Netspar

Network for Studies on Pensions, Aging and Retirement

Netspar THESES

Hana Voňková

The Use of Subjective Survey Data

Anchoring Vignettes and Stated Preference Methods

PhD Thesis 2011-009

The Use of Subjective Survey Data: Anchoring Vignettes and Stated Preference Methods

Hana Voňková

The Use of Subjective Survey Data: Anchoring Vignettes and Stated Preference Methods

Proefschrift

ter verkrijging van de graad van doctor aan de Universiteit van Tilburg op gezag van de rector magnificus, prof.dr. Ph. Eijlander, in het openbaar te verdedigen ten overstaan van een door het college voor promoties aangewezen commissie in de aula van de Universiteit op vrijdag 13 mei 2011 om 10:15 uur door

Hana Voňková

geboren op 13 mei 1980 te Tábor, Tsjechië.

Promotores:	prof.dr. A.H.O. van Soest
	prof.dr.ir. A. Kapteyn
Promotiecommissie:	prof.dr. R.J.M. Alessie
	prof.dr. P. Kooreman
	prof.dr. J.P. Smith

Acknowledgements

I would like to thank all the people without whom this thesis would not have come to fruition.

The first person who should be mentioned is without any doubt my great supervisor Arthur van Soest. When I was lost in some problem and needed to organize my ideas he always helped me to understand what I was actually doing. He was always patient and would explain the problem one more time, if necessary. Meetings with my supervisor were intellectually challenging from many points of view. Sometimes he would be falling asleep during our discussions. Perhaps they were not always interesting for him or maybe he thought that I could solve the problems myself; this is still a puzzle for me. I also appreciate his sense of humor. Once David and I received a postcard from him and Josette with one short but important sentence: "Praying for all PhD students." All the prayers were indeed necessary. Although most of us worked hard, our visible progress was sometimes small. Arthur sent us the postcard after I complained that we had sent him many postcards and had not received any in return. I am glad that he was able to deal with my sense of humor.

I was fortunate to have the opportunity to collaborate with an excellent researcher and a charismatic, active, and impatient person, Arie Kapteyn. Our discussions about the anchoring-vignettes method and its assumptions were always fruitful. We either found a solution or came up with new ideas. We also discovered that having discussions while warming milk is not a good idea. Milk can actually boil over. I also appreciate Arie's broad general education. He is not only an expert in economics and econometrics but also knows a great deal about medical science, history, politics, geography, etc. Such universally educated people are becoming rare at universities. I would also like to thank Arie for his support during my stay at RAND Corporation in Santa Monica. It was impressive: he introduced me to RAND people, explained how RAND is organized, helped me to deal with all the formalities, etc. I never hesitated to ask him questions. He has a t-shirt that says "never stop asking" so hopefully he did not mind being bothered so many times. I was also glad that I could stay

in his home in the beautiful and wild Topanga throughout my three months at RAND. I hope that in the future, I will be able to offer this kind of support to my students at Charles University in Prague.

Another outstanding researcher I collaborated with was James P. Smith (just call him Jim) from RAND. Jim combines academic discussions with numerous jokes. Talking to him was both useful and fun. Who would not wish to have such a co-author?

I also worked with my colleague and good friend Patrick Hullegie. He started his PhD at Tilburg University at the same time as I did. During the first year of our study we considered starting a project together. We complement each other well: Patrick is good at writing articles and studying literature, and I could contribute some math and programming. We learned much from each other, wrote an article, and remained good friends.

The collaboration with Arthur, Arie, Jim, and Patrick was simply perfect and I hope that it will continue in the future.

My PhD at Tilburg University was mainly about studying econometrics, programming, and writing articles. But not all of my life was that boring, since I had several free-time activities. The most important of them helped me to live a life outside the university and to understand the Netherlands by reading newspapers and watching TV. I am referring to learning the Dutch language. I would like to thank all the teachers at the Talencentrum at Tilburg University; they organized Dutch-as-a-second-language courses for employees of the university and supported us in this important activity.

PhD students at Tilburg University typically share their office with at least one other student. I shared my office with Chris Müris, a bright and crazy person. Sometimes he worked hectically, sometimes not at all. During these "brain out of office; body inside" times he came up with several interesting challenges, including jumping over his desk without touching it (he didn't manage) and doing a headstand in the middle of the office (he did manage). I thank him for a great time.

During my three-month internship at RAND corporation I shared an office with Luc Bissonnette, another PhD student at Tilburg University, who was visiting RAND at the same time. Every morning when I entered our office, he either told me a joke, tried to insult me in a clever and gentle way, or looked desperate. The reason for his desperation was typically his laptop, which averaged two crashes a day, destroying his work every time. Luc's life (and mine) improved considerably when Arthur helped him to get a new laptop (read: bought him a new laptop). I thank Luc for all the funny moments and I thank Arthur for Luc's

new laptop.

I am grateful to all my other friends at Tilburg University, who improved my mood, cooked for me, and danced with me: Tunga, Amar, Maria, Andrea, Sara, Pavel, Tobias, Otilia, Martin, Kim, Salima, John, Mohammed, Marco, Guillaume, Miguel, Cristian, Jan, Nathanael, Gerard, Fangfang, Ting, Yang, and the Czech students: Honzík Kabátek and Jarda Pazdera.

Finally, I would like to thank my closest family, who for some reason believed in my ability to get a degree from such a demanding school. Thank you Mom, Dad, Grandma, Grandpa, Sister, Uncle Jan, Aunt Zdena, Mom-in-law Jana, Dad-in-law Petr, and Sister-in-law Katka.

My husband David Voňka, a miserable PhD student, deserves no thanks. I met him during my mathematical studies at Charles University in Prague and married him after long consideration. His help with programming and our discussions about math continually slowed me down.

Contents

Chapter 1

Introduction

It's fantastic, incredible, unbelievable
and I might even say . . . ausgezeichnet.

A computer game

This dissertation contains empirical analyses that use subjective survey data. The anchoring vignette method is used in Chapters 2, 3, and 4. In Chapter 5, we use the stated preference method to study the sensitivity of retirement decisions to financial incentives. In the next section we introduce these topics and briefly describe the content of each chapter.

1.1 On anchoring vignettes

In many surveys, people are asked simple and understandable questions. Examples include:

- How much say do you have in getting the government to address issues that interest you?
- Do you have any impairment or health problem that limits the kind or amount of paid work you can do?
- In the last 30 days, how much difficulty did you have in seeing and recognizing from across the road a person you know (i.e., from a distance of about 20 meters)?

Such questions are preferred to other measurements of the concepts for many reasons. For example, some concepts (such as political efficacy or freedom) are too abstract for the average respondent and therefore he/she would not clearly understand an abstract question such as "What is the level of political efficacy in your country?" The assessment of other concepts (such as work disability, quality of health-care system) can be done via a large set of questions. However, this method may be too expensive, and respondents may become bored by the questions and bias their responses. Some concepts (such as visual acuity) can be objectively measured but the cost may be high. A well-chosen example question can summarize a broad area.

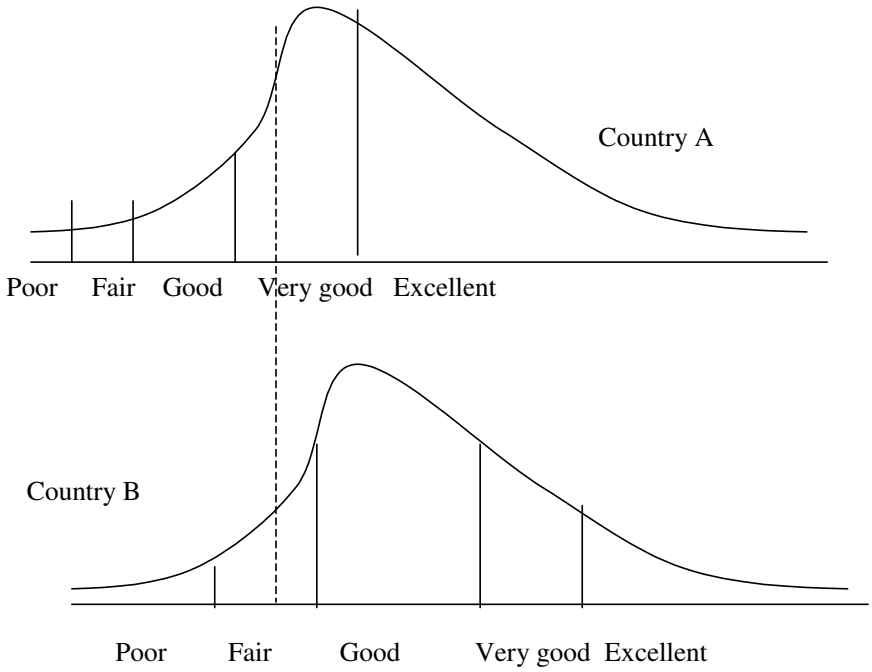
However, the answers to questions may depend not only on the objective situation (concentration and memory) in which we are typically interested but also on the response scales of the respondents. These scales can differ in different countries and/or different socio-economic groups within a country. The anchoring-vignette method was introduced by King *et al.* (2004) as a tool to separately identify the objective situation and the response behavior of the respondents.

Vignette researchers often use the following example (see Kapteyn *et al.* (2007)) to explain the basic idea of the vignette method. Imagine that we would like to compare health in two hypothetical countries. The health distribution in the two countries is depicted in Figure 1.1. The distribution in country A is shifted to the left compared to country B. This implies that people in country A are, on average, less healthy than people in country B.

In our data we do not have a true measure of the health because it would be too expensive to obtain. Instead we have the self-assessed health on a five-point scale (poor, fair, good, very good, excellent) for each individual. But the people in the two countries use different response scales to assess their health. People in country A are more positive about their health status than those in country B. The dashed line in Figure 1.1 represents a person in country A with a given health status who would assess his/her health as very good, whereas a person in country B with the same health status would assess his/her health as fair. Considering only the self-assessments in this hypothetical example, we would conclude that the people in country A have better health than those in country B, but this is an incorrect conclusion. Ignoring the different response scales of the people in the two countries would completely mislead us. Later we will discuss real examples of this problem.

The vignette method provides a solution for this problem. A vignette question gives a scenario describing the health of a hypothetical person and then asks

Figure 1.1: Comparing self-reported health across two countries in case of DIF



Note: The figure is taken from Kapteyn *et al.* (2007).

respondents to evaluate that person's health on the same scale used for a self-report on their own health. Suppose that we construct a vignette describing the health corresponding to the dashed line. People in country A would evaluate the vignette person's health to be very good, whereas people in country B would describe it as fair. Given that the health of the vignette person is the same, the difference in the evaluations between the two countries must be caused by different response scales. Therefore, the vignette evaluations help us to identify the differences between the response scales. Using this information we can adjust the self-assessed health in the two countries. For example, we could choose a scale in country A as the benchmark and express the evaluations in country B on the benchmark scale. We would then conclude that the health in country B is better than that in country A.

If you read this hypothetical example carefully, you might notice that the vignette method would not work without two underlying assumptions. First, people must use the same reporting behavior for both the self-assessments and the vignette evaluations. Suppose that people in country A are not consistent from this point of view and use the scale of country A for their self-assessments and the scale of country B for their vignette evaluations. If we then compare the vignette

evaluations in the two countries, we would conclude that no adjustment of the self-assessments is necessary because the people in the two countries evaluate the vignettes in the same way. If we do not perform any adjustment of the self-assessments we will get an incorrect conclusion as described in the example. The assumption that respondents use the same scale for both self-assessments and vignette evaluations is called response consistency.

Second, each person must interpret the true health status described by the vignette in the same way. In our hypothetical example, in both countries a given health vector must correspond to the same point on the horizontal axis. Suppose that people in country A do not interpret the vignette person's health correctly and interpret the health depicted by the dashed line to be poor (the dashed line would not be vertical but sloped to the left). If there is such a misinterpretation for all the health vectors, we would conclude that people in country A are much healthier than people in country B. The assumption that the concept described in the vignette must be interpreted in the same way by all respondents is called vignette equivalence.

