Cross-National Comparability in AMANDA and SHARE

Axel Börsch-Supan
Mannheim Research Institute for the Economics of Aging

Paper for the AMANDA Final Conference in Como, Italy, 24 February 2006

This version: 14 February 2006

Abstract: An overarching aim of the analyses in the AMANDA project is to understand how the diverse historical, cultural and institutional settings – for example the diverse welfare policies and their current reforms – influence the ageing process of individuals and societies. The AMANDA analyses heavily rely on cross-national variation to identify causal effects of policy interventions.

Cross-national analysis provides chances and poses challenges. They are the subject of this paper. It describes which measures the SHARE and AMANDA sister projects have taken to cope with the challenges and to maximize the chances.

Keywords: Cross-national research; Ageing; European welfare states

Acknowledgements: The work in this paper has been supported financially through the 5th framework programme under the project name of AMANDA (“Advanced Multidisciplinary Analysis of New Data on Ageing”, QLK6-CT-2002-002426).

The SHARE data collection has been primarily funded by the European Commission through the 5th framework programme (project QLK6-CT-2001-00360 in the thematic programme Quality of Life). Additional funding came from the US National Institute on Aging (U01 AG09740-13S2, P01 AG005842, P01 AG08291, P30 AG12815, Y1-AG-4553-01 and OGHA 04-064). Data collection was nationally funded in Austria (through the Austrian Science Fund, FWF), Belgium (through the Belgian Science Policy Office) and Switzerland (through BBW/OFES/UFES).

Parts of this paper rest on contributions by Kirsten Alcser, Grant Benson, Marcel Das, Janet Harkness, Hendrik Jürges, Arie Kapteyn, Anders Klevmarken, Giuseppe de Luca, Franco Peracchi and Arthur van Soest to various documents underlying the AMANDA project. The author gratefully acknowledges these contributions and is solely responsible for all remaining errors.
Special thanks go to Stephanie Stuck who manages the AMANDA data base of generated variables.
1. Introduction

Ageing is one of the greatest social and economic challenges of the 21st century, particularly so in Europe. By 2025, about one third of Europe’s population will be aged 60 or over, with a particularly rapid increase in the number of oldest-old citizens (cf. WHO 2002). While the demographic mechanisms driving this development – lower fertility rates and longer life-spans (e.g., Hayward & Zhang 2001) – are well-known, our knowledge about the social and economic consequences of population ageing is yet incomplete (e.g., Börsch-Supan 2004). Public policy clearly plays a key role here. The typical European combination of an ageing population and widespread early retirement (cf. Kohli et al. 1991), for example, puts severe strains on our social security systems’ capacity to maintain today’s standard of living for future generations of older people.

To cope with these and other challenges, such as growing long-term care needs (e.g., Batljan & Lagergren 2005), it is important to achieve a better understanding of the complex linkages between economic, health, and social factors that determine the quality of life of the older population. These interactions take place at the individual level in the first place, they are dynamic as ageing is a process, not a state in time, and they are fundamentally affected by a country’s welfare regime, e.g. its labour market institutions, social security and health care systems.

SHARE -- the longitudinal Survey of Health, Ageing and Retirement in Europe – is designed to provide an infrastructure to help researchers to better understand the individual and population ageing process: where we are, where we are heading to, and how we can influence the quality of life as we age – both as individuals and as societies. A particular aim of the analyses based on SHARE, notably the EU-sponsored project on Advanced Multidisciplinary Analyses of New Data on Ageing (AMANDA), is to understand how the diverse historical, cultural and institutional settings in Europe influence the ageing process and its implications for individuals.

This paper reflects on the chances and challenges of cross-national research, and describes which measures the SHARE and AMANDA sister projects have taken to cope with the challenges and to maximize the chances.

The paper begins by pointing out the general chances and challenges of cross-national research (section 2). Section 3 describes the SHARE survey underlying most analyses. The three elements of constructing a cross-nationally comparable survey consist of a common survey instrument (section 4), a cross-nationally comparable sampling
scheme (section 5) and harmonized interviewing procedures (section 6). Departing from the SHARE raw data, the AMANDA project generated cross-nationally comparable indicators (section 7). Since the AMANDA project puts special emphasis on cross-cutting analyses encompassing health as well as socio-economic factors, particular attention was given to cross-national comparability of health measures. They include attempts to cope with differential item functioning (section 8) and attempts to compensate for the subjectivity of self-assessed health ratings through physical measures (section 9). Section 10 concludes with an outlook on the future of SHARE and AMANDA.

