly higher rates of item nonresponse for detailed questions about household financial circumstances than to other less private and less sensitive questions. Taking the 2003/2004 wave as an example, tables 8 and 9 show that these questions can have a missing rate of over 40%. Similar missing rates to questions about financial circumstances have been documented in various survey contexts (e.g., Bover, 2004; Hoynes et al., 1998; Juster and Smith, 1997; Kalwij and van Soest, 2005).

Table 8: Response rates for monthly net income and for the question about total annual savings in 2003/2004.

	Value	Bracket	Unknown
Net income	69%	25%	6%
Annual savings	88%		12%

Note: All calculations are unweighted.

Table 9: Response rates for financial and real wealth items in 2003/2004.

	Have item			Value reported for those having the item
	Yes	No	Unknown	
Savings/term accounts	56%	36%	8%	74%
Building society savings agreements	26%	66%	8%	67%
Whole life insurance policies	28%	64%	8%	57%
Bonds	8%	84%	8%	57%
Shares & real-estate funds	18%	74%	8%	61%
Owner occupied housing	47%	49%	4%	96%

Note: All calculations are unweighted.

For studies that use the detailed financial information in the SAVE study, missing information on one of those variables is a problem. It is tempting to simply delete all observations with missing values. But deleting observations with item nonresponse, i.e. relying on a complete-case analysis might lead to an efficiency loss due to a smaller sample size and to biased inference when item nonresponse is related to the variable of interest.¹ Particularly for multivariate analyses that involve a large number of covariates, case deletion procedures discard a considerably high proportion of subjects, even if the per-item rate of missingness is rather low.

To prevent biased inference based on an analysis of only complete cases, an iterative multiple imputation procedure has been applied to the SAVE data (Schunk, 2007). Iterative multiple imputation methods have recently been applied to other large-scale socio-economic survey data (see Bover, 2004; Kennickell, 1998). The imputation method for the U.S. Survey of Consumer Finances, developed by Arthur Kennickell, has been applied to the Spanish Survey of Household Finances (Bover 2004), and it has also inspired the development of the imputation method that is used for SAVE. Multiple imputation (Rubin, 1987) simulates the distribution of missing data to allow for a more realistic assessment of variances than single imputation. The procedure uses a Markov-Chain Monte-Carlo method to replace missing data by draws from an estimate of the conditional distribution of the data. The multiple imputation algorithm generates five complete data sets with all missing values replaced by imputed values. The differences between the imputed values across those five datasets reflect the uncertainty about their true value.

That is, for each wave, five imputed versions of the SAVE-data are available. Additionally, we provide a so called indicator data set. This data set (“SAVE_[year]_indicator.dta”) indicates whether a certain value is original (0) or has been imputed (1).

¹ See, e.g., Rubin (1987) and Little and Rubin (2002) for discussions about efficiency and bias in a missing data context.

5. Design of Weights

5.1 Preliminary Remarks

For reasons of representativeness, observations are weighted when doing computations with SAVE data. To calculate the weights, Mikrozensus surveys from the Statistisches Bundesamt are taken into account as a representative standard of comparison.

There are two types of weights, each of which compare SAVE to the Mikrozensus in two dimensions. The first type of weights compares SAVE to the Mikrozensus dependent on the dimensions age and income, the second type dependent on household size and income.

5.2 Calculation of Weights Dependent on Age and Income

5.2.1 Basic Method

The observations in SAVE are split into 9 categories („cells“) according to 3 age classes and 3 income classes:

	Income class 1	Income class 2	Income class 3
Age class 1	<i>cell 1</i>	<i>cell 2</i>	<i>cell 3</i>
Age class 2	<i>cell 4</i>	<i>cell 5</i>	<i>cell 6</i>
Age class 3	<i>cell 7</i>	<i>cell 8</i>	<i>cell 9</i>

The number of observations in each cell is divided by the total number of observations in the SAVE sample in order to calculate each cell's relative frequency in the sample. Thus, there are 9 relative frequencies which add up to 1. For the Mikrozensus, the observations are split into the 9 cells accordingly (3 age classes, 3 income classes) to determine each cell's relative frequency in the Mikrozensus sample.

Dividing the relative frequency of each cell in the Mikrozensus by the relative frequency of the corresponding cell in SAVE yields the weight for each cell. One weight is assigned to each observation according to the observation's cell. Since there are 9 cells, there exist 9 weights per sample.