Anchoring vignettes have been applied in various domains. Here are three applications where the vignette method helped to adjust the scale of self-assessments correctly, so that the adjusted self-assessments are on average closer to the objective reality. The first example is an application in political science, specifically the measurement of political efficacy (see King *et al.* (2004)). The second example is an application in health economics, specifically a measurement of work disability (see Kapteyn *et al.* (2007)). The third example is a health application involving the measurement of visual acuity (see King *et al.* (2004)).

Example 1: Political efficacy

King *et al.* (2004) measured political efficacy in China and Mexico. They asked respondents: *How much say do you have in getting the government to address issues that interest you? (1) No say, (2) Little say, (3) Some say, (4) A lot of say, (5) Unlimited say.* A comparison of the raw responses was surprising. The Mexicans judged themselves to have lower levels of political efficacy than the Chinese. However, the actual level of democracy and freedom in these two countries suggests the opposite conclusion.

To explain this apparent paradox, King *et al.* (2004) used the vignette method. They asked respondents to evaluate five vignettes concerning "say in government" on the same scale as the self-assessment question. An example vignette is: *[Imelda] lacks clean drinking water. She and her neighbors are drawing attention to the issue by collecting signatures on a petition. They plan to present*

the petition to each of the political parties before the upcoming election. Using these evaluations they show that the Chinese have lower standards for the level described by any given response category. After adjusting the self-assessments for heterogeneity in reporting behavior they conclude that Chinese have lower levels of political efficacy than Mexicans. Using only the self-assessments would be seriously misleading in this case.

Example 2: Work disability

Kapteyn *et al.* (2007) try to understand why workers in different Western countries report different rates of work disability in contrast to the believed similarity in their "objectively" measured health status. Specifically, they compare two countries: the Netherlands and the US. Respondents in these two countries were asked: *Do you have any impairment or health problem that limits the kind or amount of paid work you can do?* The raw data show that self-reported work disability is much higher in the Netherlands than in the US. In contrast, a comparison of "objectively" measured health conditions such as diabetes, arthritis, hypertension, heart problems, stroke, and emotional problems suggests that the Dutch population is healthier than the US population.

Kapteyn *et al.* (2007) implement a vignette methodology. The respondents were given five vignettes in each of the three domains considered to be the most important determinants of work disability: emotional problems, pain, and heart disease. An example pain vignette is: *[Catherine] suffers from back pain that causes stiffness in her back especially at work but is relieved with low doses of medication. She does not have any pain other than this generalized discomfort.* Using the vignette evaluations, they found strong evidence that American respondents use a tougher standard when assigning a work-disability status. Especially in the more subjective domains, emotions and pain, this heterogeneity in reporting behavior is large. Accounting for the fact that Dutch respondents use lower thresholds than Americans do explains a substantial part of the observed differences in reported work disability.

They also provide evidence that different groups within a country use different thresholds. In the US, they found a separate significant effect of sex, age, and education on the use of response scales. Women, people with low education, and older people are tougher, i.e., they use higher thresholds when evaluating their work-disability status.

Example 3: Visual acuity

King *et al.* (2004) included self-assessment and vignette questions to measure visual acuity on surveys for the World Health Organization in China and Slovakia. The data for China were collected in 2001 and those for Slovakia in 2000.

The vision self-assessment question was: *In the last 30 days, how much difficulty did you have in seeing and recognizing a person you know across the road (i.e., from a distance of about 20 meters)?*, with response categories none, mild, moderate, severe, and extreme/cannot do. Half of the respondents, randomly chosen, evaluated eight vignettes. An example vignette is: *[Angela] needs glasses to read newsprint (and to thread a needle). She can recognize people's faces and pick out details in pictures from 10 meters quite distinctly. She has no problem with seeing in dim light.*

In these surveys the standard test of vision—the Snellen Eye Chart test—was also included. The result of the test was not surprising: the Chinese have considerably worse vision than the Slovaks. In China glasses are not generally available and in general the health care system is inferior to that in Slovakia.

They estimated both an ordered probit model (vignette adjustment is not used) and a parametric model for anchoring vignettes. While the probit model indicated that there is no significant difference in vision between the two countries, the parametric model was in line with the measured test, i.e., the conclusion was that Chinese have significantly worse vision than Slovaks.

The vignette method has been used in many other domains. See Gary King's website http://gking/harvard.edu/vign/eg for lists of vignettes that have been used in specific areas.

In the three examples we showed that the anchoring-vignette method helped to solve the puzzle when self-assessments were not in line with objective measures. However, the question is whether the method is valid in other situations. A formal validation of the method was introduced by Van Soest *et al.* (2011). They validated the method with the use of an objective variable measured for each respondent. Specifically, they collected self-assessments of drinking patterns over the course of a year and four vignettes describing the number of drinks per occasion. They also asked the students to give the number of drinks typically consumed per occasion, which they took to be an objective measure of drinking behavior. To adjust the self-assessments they use all four vignettes together. The results suggest that allowing for heterogeneous reporting behavior substantially improves the fit of the model as well as the correlation between the self-assessments and the objective measure.

Datta Gupta *et al.* (2009) use a similar approach to that of Van Soest *et al.* (2011). They work with data from the first wave of SHARE. Specifically, they use self-assessments and nine vignette evaluations of work disability and grip strength as an objective measure. They find that the DIF-adjusted self-assessments are not more in agreement with the objective information than the unadjusted self-assessments. This result shows that the method does not always help, and the basic question is why not.

In this dissertation we validate the anchoring-vignette method. In Chapter 2 we try to answer the following questions: Do different vignettes within a domain help to bring self-assessments closer to reality? Do we get the same conclusion for different domains? Are these results the same in different years? We use rich datasets from both waves (2004, 2007) of the Survey of Health, Ageing and Retirement in Europe (SHARE) for three health domains: cognition, breathing, and mobility. The main results are the following. For cognition the method is sensitive to the choice of the vignette: one vignette brings the self-assessments closer to the objective situation, while two others do not. When possible the results are found to be consistent over time. The breathing vignette collected in wave 2 brings the self-assessments closer to the objective situation. However, our results based on data from wave 1 indicate that this might be sensitive to the choice of vignette. The most positive results are found for mobility, for which all the vignettes bring the self-assessments closer to the objective situation.

The answers to the previous questions give us an intuition about whether or not the method is valid in general. We conclude that the method does not always help. There are also new questions: Are the main underlying assumptions (response consistency and vignette equivalence of the method in different domains) satisfied? How can we formally test them? Further questions include: Is the commonly used parametric model for anchoring vignettes correctly specified? Are the statistical assumptions of this model satisfied?

In Chapter 3, we try to answer the first set of questions related to the test of underlying assumptions. We test response consistency and discuss vignette equivalence using data from an experiment. Specifically, respondents in an Internet panel are asked to describe their health in a number of domains and to rate their health in those domains. In a subsequent interview respondents are shown vignettes that are in fact descriptions of their own health. Under response consistency and some auxiliary assumptions on the validity of the experiment, there should be no systematic differences between the evaluation of the vignettes in the second interview and the self-evaluations in the first interview. Our non-parametric analysis suggests that response consistency is satisfied for sleep but

not for other health domains. Using a parametric model gives some insight into why this is the case.

In Chapter 4, we try to answer the second set of questions related to the correct specification of the parametric model. We use the chi-square test for a parametric model with covariates, introduced by Andrews (1988). The cells for the tests are here constructed mainly nonparametrically. Specifically, the nonparametric approach to anchoring vignettes is used for the construction of the cells. It does not require any explanatory variables and makes no statistical assumptions. It allows several diagnostic tests of the statistical assumptions of the parametric model. If the parametric model is rejected, the nonparametric approach is still a feasible alternative. We run the tests for six health domains (breathing, cognition, depression, mobility, sleeping, and bodily pains) using data from the Survey of Health, Ageing and Retirement in Europe (SHARE), collected in 2004. P-values of the chi-square test for a parametric model indicate that the use of the random effect in thresholds plays a substantial role. Without the random effect every parametric model is rejected; adding the random effect is already (a modest) part of the solution.

1.2 On stated preferences

Stated-preference (SP) data arise when people state how they would decide in hypothetical situations. SP data are generally good for studying the preferences of people in settings that differ considerably from the current state. The SP method is commonly used in marketing research and transport economics (e.g., Louviere *et al.* (2002)) and is gaining ground in economics (e.g., Barsky *et al.* (1997) or Revelt and Train (1998)). An alternative to SP data is revealed-preference (RP) data. The term "revealed preferences" refers to the preferences that people reveal in real-world settings. RP data are data on the observed actual behavior of individuals. These data are well suited to short-term forecasting of the effects of small departures from the current state of affairs. There are many studies in the economics literature based on RP data, e.g., Lumsdaine and Mitchell (1999) and Kapteyn and De Vos (2004).

In Chapter 5 we use stated-preference data to analyze the preferences of Dutch people for early, late, and gradual retirement. The main reason for using stated rather than revealed preferences is that we want to estimate preferences for pension plans that do not exist or to which many workers do not have access, such as retirement after age 65 or gradual retirement. Moreover, stated-preference data allow for a design where the choice opportunities are exactly known and the

variation in choices is substantial and by construction exogenous to preferences.

The basic idea of our experiment is as follows: Survey respondents aged 25 and older in the Netherlands were given hypothetical retirement scenarios describing the age(s) of (partial and full) retirement and corresponding replacement rates. Several types of retirement trajectories were considered: retirement before, at, or after the standard retirement age (65 years), with and without gradual retirement, and with various replacement rates during partial and full retirement. The data were collected in 2006, 2007, and 2008, partly for the same respondents.

The SP data are used to estimate an intertemporal utility model in which the individual's utility is the discounted sum of within-period utilities that depend on employment status (working, partially retired, or (fully) retired) and income in that period. The parameters of the utility function vary with observed and unobserved respondent characteristics and the year of data collection. The estimated model is used to analyze how retirement preferences differ by background characteristics and how they evolve over the survey years. Simulating the choice of the retirement age under actuarially fair and unfair trade-offs, we then analyze how the preferred retirement age changes if pension-income levels change irrespective of the retirement age (the "(pension) income effect"), or if the pension-benefit accrual induced by delaying retirement changes (the "price" or "substitution effect").

Our experiment shows that financial incentives have a large effect on the preferred retirement age, often even larger than the effects found with revealed preferences, in line with the fact that we allow for flexible choices without imposing restrictions such as mandatory retirement at age 65. Introducing gradual retirement opportunities after the normal retirement age would stimulate participation after age 65. We find that for trade-offs involving gradual retirement, the replacement rate after full retirement is given much more weight than the replacement rate during gradual retirement. Our simulations with choices between actuarially fair retirement scenarios at ages between 60 and 70 show that an increase in lifetime pension incomes by 10 % would lower the average retirement age by 3 months (the "income effect"). Changing the compensation for delaying retirement from actuarially fair to 50 % of what would be actuarially fair would reduce the average retirement age by 9.7 months.

Chapter 2

Do vignette descriptions matter

2.1 Introduction

Survey respondents are commonly asked to self-assess their health, work disability, job/life satisfaction, and other concepts. Consider, for example, the typical survey question that asks respondents to self-assess their health: "Would you say your health is ...," with answers ranging from "very bad" to "very good." Researchers frequently use the answers to these questions to study differences between countries or between groups within a country. When the goal is to draw conclusions about actual differences, the results from direct comparison of self-assessments may be biased if respondents use the response categories in different ways. This interpersonal incomparability is referred to in the literature as differential item functioning (DIF) or as heterogeneity in reporting behavior.

King *et al.* (2004) introduced anchoring vignettes as a tool to correct self-assessments for heterogeneity in reporting behavior. An anchoring vignettes is a short description of aspects of a hypothetical person's life which are relevant to the domain of interest. Application of the idea means that survey respondents not only assess their own situation but also the described situation of person in the vignette. Both situations should be assessed on the same scale. Intuitively, the method can be understood as follows: since the situation described in the vignette is the same for every respondent, vignette evaluations provide information about response styles of respondents. Therefore, we can identify and adjust self-assessments for heterogeneity in reporting behavior. The anchoring vignettes method requires the following two assumptions: (1) response consistency, which is the assumption that individuals use the same reporting behavior for self-assessments and vignettes evaluations; (2) vignette equivalence, which is the assumption that the level of the variable represented in the vignette is

understood in the same way by all respondents.