2. Chances and challenges of cross-national research

The case for cross-national research has been made forcefully by the National Research Council (National Research Council, 2001). The essential argument is that the variety of circumstances and policies is in general much larger across countries than within a single country such that policy makers and researchers can learn from what has happened and what has been tried elsewhere. A straightforward example is the fact that population aging varies among countries. Hence, countries with relatively prolonged baby booms will experience the aging of their populations later (e.g. the United States) and may therefore be in a position to evaluate the successes and failures of policies adopted in nations with higher proportions of older populations (e.g. Germany and Italy).

An important arena for cross-national research is public policy evaluation. The impact of economic and social policies can only be understood if we observe one policy in contrast to other policies. The U.S., for instance, has only one Social Security system, and it has only experienced minor changes during the time since good micro data has been available. A counterfactual of retirement policy is hard to construct using U.S. data. International variation provides this contrast (a formal argument is provided in National Research Council 2001).

There are striking examples in pension policy and health care provision for this variation. The share of retirement income provided by public pensions varies from more than 90 percent in the large Continental European countries to less than 50 percent in the Netherlands and Switzerland where more retirement income is provided privately (Börsch-Supan and Miegel, 2001). The extent to which specific retirement ages are induced varies greatly across the major industrial countries (Gruber and Wise, 1999, 2004, 2005). The percentage of GDP spent on health care varies considerably across the industrialized
countries: within the European Union from 7.4% in Finland to 11.1% in Germany, while this percentage is 15.0% in the United States (OECD, 2005). Uptake rates of disability insurance vary dramatically across EU member states without discernible differences in corresponding health care spending (Aarts, Burkhauser, de Jong, 1996, Börsch-Supan, 2005). SHARE 2004 also detected a remarkable North-South gradient in health (Mackenbach et al., 2005).

The work by Gruber and Wise (1999, 2004, 2005) on retirement and the work by Poterba (1994), Börsch-Supan (2001, 2003), Kapteyn and Panis (2002), and Hurd and Kapteyn (2003) on savings and pensions are examples that interpret the variation in institutions across countries as “historical experiments” which help to identify the effects of institutions on societal processes, such as early retirement and disability uptake. The work by the Gruber and Wise group has made significant inroads into pension policy recommendations (Blöndal and Scarpetta, 1998; OECD, 2000; European Commission, 2004). A crucial policy question is whether different international outcomes, such as disability insurance uptake rates, are generated by different economic incentive effects, by preferences that have led to the creation of different policy rules, or by differences in the actual health status.

While the theoretical benefits of comparative international research are becoming increasingly clear, the conduct of such research poses a number of challenges. The basic challenges for cross-national research are the development of survey designs that can be readily adapted to different social and cultural settings; the harmonization of concepts and measures that provide a reasonably acceptable level of cross-national comparability; and the coordination of data collection and analysis across countries (National Research Council, 2001).

Cross-national comparability is achieved by using survey methods that generate, firstly, comparable sample outcomes (such as unit and item response rates, hence representativeness), secondly, by measuring the same concepts, and, thirdly, by creating secondary variables and meta-data that correct for remaining cultural and institutional cross-national variation. The work in SHARE has addressed the first two dimensions, AMANDA the third.

As to the first dimension, the willingness to respond varies a great deal across countries and reflects historical experiences that have shaped cultures of openness or privacy vis-à-vis interviewers. They also result in data confidentially laws that are
strikingly different across countries and create very different environments for actual survey practice. Sections 5 and 6 describe how SHARE has tried to deal with this.

Concerning the second dimension, SHARE has developed several mechanisms to minimize purely linguistic and semantic differences in this multi-language survey. They are described in section 4. Asking the same set of questions, however, is not necessarily sufficient for measuring the same concepts. For many variables used in studies of health, psychology, and economics, common measurements are hard to obtain, especially for subjective items such as health and well-being. SHARE and AMANDA employ two complementary approaches to cope with this. They are described in sections 8 and 9.

Finally, AMANDA has generated a large set of electronic tools, variables and programming procedures for data analysis which are designed to aid researchers in interpreting the SHARE data, see section 7. Since SHARE is a very complex survey, these AMANDA tools are essential for an efficient data analysis.