A weight greater than 1 implies that the cell's appearance in the representative Mikrozensus is higher than in SAVE. Thus, SAVE observations in this cell are weighted relatively high. A weight smaller than 1 implies that the cell's appearance in the representative Mikrozensus is lower than in SAVE. Therefore, SAVE observations are weighted relatively low. A weight equal to 1 implies that the cell's appearance in SAVE corresponds to the representative appearance in the Mikrozensus.

5.2.2 Calculation

Method 1:

The weights resulting from this method are the most common ones used in computations with SAVE data.

The following three age classes are applied:

Age class 1: under 35 years of age

Age class 2: 35 to 55 years of age

Age class 3: 55 years or above

The following three income classes are applied:

Income class 1: below 1300 € of net income per month

Income class 2: 1300 € to 2600 € of net income per month

Income class 3: 2600 € of net income per month and above

As described in 7.2.1, the weight of each cell is determined and each observation is assigned one of the nine different weights according to which cell they belong.

Method 2:

This method corresponds to method 1 except for the age classes applied. Method 2 uses the following age classes:

Age class 1: under 35 years of age

Age class 2: 35 to 65 years of age

Age class 3: 65 years or above

The three income classes remain the same.

5.3 Calculation of Weights Dependent on Household Size and Income

The calculation of weights dependent on household size and income corresponds to the calculation dependent on age and income. Instead of age classes, however, 3 different household sizes are used to divide the observations into 9 cells.

	Income class 1	Income class 2	Income class 3
Household size 1	<i>cell 1</i>	<i>cell 2</i>	<i>cell 3</i>
Household size 2	<i>cell 4</i>	<i>cell 5</i>	<i>cell 6</i>
Household size 3	<i>cell 7</i>	<i>cell 8</i>	<i>cell 9</i>

The following household sizes are applied:

Household size 1: one person

Household size 2: two persons

Household size 3: three persons or more

The three income classes remain the same.

5.4 Weights in SAVE 2001

The SAVE 2001 income classes differ slightly from the income classes used in the other SAVE surveys. The following income classes are applied:

Income class 1: under 1278 € of net income per month

Income class 2: 1278 € to 2556 € of net income per month

Income class 3: 2556 € of net income per month and above

Weights are calculated for the entire 2001 survey on the one hand, and for each sample in the 2001 survey (Quota Sample and Access Panel) separately on the other.

The following weight variables appear in the imputed 2001 dataset:

„weights_2001_age_inc_1”

- *Weights dependent on age and income*
- *Computation of weights according to method 1*
- *Computation of weights separately for each sample in the 2001 survey*
- *Each observation is assigned one weight, there are 18 different weights in total (2 samples, 9 cells)*
- *Mikrozensus 2001 as benchmark*

„weights_2001_age_inc_all_1”

- *Weights dependent on age and income*
- *Computation of weights according to method 1*
- *Computation of weights for the entire 2001 survey*
- *Each observation is assigned one weight, there are 9 different weights in total (1 sample, 9 cells)*
- *Mikrozensus 2001 as benchmark*

„weights_2001_age_inc_2”

- *Weights dependent on age and income*
- *Computation of weights according to method 2*
- *Computation of weights separately for each sample in the 2001 survey*
- *Each observation is assigned one weight, there are 18 different weights in total (2 samples, 9 cells)*
- *Mikrozensus 2001 as benchmark*

„weights_2001_age_inc_all_2”

- *Weights dependent on age and income*
- *Computation of weights according to method 2*
- *Computation of weights for the entire 2001 survey*
- *Each observation is assigned one weight, there are 9 different weights in total (1 sample, 9 cells)*
- *Mikrozensus 2001 as benchmark*

„weights_2001_hhsize_inc”

- *Weights dependent on household size and income*
- *Computation of weights separately for each sample in the 2001 survey*
- *Each observation is assigned one weight, there are 18 different weights in total (2 samples, 9 cells)*
- *Mikrozensus 2001 as benchmark*

„weights_2001_hhsize_inc_all”

- *Weights dependent on household size and income*
- *Computation of weights for the entire 2001 survey*
- *Each observation is assigned one weight, there are 9 different weights in total (1 sample, 9 cells)*
- *Mikrozensus 2001 as benchmark*

5.5 Weights in SAVE 2003 / 2004

Weights are calculated for each sample in the 2003/04 survey (Quota Sample 2003, Random Route Sample 2003 and Access Panel 2004) separately; weights are also calculated for the entire 2003 and the entire 2004 sample, and for the entire 2003/04 SAVE survey.