The vignette method has been applied in different domains like politics (e.g., King *et al.* (2004), Hopkins and King (2010)), health (e.g., Salomon *et al.* (2004), Bago d'Uva *et al.* (2008b)), work disability (e.g., Kapteyn *et al.* (2007)), satisfaction with the health care system (e.g., Murray *et al.* (2003), Sirven *et al.* (2008)). See Gary King's website (http://gking/harvard.edu/vign/eg) for a large collection of vignettes used in different settings.

This chapter studies the validity of the parametric model for the anchoring vignette method, called the CHOPIT model, because it is used most often in applications of anchoring vignettes. In addition to the response consistency and vignette equivalence assumptions, the CHOPIT model makes functional form and distributional assumptions. See Section 2.3 for more details. If all of the assumptions of the model do not hold, we can get wrongly adjusted self-assessments. We do not test the assumptions of the model separately, rather we take the following two approaches to assess the validity of the anchoring vignette method.

First, we study whether different vignettes within a certain domain lead to similar adjusted self-assessments. This idea requires that more than one vignette is collected within a domain, which is the case for the data we use. After estimating the CHOPIT model we compute the correlation coefficient between any pair of DIF-adjusted self-assessments (each adjusted using a single vignette) within a domain. If different vignettes would lead to similar adjusted self-assessments then the correlation coefficient between any pair of DIF-adjusted self-assessments would be close to one. As far as we are aware this approach has not been taken before.

This first approach is uninformative about the question whether different DIF-adjusted self-assessments are closer to the actual situation than unadjusted self-assessments. That is what we study in the second approach, details of which are discussed in Section 2.3. Assessing the validity of the anchoring vignette method by means of a measure of the actual situation has been suggested before in the literature. In fact, we follow Van Soest *et al.* (2011). The novelty of our approach is that we study the performance of a single vignette, as we did in the first approach. The credibility of this approach is closely connected with the quality of the chosen objective measure(s) as well as with the assumptions of the CHOPIT model. The quality of the objective measure(s) depends on how closely they correspond with the health dimensions elicited in the self-assessment and vignette questions. If the correspondence is strong, then the results of this approach show whether or not the DIF-adjusted self-assessments are closer or not to the actual situation than the unadjusted self-assessments. If the correspondence is weak,

then the results of this approach may not be valid. In that case, or in the case that there is no measure of the actual situation available, results of our first approach will still indicate whether or not the vignette method is sensitive to the choice of the vignette (as long as more than 1 vignette is collected).

Studying whether the method is sensitive to the choice of the vignette and studying whether DIF-adjusted self-assessments are closer to an objective measure than unadjusted self-assessment are both important issues, because researchers who apply the method should be confident that the method works properly.

The first comparison of unadjusted and DIF-adjusted self-assessments with a measure of the actual situation is, as far as we are aware, reported by King *et al.* (2004). They use self-assessments and vignette evaluations on visual acuity collected by the WHO for China and Slovakia. On average, the self-assessments do not show a significant difference in visual acuity between Chinese and Slovak respondents. However the measured test for vision - the Snellen Eye Chart test - shows that respondents from China have, on average, substantially worse vision than those from Slovakia. Self-assessments are adjusted using the eight vignette evaluations simultaneously. Comparison of these DIF-adjusted self-assessments confirms the conclusions from the measured test.

Van Soest *et al.* (2011) propose a formal test for the response consistency assumption. Additionally, they show whether the distribution of DIF-adjusted self-assessments is "closer" to the distribution of an objective measure than the unadjusted distribution. They collected self-assessments and four vignette evaluations on drinking behavior among students at a large Irish university. Additionally, they collected a measure of actual drinking behavior: the self-reported number of drinks typically consumed per occasion. Self-assessments are adjusted using all four vignettes simultaneously. Their results suggest that allowing and adjusting for heterogeneous reporting behavior, as well as assuming response consistency, substantially improves the fit of the model as well as the correlation between the self-assessments and objective measure.

Datta Gupta *et al.* (2009) take a similar approach as Van Soest *et al.* (2011), using data from the first wave of Survey of Health Ageing and Retirement in Europe (SHARE). The paper focuses on work disability. Self-assessments are adjusted using all nine vignette evaluations simultaneously and grip strength is used as an objective measure. Their finding is that DIF-adjusted self-assessments are not closer to the objective measure than the unadjusted self-assessments.

Using data from both waves (2004 and 2007) of SHARE we study the validity of the vignette method for three domains not studied before: cognition, breathing

and mobility. SHARE collected data on self-assessments, vignette questions and objective measures for the three domains studied in this chapter. More details about the data collected in both waves is given in Section 2.2.

For cognition we find that different vignettes lead to different adjusted self-assessments. One vignette brings the self-assessment closer to a measure of actual cognition, while two others do not. For breathing we find that different vignettes lead to different adjusted self-assessments. However, the vignette collected in the 2007 wave brings the self-assessment closer to a measure of actual breathing. Our findings for mobility are most encouraging. Here, all vignettes bring the self-assessment closer to a measure of actual mobility.

The remainder of this chapter is organized as follows. Section 2.2 provides more information on the SHARE data. The CHOPIT model as well as our two approaches to validate the vignette method are discussed in Section 2.3. Results are discussed in Section 2.4 and Section 2.5 concludes.

2.2 Data

The chapter uses data from both waves of the Survey of Health, Ageing and Retirement in Europe (SHARE), a nationally representative sample of the population 50 and older, which provides detailed information on health, socioeconomic status, and social and family networks of more than 45,000 individuals. In 2004 data for the first wave were collected in eleven European countries. In 2006-07 data for the second wave were collected in the same eleven countries and three new countries.

For each of the three health domains we focus on, three vignette questions were collected in the first wave. By contrast, the second wave collected one vignette per domain, which was chosen out of the three from the first wave. In both waves the data on self-assessments and vignette questions are only collected in subsamples of the overall SHARE samples. In the remainder of this chapter we will refer to these subsamples, which are different for both waves, as the vignette samples. Self-assessments and vignette evaluations were collected in both waves for Belgium, France, Germany, Greece, Italy, the Netherlands, Spain and Sweden, and only in the second wave for the Czech Republic, Denmark and Poland.[1] The SHARE data also contain information on objective measures for all three domains. The objective measures for cognition are available in both waves, whereas those for breathing and mobility only in the second wave. Moreover,

[1] In Greece, self-assessments and vignette evaluations were collected in both waves, but the data of the second wave were not available in the release we use.

only respondents younger than 75 were asked to participate in the objective measurement task for mobility.

To keep as much information as possible we use a different sample for each (domain,wave) combination. That is, we have a sample for cognition for wave 1, and another sample for cognition for wave 2. This is also the case for the other two domains. Each (domain, wave) sample is selected on the self-assessment and vignette(s) available for that combination, on available objective measure(s), and on a set of common covariates. The samples are generally distinct.[2]

Descriptive statistics for the self-assessments, vignette evaluations and objective measures, as described below, are based on the relevant (domain,wave) sample. Descriptive statistics of the covariates are based on the vignette samples of both waves. As it turns out only for the mobility sample of the second wave the distribution of covariates is substantially different from the one for the vignette sample of the wave 2. Below we discuss on which aspects it differs.

2.2.1 Self-assessments and vignette ratings

To begin, consider an example of a self-assessment question for concentration: "Overall in the last 30 days how much difficulty did you have with concentrating or remembering things?". Self-assessment questions for other two domains studied in this chapter, breathing and mobility, can be, for example, found in Appendix 4.B. In all cases the possible answer categories are 'none', 'mild', 'moderate', 'severe', and 'extreme'.

As noted already, the vignette collected in the second wave of SHARE was chosen out the three vignettes collected in the first wave. Each vignette describes aspects of a hypothetical person's life relevant to the domain. The exact wording of all three anchoring vignettes collected for all three domains can be, for example, found in Appendix 4.B.[3]

The percentage of missing observations for the self-assessments and the vi-

[2]We test whether DIF-adjusted self-assessments are closer to the actual situation than unadjusted self-assessments. Theoretically, if the DIF-adjustment of self-assessment helps in the whole population, then it also helps in any sample of the population. If the adjustment helps in a sample of the population (in particular a nonrepresentative sample), it is only an indication that it may help in the whole population. This should be taken into account while working with different samples.

[3]To distinguish vignettes for specific domains, vignettes for breathing are labeled b_1, b_2, b_3, for concentration c_1, c_2 and c_3 and for mobility m_1, m_2 and m_3 in this chapter. These labels correspond to labels v_1, v_2 and v_3 introduced in Appendix 4.B. Health problems of hypothetical person described in vignette v_1 are mild. In vignette v_2 a person has more health problems than in vignette v_2. Vignette v_3 describes the most extreme health problems among all the three vignettes.

gnette questions is very low, it is at most 1.9 percent. It is also very similar across countries. Descriptive statistics for the self-assessments and vignette evaluations are given in Tables 2.1, 2.2 and 2.3 for wave 1 and in Table 2.4 for wave 2.

In all cases, most respondents report to have either no or only a mild problem. Few respondents report to have a severe or extreme problem. From the vignette evaluations of wave 1 it becomes clear that, within each domain, the vignette numbered "1" is, on average, considered to be the vignette describing the most mild problem within that domain. The vignette numbered "2" describes, on average, a more severe problem than the vignette numbered "1", and the vignette numbered numbered "3" described, on average, the most extreme health problem.

In our empirical analyses, we always combine the two categories "severe" and "extreme" for both self-assessments and vignette evaluations, because especially in the latter category there are few observations.

2.2.2 Objective measures

One of the two validation approaches taken in this chapter studies whether DIF-adjusted self-assessments are closer to a measure of the actual situation than unadjusted self-assessments. The measures we use are discussed below for each domain.

Cognition

For cognition we use the following four objective measures: immediate and delayed verbal memory, verbal fluency and numerical ability.[4] Specifically, respondents were asked to do the following tasks:

- Immediate recall: Now, I am going to read a list of words from my computer screen. We have purposely made the list long so it will be difficult for anyone to recall all the words. Most people recall just a few. Please listen carefully, as the set of words cannot be repeated. When I have finished, I will ask you to recall aloud as many of the words as you can, in any order. Is this clear?

- Delayed recall: A little while ago, I read you a list of words and you repeated the ones you could remember. Please tell me any of the words

[4]The variables used to measure cognitive functioning in SHARE are very similar to those used in other, well-known, surveys such as the English Longitudinal Study of Ageing (ELSA), the Asset and Health Dynamics Among the Oldest Old (AHEAD) study and the Health and Retirement Study (HRS). See e.g., Mehta *et al.* (2003), Llewellyn *et al.* (2008), Herzog and Wallace (1997).

you can remember now? (This question is directly asked after the final question assessing numerical ability.

- Verbal fluency: Now, I would like you to name as many different animals as you can think of. You have one minute to do this.

- Numeracy: Next, I would like to ask you some questions which assess how people use numbers in everyday life.

  1) If the chance of getting a disease is 10 per cent, how many people out of 1,000 would be expected to get the disease?

  2) If the respondent's answer to question 1 was incorrect, the next question is: In a sale, a shop is selling all items at half price. Before the sale, a sofa costs 300 (local currency). How much will its cost in the sale?

  3) If the respondent's answer to question 1 was correct, the next question is: A second hand car dealer is selling a car for 6,000 (local currency). This is two-thirds of what it costs new. How much did the car costs new?

  4) If the answer to question 3 is correct, the next question will be: Let us say you have 2,000 (local currency) in a savings account. The accounts earns ten per cent interest each year. How much will you have in the account at the end of two years?