3. The SHARE 2004 baseline wave
SHARE is modelled closely after the U.S. Health and Retirement Study (HRS; see Juster & Suzman 1995) and the English Longitudinal Study of Ageing (ELSA; see Marmot et al. 2003). Yet, SHARE delivers another dimension: It is the first data set to combine an interdisciplinary with a cross-national dimension, providing extensive information on the socio-economics status, health, and family relationships of the elderly population in a large number of European countries, representing Europe’s economic, social, institutional, and cultural diversity from Scandinavia (Denmark, Sweden) across Western and Central Europe (Austria, Germany, France, Belgium, The Netherlands, Switzerland) to the Mediterranean (Greece, Italy, Spain). Additional data will come from Israel in 2006, and from the Czech Republic, Ireland and Poland in 2006/07.

At this point, SHARE has collected a first baseline wave of data. They include health variables (e.g. self-reported health, physical functioning, cognitive functioning, physical measures such as grip strength and walking speed, health behaviour, use of health care facilities), psychological variables (e.g. psychological health, well-being, life satisfaction, control beliefs), economic variables (e.g. current work activity, job characteristics, job flexibility, opportunities to work past retirement age, employment history, pension rights, sources and composition of current income, wealth and consumption, housing, education),
and social support variables (e.g. assistance within families, transfers of income and assets, social networks, volunteer activities, time use).

The SHARE data is freely available for all researchers; see the application procedure on the project’s website [http://www.share-project.org](http://www.share-project.org). A preliminary release of the data collected until the late Fall of 2004 contains information on some 22,000 individuals aged 50 and older (including spouses, irrespective of age) from 15,000 households in 10 countries. Further data collection has been undertaken until mid 2005, and a second release in 2006 will feature some 27,000 households in 11 countries. A large number of descriptive analyses has been conducted as part of the AMANDA project, see Börsch-Supan et al. (2005).

4. Cross-national comparability of the survey instrument

The first step in the development of a cross-national survey is the construction of a common survey instrument. SHARE developed the questionnaire in English. At this beginning stage, the attention was focused on which variables to include into survey (where questionnaire time is a scarce resource). This selection process was dominated by scientific considerations (i.e., potentially valuable research questions). Already at this stage, however, only questions were selected that were relevant for all participating countries. As a general principle, questions had to be “generic”, i.e. referring to generalizable (often abstract) concepts, rather than country-specific. There were only few exceptions to the generic blueprint of the questionnaire. Country-specific parts were introduced when institutions were fundamentally different, e.g. institutional details of the pension and health care systems. The general principle was to ask for the generic concept, and then to add country-specific examples for that concept. Second, country specifics could be introduced by skipping irrelevant answer categories and by adding new country specific answer categories in the LMU. These exceptions never led to a different sequence of questions for a specific country.

The main challenge for cross-national survey design, however, is of course the multitude of different languages in Europe. In addition to the difficulties of exact translation, the main problems are cross-cultural differences in the underlying concepts and in the way, in which even semantically identical phrases and expressions are perceived in different countries. SHARE – and this is typical for the general strategy in SHARE to

---

Somewhat ironically: English was not spoken in any of the countries participating in the 2004 wave.
minimize cross-national differences – adopted a two-track approach to keep the centrifugal forces through translation and country-specific adaptation in check: develop electronic instruments that make cross-national differences transparent and easily detectable (the language management utility, LMU, developed by CentERdata in Tilburg, Netherlands), and install procedural rules that minimize the differences so detected (a multistage translation process, developed by Janet Harkness at ZUMA, Germany).

The language management utility is essentially a matrix-shaped data base. Each column represents a language, each row an element of the survey instrument (an introductory text, a question, or an answer). The actual survey instrument picks the elements from a language column of the LMU and fits it into the logical structure (the skip pattern of the interview flow). In this way, the LMU forces each national survey instrument into the same logical structure, and SHARE avoids country-specific survey parts. Moreover, the LMU also works as a visual instrument to compare several languages with each other and therefore strengthens cross-national comparability. Finally, the LMU permits a very elegant addition of languages and countries without major programming effort.

The first column of the LMU is the English base version of the survey instrument. The remaining columns were filled in the translation process. Although each participating country in SHARE organised its own translation effort, SHARE initiated several activities to strengthen cross-national comparability. First, all countries were provided with guidelines recommending how to go about hiring translators, testing translators, organising the translation, and reviewing and assessing the translation. SHARE essentially followed the model used in the European Social Survey, although due to its much more complex interdisciplinary content, more freedom was given to the participating countries. Second, the project co-ordinator commissioned an expert in survey translation to advise SHARE participants on any translation queries they might have. Third, the project co-ordinator commissioned a professional review of a sample of the first draft of SHARE translations. SHARE countries were provided with feedback from an external set of translators, each working in their language of first expertise. The translators commented in detail on selected questions and submitted a brief general appraisal of the translation draft, pointing out areas where improvements could be made. This procedure was repeated for a later draft of the questionnaire and feedback again provided to SHARE participants. The pretest-and-pilot design of the SHARE study, coupled with the translation guidelines and appraisals,
provided the SHARE project with a rare opportunity to refine and correct the source questionnaire and the translated versions.