The following weight variables appear in the imputed 2003/04 dataset:

„weights_2003_2004_age_inc_1”

- *Weights dependent on age and income*
- *Computation of weights according to method 1*
- *Computation of weights separately for each sample in the 2003/04 survey*
- *Each observation is assigned one weight, there are 27 different weights in total (3 samples, 9 cells)*
- *Mikrozensus 2002 serves as the benchmark for both 2003 samples, Mikrozensus 2003 as the benchmark for the Access Panel 2004*

„weights_2003_2004_age_inc_03_1”

- *Weights dependent on age and income*
- *Computation of weights according to method 1*
- *Computation of weights separately for the 2003 sample and the 2004 sample*
- *Each observation is assigned one weight, there are 18 different weights in total (2 samples, 9 cells)*
- *Mikrozensus 2002 serves as the benchmark for the 2003 sample, Mikrozensus 2003 as the benchmark for the 2004 sample*

„weights_2003_2004_age_inc_all_1”

- *Weights dependent on age and income*
- *Computation of weights according to method 1*
- *Computation of weights for the entire 2003/04 survey*
- *Each observation is assigned one weight, there are 9 different weights in total (1 sample, 9 cells)*
- *Mikrozensus 2002 as benchmark*

„weights_2003_2004_age_inc_2”

- *Weights dependent on age and income*
- *Computation of weights according to method 2*
- *Computation of weights separately for each sample in the 2003/04 survey*
- *Each observation is assigned one weight, there are 27 different weights in total (3 samples, 9 cells)*
- *Mikrozensus 2002 serves as the benchmark for both 2003 samples, Mikrozensus 2003 as the benchmark for the Access Panel 2004*

„weights_2003_2004_age_inc_03_2”

- *Weights dependent on age and income*
- *Computation of weights according to method 2*
- *Computation of weights separately for the 2003 sample and the 2004 sample*
- *Each observation is assigned one weight, there are 18 different weights in total (2 samples, 9 cells)*
- *Mikrozensus 2002 serves as the benchmark for the 2003 sample, Mikrozensus 2003 as the benchmark for the 2004 sample*

„weights_2003_2004_age_inc_all_2”

- *Weights dependent on age and income*
- *Computation of weights according to method 2*
- *Computation of weights for the entire 2003/04 survey*
- *Each observation is assigned one weight, there are 9 different weights in total (1 sample, 9 cells)*
- *Mikrozensus 2002 as benchmark*

„weights_2003_2004_hhsize_inc”

- *Weights dependent on household size and income*
- *Computation of weights separately for each sample in the 2003/04 survey*
- *Each observation is assigned one weight, there are 27 different weights in total (3 samples, 9 cells)*
- *Mikrozensus 2002 serves as the benchmark for both 2003 samples, Mikrozensus 2003 as the benchmark for the Access Panel 2004*

„weights_2003_2004_hhsize_inc_03”

- *Weights dependent on household size and income*
- *Computation of weights separately for the 2003 sample and the 2004 sample*
- *Each observation is assigned one weight, there are 18 different weights in total (2 samples, 9 cells)*
- *Mikrozensus 2002 serves as the benchmark for the 2003 sample, Mikrozensus 2003 as the benchmark for the 2004 sample*

„weights_2003_2004_hhsize_inc_all”

- *Weights dependent on household size and income*
- *Computation of weights for the entire 2003/04 survey*
- *Each observation is assigned one weight, there are 9 different weights in total (1 sample, 9 cells)*
- *Mikrozensus 2002 as benchmark*

5.6 Weights in SAVE 2005

Weights are calculated for the entire 2005 survey on the one hand, and for each sample in the 2005 survey (Random Route Sample and Access Panel) separately on the other.