  The algorithm SHARE uses to compute the score for numeracy is as follows. Every respondents is asked the first question. If s/he gives the correct answer the score for numeracy equals 3 and the next question is question 3. If question 3 is answered correctly, the score for numeracy becomes 4 and the next question is question 4. If the respondent answers question 4 correctly, then the score for numeracy becomes 5. If the first question is answered incorrectly, then the score for numeracy equals 1 and the next question is question 2. If this question is answered correctly the score for numeracy will become 2.

All objective measures described above are available in both waves. Whereas immediate and delayed recall seem to be closely related to the cognition question, which asks about "concentrating and remembering things," numeracy and verbal fluency seem to be less related. Still we included them into our analyses to see whether the results would be the same. Anticipating our results, we find that the

conclusions are the same irrespective of the objective measure except for verbal fluency for wave 1.

The percentage of missing observations for the four objective measures of cognition is very low. In both waves most respondents are able to immediately recall 4 words or more, however few respondents recall more than 7 words immediately. As expected, after a short delay, respondents recall fewer words. Most of them recall 2 up to 5 words after a short delay. The median score for numeracy is 3 in both waves. The number of animals respondents can mention in one minute is 19 on average.

Table 2.5 shows descriptive statistics by country for delayed recall information collected in the second wave. It reveals that in Italy and Spain respondents recall, on average, relatively few words after a short delay, whereas in Denmark, the Netherlands and Sweden respondents recall relatively many words after a delay.

In our empirical analyses, immediate recall is coded into 5 different categories. Delayed recall and numeracy into 4 categories, and verbal fluency (the number of animals mentioned) into 5 different groups.[5]

Breathing As objective measure for breathing we use the result of a so-called peak flow test, which measures how fast respondents can exhale while breathing out as hard and fast as possible.[6] Specifically, respondents were asked to do the following task:

The next test that I am going to ask you to perform will measure how fast you can expel air from your lungs. It is important that you blow as hard and as fast as you can. I would like you to perform the test two times. When we are ready to begin, I will ask you to stand up. Take as deep a breath as possible. Open your mouth and close your lips firmly around the outside of the mouthpiece, and then blow as hard and as fast as you can into the mouthpiece.

If two measurements are available for a respondent we take the maximum value, otherwise we take the single measurement available as value. The unit of measurement of the peak flow test is liters/minute and it ranges from 60 to 880.[7] Note that this measure is only available for the second wave.

[5]We merge the categories containing few observations.

[6]This test has been widely used in clinical practice to assess airflow obstruction and for monitoring patients with asthma Nunn and Gregg (1989),Quanjer *et al.* (1997).

[7]Interviewers were instructed to record a value of 30 if the respondent's measurement was less than 60, and to record a value of 890 if the respondent exhaled more than 880 liters per minute. In our study we observe this for a negligible (less than 2 percent) number of observations.

The percentage of missing observations for the peak flow test is 7.3. The most important reason why an observation is missing is that the respondent thinks it is not safe to do the test. Here there are country differences. Whereas in most countries the percentage of respondents who think it is not safe to do the test is 7 percent or lower, in France and Italy it is 15 percent. Descriptive statistics for the peak flow test are provided in Table 2.6. It discloses that in Italy and Spain the lung capacity is, on average, relatively low, and that is relatively high, on average, in Denmark, the Netherlands and Sweden.

Mobility

As objective measure for mobility we use the result of a so-called stand-up test:

The next test measures the strength and endurance in your legs. I would like you to fold your arms across your chest and sit so that your feet are on the floor; then stand up keeping your arms folded across your chest. The respondent is then asked whether s/he thinks it is safe to stand up five from a chair five times without using their arms. When the answer is affirmative the respondent is asked to do the test and the interviewer records the time (in seconds) used for five stands.

Our measure is the time in seconds needed for five stands.[8]

The percentage of missing observations for the stand-up test is 18.4. This percentage is computed for the group of respondents younger than 75, as only they are asked to participate. Approximately 82 percent of the respondents of the vignette sample of wave 2 is younger than 75. The most important reason that an observation is missing is either that the respondent thinks it is not safe to do a single test, or that s/he is not able to do the single test according to instructions (having their arms fold across their chest), or because the respondent thinks it is not safe to stand up five times. Here, there are also country differences. Whereas on average the percentage of missing observations is around 18 percent, in France it is 28 percent and in Italy it is 36 percent.

Descriptive statistics for the stand up test are given in Table 2.6. On average, respondents need 11 seconds to finish the test. Respondents in Denmark and Sweden are, on average, relatively fast, whereas respondents in Belgium and the Netherlands are, on average, relatively slow.

[8]If the test was not completed within one minute we only observe that the respondent needed more than one minute and not the exact time. In our study we observe this for a negligible (0.5 percent) number of observations.

2.2.3 Covariates

The parametric version of the anchoring vignette method models the actual level of health and reporting heterogeneity using a vector of covariates. The model will be discussed in more detail in Section 2.3. In this chapter we include the following covariates, which are commonly used in applications of this model to health: country, age in groups of 5 years, gender, low/mid/high education, living alone, suffering from a long-term illness, never/sometimes/often engaged in physical activity.

Further information regarding the "construction" of our covariates can be found in Table 2.7.

Descriptive statistics of the covariates are given in Table 2.8. This table reveals that all (domain,wave) specific samples, except the one for mobility in wave 2, are similar to the corresponding vignette samples. The mobility sample contains fewer observations, respondents are slightly better educated, live less often alone suffer less often from a long-term illness, are more often engaged in physical activity, and are on average younger. The different composition of this sample is likely to be due to the selection rules for the objective measure (stand-up test).

2.3 Model

The approach taken in this chapter to validate the anchoring vignettes method requires the availability of an objective measure. We follow Van Soest *et al.* (2011), who extend the compound hierarchical ordered probit (CHOPIT) model, by also modeling the objective measure.

2.3.1 Model for self-assessments

The self-assessment, y_{si}, of individual i is modeled as an ordered response equation with latent variable

$$y^*_{si} = \mathbf{x}'_i\beta_s + \varepsilon_{si},$$

where $\mathbf{x}_i$ is a vector of covariates including a constant term, and β_s a vector of parameters. The error term, ε_{si}, is assumed to be normally distributed with mean zero and variance σ^2_s, and independent of the covariates $\mathbf{x}_i$. The reported and observed responses, y_{si}, are generated by the following mechanism

$$y_{si} = k \Leftrightarrow \tau^{k-1}_{si} < y^*_{si} \leq \tau^k_{si}, \qquad k = 1, \ldots, K$$

where $-\infty = \tau_{si}^0 < \tau_{si}^1 < \ldots < \tau_{si}^K = +\infty$. The thresholds are modeled as

$$\begin{aligned}\tau_{si}^1 &= \mathbf{x}_i'\gamma_s^1 + u_i,\\ \tau_{si}^k &= \tau_{si}^{k-1} + \exp\left(\mathbf{x}_i'\gamma_s^k\right), \quad k = 2, \ldots, K-1,\end{aligned}$$

where $\mathbf{x}_i$ is a vector of covariates, and γ_s^k, for $k = 1, \ldots, (K-1)$, are vectors of parameters. The random effect, u_i, is assumed to be normally distributed with mean zero and variance σ_u^2, and independent of the covariates $\mathbf{x}_i$.

The idea that reporting behavior varies across individuals is formalized by modeling the thresholds to be individual-specific. The latent variable, y_{si}^*, can be interpreted as the true level of health as perceived by the individual. Note that using only self-assessments, the parameter vectors β_s and γ_s^1 are not separately identified, but the parameter vectors γ_s^k, for $k > 2$, are. That is, using only self-assessments we are not able to "decompose" the self-assessments in a part that is due to differences in "true" health (β) and a part due to heterogeneity in reporting behavior ($\gamma_s^k,\ k = 1, \ldots, K-1$).

2.3.2 Model for vignettes

Although in SHARE wave 1 three vignettes were collected for each domain, we only estimate models using a single vignette at a time. The reason is that we want to study whether the method is sensitive to the choice of the vignette. The discussion below is based on the availability of a single vignette.

Under the assumption that there is an actual level of health, ϑ, associated with the hypothetical person described in the vignette, vignettes can be used to correct self-assessments for heterogeneity in reporting behavior. The assumption that the actual level of health of the hypothetical person described in the vignette is the same for every individual formalizes the vignette equivalence assumption. Each respondent perceives the actual level of health only with random error, i.e.,

$$y_{vi}^* = \vartheta + \varepsilon_{vi},$$

where the error term, ε_{vi} is assumed to be normally distributed with mean zero and variance σ_v^2 and independent of the covariates $\mathbf{x}_i$. The observed vignette evaluations are generated by the following mechanism

$$y_{vi} = k \Leftrightarrow \tau_{vi}^{k-1} < y_{vi}^* \leq \tau_{vi}^k, \qquad k = 1, \ldots, K$$

where $-\infty = \tau_{vi}^0 < \tau_{vi}^1 < \ldots < \tau_{vi}^K = +\infty$. The thresholds are modeled similarly

as in the self-assessment model, i.e.,

$$\begin{aligned}\tau_{vi}^1 &= \mathbf{x}_i'\gamma_v^1 + u_i,\\ \tau_{vi}^k &= \tau_{vi}^{k-1} + \exp\left(\mathbf{x}_i'\gamma_v^k\right), \qquad k = 2,\ldots,(K-1),\end{aligned}$$

where the term u_i is assumed to be the same in the thresholds of the self-assessment and vignette model. It introduces unobserved individual heterogeneity and implies that the vignette evaluation is correlated with the self-assessment (conditional on the covariates $\mathbf{x}_i$).

The response consistency assumption is formalized by assuming: $\tau_{si}^k = \tau_{vi}^k$, for $k = 1,\ldots,(K-1)$. In terms of the parameters this amounts to assuming $\gamma_s^k = \gamma_v^k$, for $k = 1,\ldots,(K-1)$.

2.3.3 Model for objective measure

To study whether the anchoring vignette method brings self-assessments closer to the objective situation we make use of measures of the objective situation. The four objective measures for cognition: immediate and delayed verbal memory, numeracy and verbal fluency, are all discrete variables. In that case we model our objective measure as follows

$$\begin{aligned}y_{oi}^* &= \mathbf{x}_i'\beta_o + \varepsilon_{oi},\\ y_{oi} &= l \Leftrightarrow \tau_o^{l-1} < y_{oi}^* \leq \tau_o^l,\end{aligned}$$

where $-\infty = \tau_o^0 < \tau_o^1 < \ldots < \tau_o^L = +\infty$, are unknown thresholds that are the same for all individuals. The thresholds are modeled as

$$\begin{aligned}\tau_o^1 &= \exp\left(\gamma_o^1\right),\\ \tau_o^l &= \tau_o^{l-1} + \exp\left(\gamma_o^l\right), \qquad l = 2,\ldots,(L-1).\end{aligned}$$

The objective measures for breathing and mobility, the result of the peak flow test and the stand-up test, respectively, are continuous variables, and are modeled as follows:[9]

$$y_{oi} = \mathbf{x}_i'\beta_o + \varepsilon_{oi}.$$

In both cases, discrete and continuous, the error term ε_{oi} is assumed to be independent of the covariates, $\mathbf{x}_i$, the unobserved heterogeneity term, u_i, and the

[9]Note that the objective measures of breathing and mobility are both, in principle, affected by censoring. However, recall that the number of censored observations is negligible and therefore we do not model the censoring.

error term of the vignette model, ε_{vi}. However, ε_{oi} is allowed to be correlated with the error term of the self-assessment model, ε_{si}, because the covariates might not capture all variation in "true" health, y^*_{si} and y^*_{oi}. The distribution of $(\varepsilon_{si}, \varepsilon_{oi})$ is assumed to be bivariate normal with mean zero, variances σ^2_s and σ^2_o, and correlation ρ.