Asking the same set of questions is not necessarily sufficient for measuring the same concepts. Two examples may make this point: asking for the “last household income” in most European situations means “last monthly income net of taxes and social security contribution”. In some situations, however, it may refer to gross income, while in some countries or for certain income sources a weekly or an annual measure is more natural. The underlying principle in SHARE was to phrase a question carefully in order to ask for the same concept in all countries, say, net monthly income.

There is a tension, however, between this principle and what is perceived as “natural” in some countries. In some instances, the SHARE survey deviated from asking strictly the same concept, but the AMANDA project developed a framework in which the concepts can be mapped into each other. This is trivial in the case of weekly, monthly or annual income. It is much more complicated in the case of net vs. gross income since many people do not know their taxes and social security contributions well enough to make a conversion straightforward.

A second example of language problems beyond translation are perceptions associated with certain answer categories. In English “excellent” is a normal way to describe very good health. The translation of this into French, German or Italian, essentially the very same Latin word, is perceived as unusual and extraordinary good health and one would not provoke fate by using this term. The first wave of SHARE provided “Rosetta stones” for cross-national and cross-cultural comparisons of such deviating scales. For example, SHARE used two different five point scales for self-assessed health, the asymmetric HRS scale ranging from “excellent” to “poor”, and a symmetric scale ranging from “very good” to “very bad”, placed at two far apart places in the questionnaire, in randomized order. The AMANDA project then provided a “translation” of the two scales into each other.

A second example for a Rosetta stone are two scales for mental health (specifically: depression). SHARE included the EURO-D scale in the CAPI and the CESD-scale in the drop-off; the AMANDA project then computed predicted EURO-D scores given the CESD items and vice versa.

The problem of scale perception is closely related to what is termed “differential item functioning”. SHARE has addressed this problem in two ways. First, SHARE employs a
new and still experimental way to use “anchoring vignettes” as an instrument to develop a common scale across languages and cultures. This is addressed in section 8. In addition, SHARE employs, wherever possible, measures that are objective and reproducible. This is addressed in section 9.

5. Cross-national comparability of the samples

The mechanics of a cross-national survey start with sampling. What can be done heavily depends on the institutional conditions in each country. So the first challenge in a cross-national survey is to harmonize sampling as far as that is institutionally possible. SHARE set up a working group led by Anders Klevmarken, University of Uppsala, Sweden, to cope with this challenge.

Given the variety of existing European sampling frames, one insight was that it is not possible to do an exact ex ante harmonization of sampling schemes because this would have implied an unsatisfactory least common denominator approach. Hence, cross-national comparability was implemented ex ante by applying a set of “iron principles” (most importantly probability sampling) flexibly to many different circumstances, and ex post by setting up sampling weights.

Most SHARE countries have registers of individuals that permitted stratification by age. In some countries these registers were administered at a regional level, e.g. Germany and the Netherlands. In these cases we needed a two or multi-stage design in which regions were sampled first and then individuals selected within regions. In the two Nordic countries Denmark and Sweden we could draw the samples from national population registers and thus use a relatively simple and efficient design. In France and Spain it became possible to access population registers through the co-operation with the national statistical offices, while in other countries no co-operation was possible.

In three countries, Austria, Greece and Switzerland, no registers were accessible. SHARE use telephone directories as a sampling frame and pre-screened eligible sample participants. As a result the sampling designs used vary from simple random selection of households to rather complicated multi-stage designs.

In order to correct for such sampling differences, SHARE constructed design weights that are all equal in Denmark (simple random sampling) but very different in Italy (complex multi-stage sampling), resulting in cross-national differences in efficiency.
Another iron principle was the selection of individuals within a household. All individuals aged 50 and over and any of their partners were included in the sample independently of whether the primary sampling units were individuals or households. Hence, the inclusion probability of a household is by design the same as that of any of the included household members.

Unit non-response was low for European standards, but somewhat higher than in the most recent HRS wave, see Table 1.