The following weight variables appear in the imputed 2005 dataset:

„weights_2005_age_inc_1”

- *Weights dependent on age and income*
- *Computation of weights according to method 1*
- *Computation of weights separately for each sample in the 2005 survey*
- *Each observation is assigned one weight, there are 18 different weights in total (2 samples, 9 cells)*
- *Mikrozensus 2004 as benchmark*

„weights_2005_age_inc_all_1”

- *Weights dependent on age and income*
- *Computation of weights according to method 1*
- *Computation of weights for the entire 2005 survey*
- *Each observation is assigned one weight, there are 9 different weights in total (1 sample, 9 cells)*
- *Mikrozensus 2004 as benchmark*

„weights_2005_age_inc_2”

- *Weights dependent on age and income*
- *Computation of weights according to method 2*
- *Computation of weights separately for each sample in the 2005 survey*
- *Each observation is assigned one weight, there are 18 different weights in total (2 samples, 9 cells)*
- *Mikrozensus 2004 as benchmark*

„weights_2005_age_inc_all_2”

- *Weights dependent on age and income*
- *Computation of weights according to method 2*
- *Computation of weights for the entire 2005 survey*
- *Each observation is assigned one weight, there are 9 different weights in total (1 sample, 9 cells)*
- *Mikrozensus 2004 as benchmark*

„weights_2005_hhsize_inc”

- *Weights dependent on household size and income*
- *Computation of weights separately for each sample in the 2005 survey*
- *Each observation is assigned one weight, there are 18 different weights in total (2 samples, 9 cells)*
- *Mikrozensus 2004 as benchmark*

„weights_2005_hhsize_inc_all”

- *Weights dependent on household size and income*
- *Computation of weights for the entire 2005 survey*
- *Each observation is assigned one weight, there are 9 different weights in total (1 sample, 9 cells)*
- *Mikrozensus 2004 as benchmark*

5.7 Weights in SAVE 2006

The weights in SAVE 2006 are calculated accordingly to the procedure in 2005. The Mikrozensus 2005 is used as benchmark

6. References

Beatty, P. and D. Herrmann (2002): To answer or not to answer: Decision processes related to survey item nonresponse. In: R. M. Groves, D. A. Dillman, J. L. Eltinge, and R. J. A. Little (Eds.), *Survey Nonresponse*, 71-85. New York: Wiley.

Biewen, M. (2001): Item non-response and inequality measurement: Evidence from the German earnings distribution. *Allgemeines Statistisches Archiv*, 85(4), 409-425.

Börsch-Supan, A. and L. Essig (2005): Household saving in Germany: Results of the first SAVE study. In: D. A. Wise (Ed.), *Analyses in the Economics of Aging*, 317-352. Chicago: The University of Chicago Press.

Bover, O. (2004): The Spanish Survey of Household Finances (EFF): Description and Methods of the 2002 Wave. *Documentos Ocasionales N. 0409*. Banco de Espana.

Cameron, A. C. and P. K. Trivedi (2005): *Microeconometrics. Methods and Applications*. New York: Cambridge University Press.

Essig, L. (2005a): Methodological aspects of the SAVE data set. *Mea-Discussion-Paper 80-05*, MEA – Mannheim Research Institute for the Economics of Aging. University of Mannheim.

Essig, L. and J. Winter (2003): Item Nonresponse to Financial Questions in Household Surveys: An Experimental Study of Interviewer and Mode Effects. *MEA-Discussion Paper 39-03*, MEA – Mannheim Research Institute for the Economics of Aging. University of Mannheim.

Ferber, R. (1966): Item nonresponse in a consumer survey. *Public Opinion Quarterly*, 30 (3), 399-415.

Geman, S. and D. Geman (1984): Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of Images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6 (6), 721-741.

Graham, J. W. and J. L. Schafer (1999): On the performance of multiple imputation for multivariate data with small sample size. In: R. Hoyle (Ed.), *Statistical Strategies for Small Sample Research*, 1-29, Thousand Oaks, CA: Sage.

Heien, T. and K. Kortmann (2003): Spar- und Anlageverhalten privater Haushalte (SAVE II). Methodenbericht. München: Infratest Sozialforschung.

Heien, T. and K. Kortmann (2005): Spar- und Anlageverhalten privater Haushalte (SAVE III). Methodenbericht. München: Infratest Sozialforschung.

Hoynes, H., M. Hurd, and H. Chand (1998): Household Wealth of the Elderly under Alternative Imputation Procedures. In: D. A. Wise (Ed.), *Inquiries in the Economics of Aging*, 229-257. Chicago: The University of Chicago Press.