2.3.4 Likelihood

The likelihood contribution of each individual i conditional on the unobserved heterogeneity, u_i, can be written as the product of a joint normal probability for the self-assessment and the objective measure, and a single normal probability for the vignette. In case of a discrete objective measure, the unconditional likelihood contribution for individual i is given by

$$\int \prod_{k=1}^{K} \prod_{l=1}^{L} \prod_{m=1}^{M} P(y_{si}=k, y_{oi}=l|\varphi, u_i)^{\mathbb{I}(y_{si}=k, y_{oi}=l)} P(y_{vi}=m|\varphi, u_i)^{\mathbb{I}(y_{vi}=m)} f(u_i)\, du_i \tag{2.1}$$

where $f(.)$ is the normal density function with variance σ^2_u, and $\mathbb{I}(.)$ the indicator function. The vector of parameters is $\varphi = (\beta', \sigma^2_s, \gamma'_s, \sigma^2_u, \vartheta, \gamma'_v, \sigma^2_v, \beta'_o, \tau'_o, \sigma^2_o, \rho)'$.

2.3.5 Identification

We use three different models to study whether the vignette method is sensitive to the domain and the choice of the vignette, and is consistent over time. Each of these three models can be considered as a "special case" of the model discussed in the previous subsections, and every model has a different set of identifying assumptions.

CHOPIT model King *et al.* (2004) have introduced the CHOPIT model, which combines the self-assessment and vignette part of the model discussed before, i.e. the objective part is not included. For identification reasons the constant term of the β_s equals zero and the variance of error term in the self-assessment part is normalized to one, i.e., $\beta_{s,1} = 0, \sigma^2_s = 1$. Because of response consistency we assume that $\gamma^k_s = \gamma^k_v$ for $k = 1, \ldots, (K-1)$.

Model A (No DIF, No RC) Reporting behavior is assumed to be homogeneous in this model, therefore $\tau^k_{si} = \tau^k_s$ and $\tau^k_{vi} = \tau^k_v$, for $k = 1, \ldots, (K-1)$, and $\sigma^2_u = 0$. The model does not impose response consistency, that is, it allows for the possibility that $\tau^k_s \neq \tau^k_v$ for $k = 1, \ldots, (K-1)$. However, for identification

reasons $\tau_s^1 = \tau_v^1 = 1$. The variances of the error terms in both the self-assessment and vignette model are normalized to one, i.e., $\sigma_s^2 = \sigma_v^2 = 1$. In case of an discrete objective measure, the first two thresholds of the objective measure are equal to one and two, i.e., $\tau_o^1 = 1, \tau_o^2 = 2$, for identification reasons.

Model B (DIF, RC) This model assumes that reporting behavior is heterogeneous across individuals and therefore the thresholds are individual-specific. In addition it assumes that response consistency holds, i.e., $\gamma_{si}^k = \gamma_{vi}^k$, for $k = 1, \ldots, (K-1)$. Furthermore the model normalizes the constant term in the parameter vector of the first threshold to one, i.e., $\gamma_{s,1}^1 = \gamma_{v,1}^1 = 1$. The variance of the error term in the self-assessment model is normalized to one, $\sigma_s^2 = 1$. In case of an discrete objective measure, the first two thresholds of the objective measure are equal to one and two, i.e., $\tau_o^1 = 1, \tau_o^2 = 2$, for identification reasons.

2.3.6 Two validation approaches

As already discussed in the introduction, this chapter takes two approaches to validate the parametric model for anchoring vignettes. Here we explain our approaches in more detail.

As a first step in validating the vignette method, we investigate whether different vignettes lead to similar DIF-adjusted self-assessments. For each domain we estimate the CHOPIT model using one vignette at a time and compute the DIF-adjusted self-assessments. That is, we compute the predicted systematic parts: $\hat{y}_{si}^* = \mathbf{x}_i' \hat{\beta}_s$. Since three vignettes were collected for each health domain in the first wave, this gives us a set of three different DIF-adjusted self-assessments. Then we compute the correlation coefficient between any pair of DIF-adjusted self-assessments within each domain. If different vignettes would lead to similar DIF-adjusted self-assessments the correlation coefficient between any pair of DIF-adjusted self-assessments (each based on a single vignette) would be close to one.

These correlations are, however, uninformative about the question whether DIF-adjusted self-assessments are "closer" to the objective situation than unadjusted self-assessments. As a second step we therefore estimate the models A and B discussed in the previous section. Model A does not allow (and adjust) for heterogeneity in reporting behavior, whereas model B does. The models are estimated for each domain separately using one vignette at a time. Each time we compute the correlation coefficients between the predicted systematic parts of

(y^*_{si}, y^*_{oi}) and between the simulated values of (y^*_{si}, y^*_{oi}).[10] If the DIF-adjustment would bring the self-assessments closer to the actual situation, then the correlation coefficient given by model B would be higher than the corresponding one given by model A.

2.4 Results

Cognition First, we report the correlations between different DIF-adjusted self-assessments, each based on one vignette. See Table 2.9. The correlations indicate that vignettes c2 and c3 lead to similar DIF-adjusted self-assessments, whereas those based on vignette c1 are different from the other two. All this reveals is that for cognition the DIF-adjustment is sensitive to the choice of the vignette.

Second, we estimate the models A and B, separately using data from wave 1 and wave 2. For wave 1, the models are estimated for each combination of one of the four objective measures and one of the three vignettes. For wave 2, the models are estimated for each of the four objective measures using the single vignette that is collected. Table 2.10 and 2.11 provide a summary of results based on data from wave 1 and wave 2, respectively.

Consider first the results for wave 1. We only discuss them for delayed recall, as they are consistent with the other objective measures except verbal fluency.[11] In case vignette c1 is used the results show that the model that corrects for reporting behavior heterogeneity (model B) gives a correlation between the predicted systematic parts of (y^*_{si}, y^*_{oi}) of 0.52 compared to around 0.76 for the model that does not make this correction (model A). The correlations between the simulated values of (y^*_{si}, y^*_{oi}) are 0.22 and 0.26 for model B and A respectively. So, both correlation coefficients are lower for model B than for model A. We therefore conclude that DIF-adjusted self-assessments based on vignette c1 are more different from the objective situation than the unadjusted self-assessments.

Adjusting self-assessments using vignette c2 leads to a different conclusion.

[10] The predicted systematic parts are: $\hat{y}^*_{si} = \mathbf{x}'_i\hat{\beta}_s$ and $\hat{y}^*_{oi} = \mathbf{x}'_i\hat{\beta}_o$. The simulated values are obtained as follows: using the estimates of σ^2_s, σ^2_o, and ρ we simulate values from the bivariate normal distribution. These simulated values of ε_{si} and ε_{oi} are then added to the predicted systematic parts to obtain simulated values for y^*_{si}, and y^*_{si}.

[11] All results for cognition, for both waves, are consistent with each other, with the exception of the results for verbal fluency using wave 1 data. In that case results indicate that for all three vignettes, DIF-adjusted self-assessments are "closer" to the objective situation than the unadjusted self-assessments. We found that these results are sensitive to the inclusion of observations from Greece, as leaving out those observations gives results that are in line with those reported for the other objective measures in both waves.

For both models the correlation between the predicted systematic parts is comparable; 0.75 for model A and 0.74 for model B. The correlation between the simulated values increases from 0.26 for model A to 0.30 for model B. On the basis of these correlations we conclude that the DIF-adjusted self-assessments based on vignette c2 are about as close to the objective situation as the unadjusted self-assessments.

Consider next the results when vignette c3 is used. The correlation between the predicted systematic parts increases from 0.76 for model A to 0.83 for model B, and the correlation between the simulated values increases from around 0.26 for model A to 0.34 for model B. Here we conclude that the DIF-adjusted self-assessments based on vignette c3 are closer to the objective measure than the unadjusted self-assessments are.

Finally, consider the results based on data from wave 2. Since the vignette collected in the second wave is chosen out of the three from the first wave, we can study whether the results are consistent over time. Vignette c1 is the vignette collected in both waves. If the results are consistent over time we expect to conclude that the vignette method for vignette c1 does not help. The first set of results provided in Table 2.11 are for cognition using data from wave 2. Here the results are consistent across the four objective measures and lead to the same conclusion as before: DIF-adjusted self-assessments based on vignette c1 are more different from the objective situation than the unadjusted self-assessments. So, the results for vignette c1 are found to be consistent over time.

To summarize, our results reveal that for cognition the vignette method is sensitive to the choice of the vignette.

Table 2.12 gives a selection of parameter estimates of model A and B, estimated using data from the second wave. The differences in the correlations are for an important part caused by the country dummies and gender dummy, as for these variables either of the two following cases occurs relatively often: (1) one of the two parameter estimates is significantly different from zero, while the other is not; (2) both are significantly different from zero, but with opposite signs.

Breathing First, we discuss the correlations between different DIF-adjusted self-assessments using data from wave 1. These correlations are reported in Table 2.9, and they show that the vignettes b2 and b3 lead to similar DIF-adjusted self-assessments. However, they are very different from the one based on vignette b1. Thus, we conclude that for breathing the DIF-adjustments are sensitive to the choice of the vignette.

Second, we investigate whether the DIF-adjusted self-assessments are closer

to the objective variable than the unadjusted self-assessments. We do this using data from wave 2, for which an objective measure is available and vignette b1 is collected. For wave 1 there is no objective measure available. The results of the models A and B are reported in Table 2.11.

The correlation coefficient between the predicted systematic parts of (y^*_{si}, y^*_{oi}) equals 0.45 for model A and 0.53 for model B. The reason for the low correlations between the self-assessment and objective measure is that the parameter estimates of certain country and age dummies and the gender dummy show the same discrepancy as described earlier for cognition. The correlation coefficient between the simulated values of (y^*_{si}, y^*_{oi}) equals 0.25 for model A and 0.27 for model B. Although perhaps low, both correlation coefficients still increase when the self-assessments are adjusted for heterogeneity in reporting behavior. So, correcting for reporting behavior heterogeneity brings the self-assessments of wave 2 closer to the objective situation. Parameter estimates of the models are given in Table 2.13.

Mobility Finally, consider the results for mobility. We first give the correlations between different DIF-adjusted self-assessments using one vignette at a time and data from wave 1. Table 2.9 reports the correlations, which are high and approximately the same. Therefore, if the method works for one of the vignettes it is likely that it will work for the other two vignettes as well.

The results of the models A and B based on data from wave 2, for which an objective measure is available and vignette m1 is collected, are given in Table 2.11. The correlation between the predicted systematic parts of (y^*_{si}, y^*_{oi}) increases from 0.54 (model A) to 0.63 (model B), and the correlation between the simulated values of (y^*_{si}, y^*_{oi}) increases from 0.18 (model A) to 0.20 (model B). So, both correlation coefficients increase when allowing for heterogeneity in reporting behavior. Based on data from wave 2 we conclude that the vignette method helps for mobility.

Parameter estimates of the models are given in Table 2.14.

2.5 Conclusion

This chapter takes two approaches to validate the parametric model for anchoring vignettes. First, we study whether different vignettes lead to similar DIF-adjusted self-assessments. Second, we study whether DIF-adjusted self-assessments are closer to a measure of the actual situation than unadjusted self-assessments. Here, we also look at the performance of a single vignette.

We use SHARE data and focus on three different domains of health: cognition, breathing and mobility.

Our results show that the method is sensitive to the choice of the vignette for cognition: DIF-adjusted self-assessments based on vignette c1 are more different from the objective situation than unadjusted self-assessments; for vignette c2 we conclude that the vignette method does not bring the self-assessments closer to the objective situation; the conclusions for vignette c3 is that the self-assessments are brought closer to the objective situation. Vignette c1, which is collected in both waves of SHARE, leads to conclusions that are consistent over time. The conclusions for cognition are the same irrespective of the objective measure used, except verbal fluency for wave 1.

For the breathing vignette collected in wave 2, vignette b1, we find that DIF-adjusted self-assessments are closer to the measure for breathing than the unadjusted self-assessments. However, our results also show that there is no guarantee that it would work with one of the two other breathing vignettes collected in the first wave.

Results are most encouraging for mobility. Adjusting the self-assessments using the vignette collected in wave 2, vignette m1, brings them closer to the measure for mobility. Moreover, the vignette method is unlikely to be sensitive to the choice of the vignettes used in wave 1.