<table>
<thead>
<tr>
<th></th>
<th>SHARE 2004</th>
<th>Eurostat surveys</th>
<th>Scientific surveys</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Denmark</td>
<td>63.2</td>
<td>62</td>
<td>75</td>
<td>68</td>
</tr>
<tr>
<td>Germany</td>
<td>63.4</td>
<td>47</td>
<td>(a)</td>
<td>57</td>
</tr>
<tr>
<td>Italy</td>
<td>55.1</td>
<td>(a)</td>
<td>-</td>
<td>44</td>
</tr>
<tr>
<td>Netherlands</td>
<td>61.3</td>
<td>(a)</td>
<td>59</td>
<td>68</td>
</tr>
<tr>
<td>Spain</td>
<td>53.3</td>
<td>67</td>
<td>(a)</td>
<td>53</td>
</tr>
<tr>
<td>Sweden</td>
<td>50.2</td>
<td>-</td>
<td>(a)</td>
<td>69</td>
</tr>
<tr>
<td>Austria</td>
<td>58.1</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>France</td>
<td>73.6</td>
<td>79</td>
<td>(a)</td>
<td>-</td>
</tr>
<tr>
<td>Greece</td>
<td>61.4</td>
<td>(a)</td>
<td>-</td>
<td>80</td>
</tr>
<tr>
<td>Switzerland</td>
<td>37.6</td>
<td>-</td>
<td>-</td>
<td>34</td>
</tr>
<tr>
<td>Weighted Average</td>
<td>61.8</td>
<td>62.0</td>
<td>63.2</td>
<td>55.6</td>
</tr>
</tbody>
</table>

Notes: (a) no pre-screening response rate reported, (-) country not in sample.

SHARE compares favorably to the other surveys, although the weighted average (62%, unweighted 60%) is still lower than what is typically seen in the US. The appropriate comparison is probably with the newest HRS cohort (the Early Baby Boomers cohort drawn in 2004) which has a response rate at baseline of 69%.

Unit non-response was compensated by adjusting the design weights. This was done in a calibration approach. In most countries the calibration was done to national population totals decomposed by age and gender, in two countries more information could be used and in two countries just national totals by gender were used.
As a result, the country samples can be pooled to a single cross-national sample in spite of very different sampling frames and substantially different unit non-response. Standard errors, however, vary across subsamples and analysts have to take account of this when they estimate population statistics.

6. Cross-national comparability of interviewing procedures

SHARE uses face-to-face interviews, mainly because of the complexity of the SHARE questionnaire. In addition, the available evidence shows that this mode maximizes unit and item response. On the other hand, however, how one asks tends to influence what answers one gets. Hence, controlling the interview process is another important step in achieving cross-national comparability. Much like the translation process, SHARE adopted a two-track strategy to ascertain such comparability: technical tools and process control.

The first technical tool is the computer assisted personal interviewing (CAPI) program. The setup of the CAPI program allowed each country involved to use exactly the same underlying structure of meta-data and routing. The only difference across countries was the language, filled by the LMU described in the previous section. The CAPI program forced the interviewers to stick to the questionnaire text; included routines e.g. for the physical measurements (see section 9), and almost always prescribed answer categories, avoiding open answers.

Next to the CAPI instrument, an electronic case management system (CMS) was developed to manage the co-ordination of the fieldwork. Only three countries used their own system: France, Switzerland, and The Netherlands. The CMS basically consists of a list of all households in the gross sample that should be approached by the interviewer. Contact notes and registrations, appointments with respondents, and area and case information could be entered in the system, and the system enforced common procedures for re-contacting respondents and how to handle non-response. CAPI and CMS interact with each other, providing an integrated electronic survey environment.

In addition to these technical tools, SHARE imposed an extensive – at least for European survey traditions – training program. Since central training is impossible with about 800 interviewers in 14 languages, SHARE utilized a train-the-trainer (TTT) program approach to facilitate decentralized training in the member countries. The Survey Research Center at the University of Michigan at Ann Arbor created a training program for use by country level trainers and provided training for these trainers. The TTT program was
scripted for ease of use and consistent cross-national implementation in subsequent training sessions by local survey agencies. The TTT program provided a set of training resources, including an Interviewer Project Manual, a Facilitator Guide with power point slides and training scripts, a CD-based training on gaining respondent cooperation, and training videos to illustrate the correct interpretation and recording of call attempts, and the administration of physical measurements. The SHARE TTT program trained trainers as if they were the interviewers. First, this afforded the trainers a better appreciation of interviewer needs and difficulties and, thus, allowed them the opportunity to strengthen their own training in areas anticipated to require additional training. Second, this was the best way for trainers to familiarize themselves with the SHARE survey requirements. Most importantly, this approach aimed to ensure standard interviewer behaviour across all member countries to increase comparability of data collected for SHARE.