- Juster, F. T. and J. P. Smith (1997):** Improving the quality of economic data: Lessons from the HRS and AHEAD. *Journal of the American Statistical Association*, 92 (440), 1268-1278.
- Kalwij, A. and A. van Soest (2005):** Item Non-Response and Alternative Imputation Procedures. In: A. Börsch-Supan and H. Jürges (Eds.), *The Survey of Health, Ageing and Retirement in Europe – Methodology*, 128-150. Mannheim: Mannheim Research Institute for the Economics of Aging.
- Kennickell, A. B. (1998):** Multiple Imputation in the Survey of Consumer Finances. *Proceedings of the 1998 Joint Statistical Meetings*, Dallas TX.
- Li, K. (1988):** Imputation Using Markov Chains. *Journal of Statistical Computing and Simulation*, 30, 57-79.
- Li, K., T. Raghunathan, and D. Rubin (1991):** Large sample significance levels from multiply-imputed data using moment-based statistics and an F reference distribution. *Journal of the American Statistical Association*, 86, 1065–1073.
- Little, R. J. A. and D. B. Rubin (2002):** *Statistical Analysis with Missing Data*. New York: Wiley.
- Little, R. J. A., I. G. Sande, and F. Scheuren (1988):** Missing-data adjustments in large surveys. *Journal of Business and Economic Statistics*, 6 (3), 117-131.
- Manski, C. (2005):** Partial Identification with Missing Data: Concepts and Findings. *International Journal of Approximate Reasoning*, 39 (2-3), 151-165.
- Riphahn, R. and O. Serfling (2004):** Item Non-response on Income and Wealth Questions. *Empirical Economics*, 30 (2), 521-538.
- Rubin, D. B. (1987):** *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley.
- Rubin, D. B. (1996):** Multiple Imputation After 18+ Years. *Journal of the American Statistical Association*, 91 (434), 473-489.
- Rubin, D. B. and N. Schenker (1986):** Multiple Imputation for Interval Estimation from Simple Random Samples with Ignorable Nonresponse. *Journal of the American Statistical Association*, 81 (394), 366-374.
- Schafer, J. L. (1997):** *Analysis of incomplete multivariate data*. London: Chapman & Hall.
- Schnell, R. (1997):** *Nonresponse in Bevölkerungsumfragen*. Opladen: Leske+Budrich.
- Schräpler, J.-P. (2003):** Gross income non-response in the German Socio-Economic Panel: Refusal or don't know? *Schmollers Jahrbuch*, 123, 109-124.
- 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.

Schunk, D. (2007): A Markov Chain Monte Carlo Algorithm for Multiple Imputation in Large Surveys. Mimeo. University of Mannheim.

Tanner, M. A. and W. H. Wong (1987): The Calculation of Posterior Distributions by Data Augmentation. *Journal of the American Statistical Association*, 82 (398), 528-550.

Appendix

A1. Questionnaire and Dataset

A German and an English version of the questionnaire of each wave of the survey can be obtained upon request:

save@mea.uni-mannheim.de

A2. Documentation of all Variables

The following pages provide a table with detailed information on the variables asked in each wave of the SAVE study, as well as their coding and variable names in the SAVE-dataset.

SONDERFORSCHUNGSBereich 504 WORKING PAPER SERIES

Nr.	Author	Title
07-08	Daniel Schunk	The German SAVE survey: documentation and methodology
07-07	Hans-Martin von Gaudecker Carsten Weber	Mandatory unisex policies and annuity pricing: quasi-experimental evidence from Germany
07-06	Daniel Schunk	A Markov Chain Monte Carlo Multiple Imputation Procedure for Dealing with Item Nonresponse in the German SAVE Survey
07-05	Hans-Martin von Gaudecker Rembrandt Scholz	Lifetime Earnings and Life Expectancy
07-04	Christopher Koch Daniel Schunk	The Case for Limited Auditor Liability - The Effects of Liability Size on Risk Aversion and Ambiguity Aversion
07-03	Siegfried K. Berninghaus Werner Gueth M. Vittoria Levati Jianying Qiu	Satisficing in sales competition: experimental evidence
07-02	Jannis Bischof Michael Ebert	Inconsistent measurement and disclosure of non-contingent financial derivatives under IFRS: A behavioral perspective
07-01	Jörg Oechssler Carsten Schmidt Wendelin Schnedler	Asset Bubbles without Dividends - An Experiment
06-16	Siegfried K. Berninghaus Hans Haller	Pairwise Interaction on Random Graphs
06-15	Markus Glaser Philipp Schmitz	Privatanleger am Optionsscheinmarkt
06-14	Daniel Houser Daniel Schunk Joachim Winter	Trust Games Measure Trust
06-13	Markus Glaser Sebastian Müller	Der Diversification Discount in Deutschland: Existiert ein Bewertungsabschlag für diversifizierte Unternehmen?