Although our results indicate that the vignette method is sensitive to the domain and choice of the vignette, this should not be taken as a reason to reject this method. Here are several ideas for future research.

First, for the cognition domain we found that different vignettes lead to different results. The vignette describing a hypothetical person with the most extreme cognitive problems (vignette c3) brings the DIF-adjusted self-assessments closer to the objective situation than the other two vignettes describing milder problems. This suggests that the level of health of the vignette person matters, at least in this case. More research should be done to find out, not only, how the level of health of the vignette person matters, but also how to formulate vignettes in general.

Second, our results show that the vignette method is sensitive to the choice of the vignette, at least for the domains of cognition and breathing. The reason may be that the CHOPIT model is incorrectly specified, in particular the response consistency and vignette equivalence assumptions may not hold for all vignettes. More research should be done to find out whether or not these two assumptions are tenable.

Third, it would be worthwhile to develop a validation method of the nonpara-

metric approach for anchoring vignettes. This approach has been introduced by King *et al.* (2004) and further developed by King and Wand (2007). It does not make any statistical assumptions, but does require the response consistency and vignette equivalence assumptions. The paper by King and Wand (2007) develops a method for evaluating and choosing anchoring vignettes, which uses entropy to measure the discriminatory power of a vignette. They recommend to use the set of vignettes that is most informative in terms of their nonparametric estimator. Although both their paper and this chapter study individual vignettes, the approaches differ. King and Wand (2007) study the amount of information in a single vignette, whereas we study whether the information of the vignette is correct.

Fourth, many other issues may be important in order to appropriately use the vignette method. For example, Buckley (2008) and Hopkins and King (2010) show several patterns of bias due to context effects. Specifically, they show that the order of the self-assessment question and the vignette questions is important.

2.A Tables

Table 2.1: Self-assessment and vignette evaluations for breathing for wave 1.

	B	F	DE	GR	IT	NL	ES	SE
Self-assessment								
None	63.93	60.86	65.32	67.56	73.82	70.29	74.34	38.85
Mild	25.14	22.09	20.16	24.58	16.04	23.50	15.57	29.32
Moderate	9.11	13.50	10.28	5.90	6.60	3.88	7.02	21.55
Severe	1.46	3.44	3.43	1.69	3.07	1.75	3.07	8.02
Extreme	0.36	0.12	0.81	0.28	0.47	0.58	0.00	2.26
Vignette b1								
None	18.94	37.06	2.42	1.26	5.66	2.14	1.10	0.75
Mild	34.43	32.52	14.52	24.58	28.54	26.60	7.02	15.54
Moderate	33.70	24.66	49.60	43.96	38.21	43.11	34.21	43.86
Severe	10.93	5.40	31.85	26.54	25.00	23.11	50.66	36.34
Extreme	2.00	0.37	1.61	3.65	2.59	5.05	7.02	3.51
Vignette b2								
None	1.82	2.94	3.83	0.70	5.19	2.72	0.66	0.75
Mild	4.92	2.21	7.66	5.34	8.96	4.27	3.51	7.02
Moderate	23.68	18.53	23.99	18.54	21.46	20.97	15.79	14.79
Severe	51.37	66.01	55.24	46.07	45.05	40.97	56.58	49.37
Extreme	18.21	10.31	9.27	29.35	19.34	31.07	23.46	28.07
Vignette b3								
None	2.00	3.31	3.63	0.56	5.66	3.30	0.66	0.75
Mild	1.64	1.72	3.43	1.26	4.25	1.75	1.32	3.76
Moderate	5.46	5.28	8.06	10.81	11.56	6.60	16.01	7.02
Severe	47.91	61.60	45.36	36.94	38.92	21.36	44.96	49.87
Extreme	42.99	28.10	39.52	50.42	39.62	66.99	37.06	38.60

The numbers in this table are proportions and based on a sample that is selected on the self-assessment, vignettes, and the objective measure, as well as on the covariates ($N = 4366$). Country abbreviations: Belgium (B), France (F), Greece (GR), Germany (DE), Italy (IT), the Netherlands (NL), Spain (ES), Sweden(SE).

Table 2.2: Self-assessment and vignette evaluations for cognition for wave 1.

	B	F	DE	GR	IT	NL	ES	SE
Self-assessment								
None	33.88	39.15	44.33	52.59	42.12	42.69	44.13	56.60
Mild	45.17	35.91	36.08	31.47	35.06	47.95	23.91	21.83
Moderate	19.31	21.45	16.08	13.29	15.76	6.82	22.17	12.44
Severe	1.46	3.24	3.51	2.66	5.41	1.95	9.57	8.38
Extreme	0.18	0.25	0.00	0.00	1.65	0.58	0.22	0.76
Vignette c1								
None	17.85	16.08	23.30	41.40	27.53	21.83	16.74	5.58
Mild	64.30	53.37	49.48	39.44	43.76	68.81	38.26	24.11
Moderate	15.30	26.06	24.33	15.80	20.24	8.19	33.26	46.19
Severe	2.37	3.87	2.27	3.36	7.76	1.17	11.52	23.60
Extreme	0.18	0.62	0.62	0.00	0.71	0.00	0.22	0.51
Vignette c2								
None	2.19	4.74	8.25	11.47	6.35	1.36	4.35	0.51
Mild	26.59	31.17	33.40	34.41	31.53	17.35	27.83	6.09
Moderate	51.55	51.62	44.74	37.62	42.12	50.88	48.04	20.81
Severe	18.76	11.60	13.20	15.94	18.59	25.15	19.13	57.87
Extreme	0.91	0.87	0.41	0.56	1.41	5.26	0.65	14.72
Vignette c3								
None	1.28	2.37	2.89	3.08	4.00	1.17	0.43	0.25
Mild	9.84	9.23	8.66	13.57	15.29	5.26	4.35	1.78
Moderate	33.15	39.28	27.01	27.83	32.47	31.77	28.26	8.88
Severe	45.90	44.39	50.72	44.06	40.24	39.38	60.87	58.63
Extreme	9.84	4.74	10.72	11.47	8.00	22.42	6.09	30.46

The numbers in this table are proportions and based on a sample that is selected on the self-assessment, vignettes, and four objective measures, as well as on the covariates ($N = 4343$). Country abbreviations: Belgium (B), France (F), Greece (GR), Germany (DE), Italy (IT), the Netherlands (NL), Spain (ES), Sweden(SE).

Table 2.3: Self-assessment and vignette evaluations for mobility for wave 1.

	B	F	DE	GR	IT	NL	ES	SE
Self-assessment								
None	55.35	66.91	46.26	74.30	58.55	57.93	52.49	38.36
Mild	26.32	15.13	27.07	15.36	20.37	24.86	19.96	38.36
Moderate	12.34	13.78	18.99	5.31	11.01	11.47	17.79	17.90
Severe	4.54	3.69	7.27	3.63	7.26	4.40	8.68	4.60
Extreme	1.45	0.49	0.40	1.40	2.81	1.34	1.08	0.77
Vignette m1								
None	11.43	9.23	5.86	9.64	21.78	4.02	3.04	14.58
Mild	43.92	32.60	26.67	33.66	36.53	43.21	21.48	40.92
Moderate	36.84	47.60	49.70	44.69	30.91	38.62	54.45	34.27
Severe	7.44	9.84	16.97	11.59	9.84	11.85	20.39	9.72
Extreme	0.36	0.74	0.81	0.42	0.94	2.29	0.65	0.51
Vignette m2								
None	2.18	2.71	3.43	1.26	4.68	1.91	1.30	1.28
Mild	13.43	7.75	11.52	17.04	11.24	9.94	8.68	15.86
Moderate	41.20	39.85	35.96	38.97	30.21	33.65	44.25	46.04
Severe	35.93	45.88	43.84	36.45	43.79	39.77	40.56	35.04
Extreme	7.26	3.81	5.25	6.28	10.07	14.72	5.21	1.79
Vignette m3								
None	1.81	2.34	1.21	0.56	4.22	1.53	0.65	0.00
Mild	3.81	8.24	7.07	5.03	12.18	2.49	5.21	2.56
Moderate	35.75	37.02	26.26	22.91	20.37	29.06	24.95	14.83
Severe	43.56	47.72	56.77	41.06	51.52	39.39	59.87	59.08
Extreme	15.06	4.67	8.69	30.45	11.71	27.53	9.33	23.53

The numbers in this table are proportions and based on a sample that is selected on the self-assessment, vignettes, and the objective measure, as well as on the covariates ($N = 4377$). Country abbreviations: Belgium (B), France (F), Greece (GR), Germany (DE), Italy (IT), the Netherlands (NL), Spain (ES), Sweden(SE).

Table 2.4: Self-assessments and vignette evaluations for cognition, breathing and mobility for wave 2.

	Self-assessment										Vignette evaluation									
	B	CZ	DK	F	DE	IT	NL	PO	ES	SE	B	CZ	DK	F	DE	IT	NL	PO	ES	SE
Cognition																				
None	30.88	39.21	54.66	35.41	43.38	37.89	39.88	37.38	43.34	41.68	25.73	29.83	32.49	21.81	26.76	28.78	32.26	12.52	15.51	27.21
Mild	48.54	43.05	30.36	42.49	38.93	38.91	50.10	28.55	28.63	36.72	61.87	57.74	53.95	55.81	53.33	47.58	62.12	41.99	50.70	48.81
Moderate	17.54	14.24	13.06	20.11	14.40	15.86	7.62	24.86	19.88	18.36	10.18	11.19	12.85	19.26	16.80	19.68	4.81	35.73	25.05	20.52
Severe	2.69	3.05	1.92	1.70	2.93	6.31	1.80	7.00	6.56	3.02	2.11	1.02	0.71	3.12	2.67	3.82	0.80	8.84	8.75	3.02
Extreme	0.35	0.45	0.00	0.28	0.36	1.03	0.60	2.21	1.59	0.22	0.12	0.23	0.00	0.00	0.44	0.15	0.00	0.92	0.00	0.43
Breathing																				
None	60.51	56.39	78.87	56.33	66.83	76.81	66.80	64.09	71.75	63.47	6.23	1.33	3.74	3.00	3.23	6.19	2.68	4.44	0.91	4.01
Mild	27.38	27.35	14.09	25.00	20.06	16.28	27.63	16.02	17.31	24.05	5.33	28.31	26.57	23.33	28.80	32.92	25.77	11.58	12.53	24.94
Moderate	9.41	13.01	5.55	16.00	9.03	4.96	3.71	11.58	7.74	8.69	1.93	50.12	41.62	50.33	39.92	38.58	53.20	30.89	27.33	28.06
Severe	2.08	3.13	1.39	2.33	3.80	1.59	1.24	6.37	2.73	3.34	5.53	18.67	25.93	23.00	26.62	20.88	16.29	50.39	53.53	35.19
Extreme	0.61	0.12	0.11	0.33	0.29	0.35	0.62	1.93	0.46	0.45	0.98	1.57	2.13	0.33	1.43	1.42	2.06	2.70	5.69	7.80
Mobility																				
None	59.20	33.28	76.96	82.71	51.00	67.19	55.83	55.27	64.16	63.17	9.29	4.18	8.23	8.41	5.72	12.24	5.10	8.55	6.14	5.71
Mild	27.20	42.44	14.94	8.88	31.59	21.09	33.50	19.37	19.11	20.32	55.56	48.87	37.97	39.25	29.48	46.61	43.20	27.64	26.28	34.60
Moderate	11.77	19.77	6.84	8.41	14.30	10.16	7.52	18.52	12.63	11.11	28.86	39.87	39.75	40.65	43.28	33.07	39.08	39.60	45.39	30.48
Severe	1.66	4.18	1.27	0.00	3.11	1.56	2.67	6.27	3.75	5.08	6.30	7.07	13.92	11.68	20.27	7.81	11.89	23.65	21.84	27.30
Extreme	0.17	0.32	0.00	0.00	0.00	0.00	0.49	0.57	0.34	0.32	0.00	0.00	0.13	0.00	1.24	0.26	0.73	0.57	0.34	1.90

The numbers in this table are proportions and based on different samples for every domain. Each sample is selected on the relevant self-assessment, vignette, and objective measure(s), as well as on the covariates. For cognition $N = 6895$, for breathing $N = 6393$, and for mobility $N = 4788$. Country abbreviations: Belgium (B), Czech Republic (CZ), Denmark (DK), France (F), Germany (DE), Italy (IT), the Netherlands (NL), Poland (PO), Spain (ES), Sweden(SE).