Another process-oriented tool in order to achieve cross-national comparability was a tight monitoring of the actual field work in real time using the information generated by the case management system. At each time during the field period, the CMS transmitted (with a maximum lag of two weeks):

- how many households had been contacted
- how many interviews had been conducted
- which interviewers were actively working on SHARE and which were currently inactive
- what were the main reasons for non-contact
- what were the main reasons for non-interviews

Given this information, the country-team leaders and the co-ordinator were able to identify possible problems in the field and their reasons very early in the process. Strategies how to cope with such problems were discussed with the country teams and survey agencies and implemented without unnecessary delay. Monthly meetings of all country-tam leaders, sometimes with the agencies, fostered the international coordination of these actions and created a learning process among country teams.

7. Generation of cross-nationally comparable indicators

The AMANDA project has generated a large set of electronic tools, variables and programming procedures for data analysis which are designed to aid researchers in interpreting the SHARE data.
First, AMANDA has created an electronically accessible synopsis of all variables in the SHARE survey, with annotations if country-specific deviations, codings, etc. were present. If so, the synopsis provides a listing of the categories and explains the differences.

Second, AMANDA has imputed missing variables throughout the data in a cross-nationally consistent fashion. Imputed values are flagged and researchers have the choice of using the imputed values or not. The imputations rest on a complex iterative and multivariate algorithm that is based on Hoynes, Hurd and Chand (1998) and the work by Arthur Kennickel on the US Survey of Consumer Finances and will be made available in Release 1.

Third, AMANDA has created a host of generated variables that make the analysis of the SHARE data more convenient for researchers. Such generated variables include technical conversion of multiple answers into more easy to use dummy variables, but also conversions of money values into a common currency (Euro) using 2004 exchange rates and purchasing power parities. Coding inconsistencies across countries were systematically checked, etc.

While these post-processing steps are all but glamorous, they are essential for an efficient data analysis of the very complex SHARE data. Box 1 shows an application of the AMANDA project that was conveniently possible due to this data post-processing. It uses a common definition of disability insurance uptake, and then corrects disability insurance uptake by the general health index based on a large set of health measurements (similar to the approach taken in section 9). The figure shows that the cross-national variation in disability insurance enrolment is not due to cross-national differences in health.
Box 1: The value of generated variables generated by AMANDA

Cross-national variation in disability insurance enrolment with and without a correction for cross-national differences in health

Source: Börsch-Supan (2005)

8. Cross-nationally different item functioning

As emphasized in section 4, asking the same set of questions is not necessarily sufficient for measuring the same concepts. For many variables used in studies of health, psychology, and economics, methods for obtaining common measurements are not well understood. The difficulties are smaller for items that are quantifiable such as retirement age, income and savings, and much larger for subjective items such as health and well-being, see e.g. Sadana et al. (2002) and Tandon et al. (2001). In the parlance of psychometrics, these items suffer from “differential item functioning”, the inter-personal and inter-cultural variation in interpreting and using the response categories for the same question.

Progress is being made on a number of frontiers: a continuing large-scale cross-national effort to construct cross-national measures of general health (WHO, 2001; Tandon
et al. 2001; Murray et al., 2002; Sadana et al., 2002); to create instruments that produce valid measures of depression that are comparable across countries, cultures, and language groups (Prince et al., 1999); and there is a NIA-sponsored program to understand, and correct for, differential item functioning through anchoring vignettes in the Global Burden of Disease program (King et al., 2003).

As a first methodological approach, SHARE has invested considerably resources in adding anchoring vignettes as a self-completion questionnaire (“drop-off”) to the SHARE CAPI instrument. Vignettes are short descriptions of hypothetical persons with some (work related) health problem, for example:

Mark has pain in his back and legs, and the pain is present almost all the time. It gets worse while he is working. Although medication helps, he feels uncomfortable when moving around, holding and lifting things at work.

Respondents are then asked to evaluate the (work related) health of the hypothetical person:

Does Mark have any impairment or health problem that limits the kind or amount of paid work he can do? (1) no, not at all, (2) yes, mildly limited, (3) yes, moderately limited, (4) yes, severely limited, (5) yes, extremely limited, cannot work.

Respondents in different countries are asked to rate the same vignettes. If they evaluate the vignettes systematically different, this points at differences in response scales (“differential item functioning”). The scale difference in the above vignette answers can then be used to re-scale the answers to self-reported (work related) health variables (see King et al., 2004; Banks et al., 2004). While the introduction of vignettes in any national survey is interesting in its own right, it is especially significant in a cross-national survey with a much larger diversity of cultures and therefore potential differential item functioning.