Table 2.5: Descriptive statistics for delayed recall for wave 2

Nr. words	B	CZ	DK	F	DE	IT	NL	PO	ES	SE	Total
0	8.54	12.77	5.06	5.38	4.53	11.45	4.61	16.21	10.74	3.24	8.18
1	9.12	7.57	3.85	9.92	4.80	11.01	3.81	11.60	15.11	3.46	7.56
2	10.64	13.67	7.69	13.88	11.20	16.89	11.02	17.50	17.89	9.50	12.50
3	20.12	20.56	15.79	20.96	19.64	18.50	13.63	21.36	21.67	18.14	18.97
4	21.99	21.58	19.64	18.70	20.09	19.38	19.04	19.34	18.89	19.87	20.07
5	14.27	13.79	20.95	16.43	20.53	9.99	15.83	8.29	8.55	21.81	15.61
6	8.89	6.33	13.97	9.07	10.67	6.46	16.23	3.68	4.77	13.39	9.47
7	4.44	2.60	7.79	3.68	5.87	3.38	9.02	1.47	2.19	6.26	4.83
8	1.29	0.79	3.14	1.70	1.87	1.47	4.01	0.55	0.20	3.24	1.81
9	0.70	0.11	1.82	0.28	0.53	0.44	2.00	0.00	0.00	0.65	0.70
10	0.00	0.23	0.30	0.00	0.27	1.03	0.80	0.00	0.00	0.43	0.30
Mean	3.51	3.17	4.32	3.53	3.96	3.16	4.40	2.66	2.79	4.25	3.62
Std.dev	1.96	1.91	2.01	1.87	1.86	2.08	2.13	1.79	1.75	1.86	2.02

Note: The numbers in the first 11 rows of this table are proportions and are based on a sample that is selected on the self-assessment, vignette, the four objective measures, as well as on the covariates ($N = 6895$). Country abbreviations: Belgium (B), Czech Republic (CZ), Denmark (DK), France (F), Germany (DE), Italy (IT), the Netherlands (NL), Poland (PO), Spain (ES), Sweden(SE).

Table 2.6: Descriptive statistics for the peak flow test (breathing) and the stand-up test (mobility) for wave 2

	Peak flow test			Stand-up test		
Country	Median	Mean	Std.dev	Median	Mean	Std.dev
Belgium	330	345.66	149.77	10.78	11.76	5.38
Czech Rep.	320	326.21	132.10	10.09	10.95	4.29
Denmark	390	394.32	145.88	9.19	9.69	3.26
France	350	358.40	174.67	10.00	10.89	6.06
Germany	350	361.42	149.19	9.50	10.98	6.09
Italy	280	294.63	145.09	10.65	12.59	7.44
Netherlands	390	403.24	149.06	10.73	11.75	5.28
Poland	305	325.34	153.01	10.01	11.17	4.60
Spain	270	328.53	225.51	11.00	12.62	6.52
Sweden	420	434.09	141.69	9.44	9.87	3.65
Total	350	356.71	158.33	10	11.10	5.33

Note: The unit of measurement for the peak flow test is liters/minute and it ranges from 60 to 880. The unit of measurement for the stand-up test is time in seconds. We only observe the exact time for those respondents who are able to complete the test in 1 minute. The numbers in this table are based on different samples for every domain. Each sample has been selected on the relevant self-assessment, vignette, and objective measure, as well as on the covariates. For breathing $N = 6393$, and for mobility $N = 4788$.

Table 2.7: Description of covariates

Covariate	Description
Age	Depending on the wave, each respondent's age is calculated as 2004 or 2007 minus the year or birth, which information is provided by SHARE. We include dummies for age groups.
Education	Based on the International Standard Classification of Education (ISCED 97) we categorize education in three dummies: low, middle and high. We define ISCED levels 0 and 1 as low education, 2 and 3 as middle and 4,5 and 6 as high education. The lowest educational group is used as the reference group in the analyzes.
Gender	We include a dummy for being male.
Not alone	SHARE contains information on the marital status of its respondents. Possible answers are: (1) married and living together with spouse, (2) registered partnership, (3) married, living separated from spouse, (4) never married, (5) divorced, (6) widowed. We include a dummy for whether a respondent is living alone or not, where we define "not living alone" if marital status is reported as either (1) or (2).
Physical activity	SHARE contains information on the frequency of physical activity, such as sports or heavy housework, of its respondents. Possible answers are (1) more than once a week, (2) once a week, (3) one to three times a month, (4) hardly ever, or never. We define that a respondents is engaged *often* in physical activity if the answer is (1), *sometimes* if the answer is (2) or (3) and *never* if the answer is (4). The group of respondents that reports to be engaged in physical activity "hardly ever or never engaged" is taken as the reference group.
Illness long	The SHARE questionnaire contains the following question: "Some people suffer from chronic or long-term health problems. By long-term we mean it has troubled you over a period of time or is likely to affect you over a period of time. Do you have any long-term health problems, illness, disability or infirmity?" The answer can be either yes or no.

Table 2.8: Descriptive statistics of the covariates for the different samples for both waves

	Wave 1			
	Cognition	Breathing	Mobility	Vignette sample
Male (%)	44.83	44.69	44.71	44.43
Education mid (%)	45.08	45.03	44.92	44.60
Education high (%)	20.22	20.22	20.20	19.86
Not alone (%)	74.76	74.94	74.80	74.42
Long-term illness (%)	45.84	46.11	46.13	46.49
Phys. act. sometimes (%)	25.33	25.24	25.22	25.08
Phys. act. often (%)	34.58	34.63	34.48	34.46
Mean age	62.92	62.91	62.94	63.06
Std.dev age	9.95	9.94	9.95	10.01
N	4343	4366	4377	4544

	Wave 2			
	Cognition	Breathing	Mobility	Vignette sample
Male (%)	44.71	45.13	45.76	44.60
Education mid (%)	59.29	59.61	59.54	59.18
Education high (%)	23.36	24.25	27.46	23.20
Not alone (%)	74.95	75.44	79.45	74.43
Long-term illness (%)	49.31	48.22	44.13	49.66
Phys. act. sometimes (%)	23.61	24.15	25.77	23.53
Phys. act. often (%)	33.50	34.24	39.81	33.10
Mean age	64.37	64.09	61.11	64.56
Std.dev age	9.74	9.57	7.16	9.86
N	6895	6393	4788	7186

The numbers in this table are based on different samples for every domain. Each sample, except the vignette sample, is selected on the relevant self-assessment, vignette(s), and objective measure(s), as well as on the covariates.

Table 2.9: Correlations between two DIF-adjusted self-assessments for wave 1.

Cognition		$\hat{y}^*_{c1}$	$\hat{y}^*_{c2}$	$\hat{y}^*_{c3}$
	$\hat{y}^*_{c1}$	1.00	0.76	0.75
	$\hat{y}^*_{c2}$		1.00	0.92
	$\hat{y}^*_{c3}$			1.00
Breathing		$\hat{y}^*_{b1}$	$\hat{y}^*_{b2}$	$\hat{y}^*_{b3}$
	$\hat{y}^*_{b1}$	1.00	0.56	0.44
	$\hat{y}^*_{b2}$		1.00	0.94
	$\hat{y}^*_{b3}$			1.00
Mobility		$\hat{y}^*_{m1}$	$\hat{y}^*_{m2}$	$\hat{y}^*_{m3}$
	$\hat{y}^*_{m1}$	1.00	0.96	0.90
	$\hat{y}^*_{m2}$		1.00	0.94
	$\hat{y}^*_{m3}$			1.00

Note: The numbers are correlation coefficients between predicted values of two DIF-adjusted self-assessments, each computed by estimating the CHOPIT-model using 1 vignette at a time. The vignette used is denoted in the subscript. By predicted values we mean: $\hat{y}^*_{si} = \mathbf{x}_i'\hat{\beta}_s$. Estimations are done separately for each domain and the samples are selected on the relevant self-assessment, vignettes, as well as on the covariates. For cognition the sample is also selected on objective measures. For cognition $N = 4343$, for breathing $N = 4366$, and for mobility $N = 4377$.

Table 2.10: Summary of results for cognition for wave 1.

Vignette	No. of parameters	Loglikelihood	AIC	Corr($\hat{y}^*_{si}, \hat{y}^*_{oi}$)	Corr($\varepsilon_{si}, \varepsilon_{oi}$)	Corr(y^*_{si}, y^*_{oi})
Immediate recall						
Vignette c1						
A	53	-16,209.12	32,462.24	0.70	0.12	0.24
B	116	-15,689.02	31,422.04	0.45	0.12	0.19
Vignette c2						
A	53	-16,255.96	32,555.91	0.70	0.12	0.24
B	116	-15,770.65	31,585.30	0.68	0.13	0.28
Vignette c3						
A	53	-15,197.01	30,438.02	0.70	0.12	0.24
B	116	-14,894.04	29,832.08	0.78	0.13	0.32
Delayed recall						
Vignette c1						
A	52	-15,137.61	30,319.22	0.75	0.15	0.26
B	115	-14,618.11	29,280.22	0.52	0.15	0.22
Vignette c2						
A	52	-15,184.44	30,412.89	0.75	0.15	0.26
B	115	-14,698.00	29,440.00	0.74	0.15	0.30
Vignette c3						
A	52	-14,125.50	28,295.00	0.75	0.15	0.26
B	115	-13,821.86	27,687.73	0.83	0.16	0.34
Numeracy						
Vignette c1						
A	52	-15,129.56	30,303.13	0.75	0.08	0.21
B	115	-14,608.73	29,261.46	0.54	0.07	0.17
Vignette c2						
A	52	-15,176.40	30,396.80	0.75	0.08	0.21
B	115	-14,692.70	2,9429.40	0.72	0.07	0.25
Vignette c3						
A	52	-14,117.46	28,278.91	0.75	0.08	0.21
B	115	-13,815.77	27,675.54	0.78	0.07	0.29
Verbal fluency						
Vignette c1						
A	53	-16,022.07	32,088.13	0.50	0.12	0.20
B	116	-15,502.38	31,048.75	0.62	0.11	0.22
Vignette c2						
A	53	-16,070.72	32,185.45	0.50	0.12	0.20
B	116	-15,587.28	31,218.56	0.73	0.12	0.30
Vignette c3						
A	53	-15,009.33	30,062.66	0.50	0.12	0.20
B	116	-14,709.81	29,463.60	0.74	0.12	0.32

Model A assumes there is no heterogeneity in reporting behavior, and no response consistency (No DIF, No RC). Model B assumes there is heterogeneity in reporting behavior, and response consistency holds (DIF, RC). Corr($\hat{y}^*_{si}, \hat{y}^*_{oi}$) is the correlation coefficient between the predicted systematic values, Corr(y^*_{si}, y^*_{oi}) is the correlation coefficient between the simulated values. All estimates are based on a sample that is selected on the self-assessment, the three vignettes, the four objective measures, as well as on the covariates. $N = 4343$

Table 2.11: Summary of results for cognition, breathing and mobility for wave 2.