Evaluation of the preliminary vignette data is still undergoing as part of the AMANDA and related projects. Box 2 shows that the vignette correction does not change the ranking of countries very much, but it amplifies the distance of the top runner (Sweden) versus the other five countries analyzed so far.
### Box 2: Scaling Health Status by Anchoring Vignettes

**General Health Status using own response scales and Dutch response scales**

<table>
<thead>
<tr>
<th>Country</th>
<th>Own</th>
<th>Dutch</th>
<th>Own</th>
<th>Dutch</th>
</tr>
</thead>
<tbody>
<tr>
<td>Netherlands</td>
<td>36.5</td>
<td>36.5</td>
<td>19.3</td>
<td>19.3</td>
</tr>
<tr>
<td>Germany</td>
<td>20.4</td>
<td>20.7</td>
<td>35.0</td>
<td>35.2</td>
</tr>
<tr>
<td>Spain</td>
<td>17.3</td>
<td>22.2</td>
<td>38.8</td>
<td>35.7</td>
</tr>
<tr>
<td>Greece</td>
<td>38.1</td>
<td>39.1</td>
<td>22.8</td>
<td>19.7</td>
</tr>
<tr>
<td>Italy</td>
<td>19.7</td>
<td>15.8</td>
<td>39.7</td>
<td>56.2</td>
</tr>
<tr>
<td>Sweden</td>
<td>42.0</td>
<td>53.7</td>
<td>19.3</td>
<td>10.4</td>
</tr>
</tbody>
</table>

**Health Ranking:**

<table>
<thead>
<tr>
<th>Rank</th>
<th>Own</th>
<th>Dutch</th>
<th>Own</th>
<th>Dutch</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>SE</td>
<td>SE</td>
<td>SE</td>
<td>SE</td>
</tr>
<tr>
<td>2</td>
<td>GR</td>
<td>GR</td>
<td>NL</td>
<td>NL</td>
</tr>
<tr>
<td>3</td>
<td>NL</td>
<td>NL</td>
<td>GR</td>
<td>GR</td>
</tr>
<tr>
<td>4</td>
<td>DE</td>
<td>ES</td>
<td>DE</td>
<td>DE</td>
</tr>
<tr>
<td>5</td>
<td>IT</td>
<td>DE</td>
<td>ES</td>
<td>ES</td>
</tr>
<tr>
<td>6</td>
<td>ES</td>
<td>IT</td>
<td>IT</td>
<td>IT</td>
</tr>
</tbody>
</table>

*Source: Arie Kapteyn and Arthur van Soest, preliminary results*

### 9. Physical measurements

A second methodology to overcome cross-national and cross-cultural differences in health assessment is to employ “objective” measures that minimize differential item functioning. SHARE has followed HRS and ELSA in administering a walking speed test to older respondents. Walking speed is reported to have substantial predictive power for self-reported difficulty (or inability) with walking medium distances (cf. Lan et al., 2003). This in its turn is a measure of mobility that is a powerful predictor of disability, defined as physiological limitations, social and environmental barriers, and “sickness” behavior (see Fried and Guralnik, 1997). Disability predicts hospital and long-term care and thus is also important for policy (Mor et al., 1994). On the other hand, the walking speed does not discriminate well among the younger respondents.
SHARE has led the effort to search for a cross-nationally comparable, more objective general health measure by introducing the measurement of grip strength in the entire sample, now being followed by ELSA and HRS. Grip strength is reported to be a very good predictor of ensuing medical problems (Christensen et al., 2000; Rantanen et al. 2001; Al Snih et al. 2002). Grip strength varies significantly even at younger (Rantanen et al. 1999). It is relatively easy to administer in a wide variety of circumstances, facilitating cross-national comparisons. Housing conditions crucial for the walking speed test vary greatly across countries, while these are irrelevant in measuring grip strength.

Grip strength (two measurements of each hand) has been measured in SHARE for all respondents in the core sample (1500 households per country) as well as in the additional vignette sample (between 350 and 500 households in 9 countries). Note that this provides an interesting cross-validation between the approaches in section 8 and 9. Refusal rates were very low and so was the number of respondents who were unable to do the grip strength test due to physical limitations, confirming the findings by Frederiksen et al. (2002). Correlations between grip strength measurements of the same hand or different hands were high. Overall the quality of the grip strength data seems excellent, and grip strength reveals interesting north south differences that are also found in other health variables.

Since health is a multidimensional variable, combinations of physical performance measures have better predictive power for general health than single measurements (e.g. Harding et al., 1994). However, there is no general agreement how many dimensions should reasonably be included in a multi-purpose general survey and which combination works best.

The MOBLI index of Lan et al. (2002) combines walking speed, chair stands and expiratory peak flow. They show that this index is a substantially better predictor of self-reported difficulty with walking medium distances than walking speed alone. In a follow up study, they also show that changes in the MOBLI index are a substantially better predictor of changes in self-reported difficulty with walking than changes in walking speed (Lan et al., 2003). Melzer et al. (2003) show that the combination of walking speed, peak flow, and chair stands has predictive validity for mortality over a four years period. On the other hand, Cress et al. (1995) find that in regressions explaining an index of self-reported physical health, chair stand is not significant once grip strength, walking speed, and
balance test score are controlled for. More and more systematic research is needed to shed light on the question of the optimal measurement combination.

Measuring mental health poses similar problems. Following the same logic, SHARE includes several performance tests for cognition and mental health, including literacy, numeracy and financial literacy, as well as short and long-term memory, dementia and depression (EURO-D and CESD).

As part of the AMANDA project, Jürges (2005) has used a broad range of more objective physical and mental health measures to construct a health index which maps the combination of these measurements into a standard scale commensurable with the more subjective self-assessed general health rating. Box 3 shows that a correction using this methodology changes both the numerical value and the ranking much more dramatically than the vignette correction. Particularly striking is the distance between Denmark (DK) and Germany (DE). Using self-assessed ratings, Denmark is the top runner while Germany is second to last. Employing the correction using more objective health measurements, Denmark and Germany feature almost identical general health.

### Box 3: Scaling Health Status by Objective Health Measures

**General Health Status using own response scale and scale derived from more objective measures**

<table>
<thead>
<tr>
<th>Overall General Health Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>(Percentage reporting very good or excellent health)</td>
</tr>
<tr>
<td>Own scale</td>
</tr>
<tr>
<td>DK</td>
</tr>
<tr>
<td>SE</td>
</tr>
<tr>
<td>CH</td>
</tr>
<tr>
<td>AT</td>
</tr>
<tr>
<td>GR</td>
</tr>
<tr>
<td>NL</td>
</tr>
<tr>
<td>FR</td>
</tr>
<tr>
<td>DE</td>
</tr>
<tr>
<td>IT</td>
</tr>
<tr>
<td>ES</td>
</tr>
</tbody>
</table>

*Note: AT = Austria, CH = Switzerland, DE = Germany, DK = Denmark, ES = Spain, FR = France, GR = Greece, IT = Italy, NL = Netherlands, SE = Sweden.*

*Source: Jürges (2005)*
10. Conclusions and Outlook

This paper has described which measures the SHARE and AMANDA sister projects have taken to cope with the challenges of cross-national analysis of ageing. Further work will (a) continue and improve the data collection in successor projects to the original SHARE project, and (b) continue and refine the substantive analyses in continuation of the AMANDA project.

From its beginning, SHARE was designed as the baseline of a longitudinal survey. The second wave of data will be collected in the fall of 2006. Further waves are planned every two years. The time dimension is essential because it allows new insights in several respects. Ageing is a process, and not a state. Processes need to be observed over time. Observing two individuals of different age at the same time is no substitute for observing the same person at two ages, since the two persons have been born in different years and thus have experienced other times.

Moreover, the time dimension provides a crucial handle to detect causality. Causality is easiest detected if one can establish that an event happened after the cause. In a single wave, however, a sequence of events is impossible to detect.

The advantages of time variation and cross-national variation are not additive, but multiplicative: Cross-national variation becomes more valuable, when combined with time variation, and vice versa. This can best seen by the way of an example. The European Union is undergoing rapid institutional change. Some countries have enacted dramatic pension reforms. All countries are working on health care reform. A host of incremental labour market reforms is going on. Data with time dimension lets researchers observe the reaction to those changes, e.g. the choice of a later retirement age or higher old-age savings in response to pension reform, different health service utilisation and corresponding health status changes in response to health care reform, and possibly higher labour force participation in response to labour market reforms. Europe with its huge policy diversity represents a ‘natural laboratory’ from which we can learn about the effects of public policy on the behaviour and the well-being of its citizens.
References


Cook, NR, MS Albert, LF Berkman, D Blazer, JO Taylor and CH Hennekens (1995), Interrelationships of peak expiratory flow rate with physical and cognitive function in the elderly, Journals of Gerontology Series A: Biological Sciences and Medical Sciences, Vol 50, Issue 6 M317-M323


Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.