Domain	No. of parameters	Loglikelihood	AIC	Corr($\hat{y}^*_{si}, \hat{y}^*_{oi}$)	Corr($\varepsilon_{si}, \varepsilon_{oi}$)	Corr(y^*_{si}, y^*_{oi})
Cognition						
Immediate recall						
A	57	-25,010.81	50,069.62	0.78	0.21	0.32
B	126	-24,558.24	49,164.48	0.40	0.21	0.24
Delayed recall						
A	56	-23,223.06	46,494.13	0.79	0.21	0.32
B	125	-22,770.11	45,588.21	0.43	0.21	0.24
Numeracy						
A	56	-23,538.05	47,124.09	0.71	0.14	0.24
B	125	-23,082.89	46,213.79	0.40	0.13	0.18
Verbal fluency						
A	57	-24,880.49	49,808.97	0.73	0.11	0.24
B	126	-24,421.47	48,890.93	0.37	0.10	0.16
Breathing						
A	57	-20,646.37	41,342.74	0.45	0.18	0.25
B	129	-20224.53	40499.06	0.53	0.18	0.27
Mobility						
A	51	-24,781.54	49,607.07	0.55	0.13	0.18
B	114	-24,516.58	49,077.16	0.63	0.13	0.20

Model A assumes there is no heterogeneity in reporting behavior, and no response consistency (No DIF, No RC). Model B assumes there is heterogeneity in reporting behavior, and response consistency holds (DIF, RC). Corr($\hat{y}^*_{si}, \hat{y}^*_{oi}$) is the correlation coefficient between the predicted systematic values, Corr(y^*_{si}, y^*_{oi}) is the correlation coefficient between the simulated values.Estimates are based on different samples for every domain. Each sample is selected on the relevant self-assessment, vignette, and objective measure(s), as well as on the covariates. For cognition $N = 6895$, for breathing $N = 6393$, and for mobility $N = 4788$.

Table 2.12: Parameter estimates for cognition for wave 2 with delayed recall as the objective measure and vignette c1.

	Model A				Model B							
	β_o		β_s		β_s		γ^1		γ^2		γ^3	
Covariates	Coeff	T-values	Coeff	T-values	Coeff	T-values	Coeff	T-values	Coeff	T-values	Coeff	T-values
Constant	2.44	49.35	3.19	52.25	3.16	31.56	1.00	-	-0.16	-1.94	0.20	4.62
Belgium	-0.11	-2.33	-0.21	-4.10	-0.37	-5.58	-0.34	-4.10	0.05	0.64	0.22	5.74
Czech Rep.	-0.34	-7.51	-0.03	-0.66	-0.21	-3.17	-0.32	-3.72	0.05	0.57	0.13	3.57
Denmark	0.14	3.13	0.20	3.90	0.03	0.52	-0.38	-4.06	0.22	2.80	-0.02	-0.47
France	-0.08	-1.35	-0.13	-1.84	-0.09	-0.92	-0.25	-2.25	0.24	2.52	0.09	1.73
Italy	-0.19	-3.73	-0.19	-3.24	-0.27	-3.45	0.03	0.37	-0.07	-0.90	-0.06	-1.39
Netherlands	0.18	3.38	-0.07	-1.18	-0.37	-4.68	-0.42	-3.41	-0.21	-1.59	0.31	7.28
Poland	-0.66	-12.47	-0.29	-4.97	0.20	2.61	0.55	7.11	0.09	1.30	-0.22	-4.48
Spain	-0.40	-6.92	-0.12	-1.87	0.14	1.66	0.37	4.13	-0.06	-0.72	-0.10	-1.89
Sweden	0.29	5.21	-0.03	-0.53	-0.04	-0.53	-0.06	-0.59	0.13	1.50	-0.10	-1.97
Male	-0.26	-10.59	0.05	1.84	0.15	4.09	0.05	1.17	0.08	2.08	-0.04	-1.78
Age ¡ 50	0.23	2.91	0.02	0.27	-0.01	-0.06	0.22	1.73	-0.19	-1.30	-0.13	-1.90
Age 50 to 55	0.12	2.89	0.00	0.07	-0.06	-1.03	-0.09	-1.17	0.01	0.09	0.03	0.78
Age 60 to 65	-0.08	-1.96	-0.04	-0.92	0.03	0.48	-0.02	-0.32	0.07	1.06	0.04	1.08
Age 65 to 70	-0.24	-5.75	-0.21	-4.54	-0.15	-2.45	-0.06	-0.85	0.11	1.71	0.05	1.47
Age 70 to 75	-0.40	-9.06	-0.30	-6.00	-0.19	-2.90	0.12	1.63	-0.04	-0.59	0.05	1.52
Age 75 to 80	-0.73	-13.85	-0.46	-7.96	-0.38	-5.03	0.04	0.54	0.08	1.02	-0.04	-0.83
Age ¿ 80	-0.90	-15.72	-0.69	-11.26	-0.56	-7.21	0.07	0.75	0.06	0.80	0.06	1.35
Education low	-0.34	-8.61	-0.08	-1.87	0.03	0.56	0.19	3.08	-0.05	-0.86	-0.06	-1.70
Education high	0.34	11.12	0.14	3.92	0.02	0.40	-0.13	-2.03	-0.03	-0.52	0.04	1.47
Phys. act. sometimes	0.14	4.51	0.13	3.74	0.16	3.58	-0.09	-1.65	0.08	1.64	0.06	2.42
Phys. act. often	0.15	5.07	0.14	4.30	0.15	3.36	-0.13	-2.52	0.08	1.71	0.07	2.77
Alone	-0.09	-3.18	-0.03	-0.93	-0.04	-1.06	0.05	0.97	-0.03	-0.74	-0.03	-1.43
Illness long	-0.08	-3.21	-0.40	-14.07	-0.44	-11.86	0.01	0.29	0.02	0.44	-0.06	-2.70

	Model A		Model B	
Threshold parameters	Coeff	T-values	Coeff	T-values
γ_s^1	1.00	-		
γ_s^2	-0.01	-0.49		
γ_s^3	0.13	7.91		
γ_v^1	1.00	-		
γ_v^2	-0.01	-0.24		
γ_v^3	0.39	29.29		
γ_o^1	0.00	-	0.00	-
γ_o^2	0.00	-	0.00	-
γ_o^3	0.07	2.56	0.07	2.57
Variances				
σ_s^2	1.00	-	1.00	-
σ_u^2	0.00	-	0.38	19.16
σ_v^2	1.00	-	0.75	43.30
σ_o^2	0.91	53.33	0.91	52.71
ρ	0.21	14.59	0.21	13.52
Vignette dummy	Coeff	Std.error	Coeff	Std.error
ϑ	2.84	0.03	2.60	0.02

Model A assumes there is no heterogeneity in reporting behavior and no response consistency (No DIF, No RC). Model B assumes there is heterogeneity in reporting behavior and response consistency holds (DIF, RC). The estimates for β_o are the same in model A and B. All estimates are based on a sample that is selected on the self-assessment, the vignette, the four objective measures, as well as on the covariates. $N = 6895$.

Table 2.13: Parameter estimates for breathing for wave 2 with vignette b1.

	Model A				Model B							
	β_o		β_s		β_s		γ^1		γ^2		γ^3	
Covariates	Coeff	T-values	Coeff	T-values	Coeff	T-values	Coeff	T-values	Coeff	T-values	Coeff	T-values
Constant	-3.37	-12.71	2.59	6.21	2.52	4.41	1.00	-	-0.44	-1.11	-0.02	-0.06
Belgium	0.01	0.14	-0.08	-1.31	-0.16	-2.06	-0.28	-6.05	0.15	2.60	0.13	2.25
Czech Rep.	-0.13	-3.34	-0.17	-2.99	-0.03	-0.39	-0.16	-3.56	0.27	4.94	0.17	2.73
Denmark	0.16	4.57	0.36	5.95	0.38	4.71	-0.01	-0.36	0.06	1.11	-0.02	-0.38
France	0.11	2.09	-0.20	-2.42	-0.04	-0.38	-0.11	-1.80	0.34	4.59	0.01	0.15
Italy	-0.28	-6.27	0.34	4.58	0.29	3.01	-0.10	-1.91	0.02	0.30	0.02	0.24
Netherlands	0.11	2.41	-0.07	-0.96	0.14	1.40	-0.23	-4.27	0.34	5.27	0.23	3.24
Poland	-0.13	-3.10	-0.03	-0.43	0.14	1.64	0.45	9.34	-0.14	-1.89	-0.50	-5.48
Spain	-0.06	-1.15	0.23	2.83	0.70	6.23	0.58	10.18	-0.11	-1.28	-0.08	-0.80
Sweden	0.48	10.46	-0.05	-0.69	0.04	0.47	0.25	5.01	-0.29	-3.65	0.01	0.13
Male	0.62	22.57	-0.05	-1.24	0.02	0.30	0.16	4.88	-0.07	-1.58	-0.04	-0.92
Age ¡ 50	0.17	2.66	0.14	1.21	0.10	0.66	-0.04	-0.49	-0.06	-0.60	0.03	0.28
Age 50 to 55	0.12	3.61	0.04	0.67	0.03	0.47	0.04	1.14	-0.08	-1.58	0.01	0.26
Age 60 to 65	-0.08	-2.34	-0.03	-0.62	0.01	0.08	0.02	0.46	0.01	0.12	0.02	0.45
Age 65 to 70	-0.21	-5.95	-0.14	-2.13	-0.04	-0.47	0.02	0.50	0.06	1.24	0.03	0.53
Age 70 to 75	-0.42	-11.32	-0.19	-3.37	-0.11	-1.41	0.02	0.51	0.05	0.99	0.04	0.61
Age 75 to 80	-0.57	-13.06	-0.37	-5.56	-0.33	-3.78	-0.05	-0.91	0.14	2.25	-0.03	-0.38
Age ¿ 80	-0.70	-14.75	-0.39	-5.60	-0.32	-3.44	-0.02	-0.39	0.18	2.73	-0.07	-0.96
Education low	-0.07	-2.04	-0.06	-1.16	-0.15	-2.29	-0.01	-0.33	-0.07	-1.32	-0.04	-0.85
Education high	0.17	6.67	0.14	3.32	0.17	3.03	0.03	1.12	-0.01	-0.25	0.01	0.32
Phys. act. sometimes	0.10	3.95	0.23	5.55	0.22	3.94	-0.08	-2.43	0.03	0.83	0.06	1.39
Phys. act. often	0.19	7.79	0.35	8.82	0.37	6.90	0.00	0.11	-0.04	-0.99	0.06	1.57
Alone	-0.08	-3.16	-0.03	-0.75	-0.04	-0.84	-0.00	-0.05	-0.00	-0.01	-0.01	-0.25
Illness long	-0.07	-3.51	-0.61	-18.42	-0.59	-12.70	0.05	2.05	0.02	0.71	-0.03	-0.83
Height/100	1.88	11.80	0.38	1.53	0.68	2.08	0.24	1.35	0.07	0.28	-0.08	-0.31

	Model A		Model B	
Threshold parameters	Coeff	T-values	Coeff	T-values
γ_s^1	1.00	-		
γ_s^2	-0.28	-7.04		
γ_s^3	-0.18	-7.55		
γ_v^1	1.00	-		
γ_v^2	0.07	4.02		
γ_v^3	0.24	10.39		
Variances				
σ_s^2	1.00	-	1.00	-
σ_u^2	0.00	-	0.27	10.70
σ_v^2	1.00	-	0.63	29.70
σ_o^2	0.79	113.04	0.79	155.22
ρ	0.18	11.47	0.18	12.52
Vignette dummy	Coeff	Std.error	Coeff	Std.error
ϑ	1.54	0.02	1.90	0.30

Model A assumes there is no heterogeneity in reporting behavior and no response consistency (No DIF, No RC). Model B assumes there is heterogeneity in reporting behavior and response consistency holds (DIF, RC). The estimates for β_o are the same in model A and B. Since the objective measure is a continuous variable there are no thresholds in the objective part of the model. All estimates are based on a sample that is selected on the self-assessment,the three vignettes, the four objective measures, as well as on the covariates. $N = 6393$.

Table 2.14: Parameter estimates for mobility for wave 2 with vignette m1

	Model A				Model B							
	β_o		β_s		β_s		γ^1		γ^2		γ^3	
Covariates	Coeff	T-values	Coeff	T-values	Coeff	T-values	Coeff	T-values	Coeff	T-values	Coeff	T-values
Constant	10.26	37.70	0.56	8.69	0.53	7.55	1.00	-	-0.07	-1.08	-0.22	-3.05
Belgium	0.67	2.37	-0.17	-2.54	0.20	2.43	0.30	4.72	0.19	3.26	0.03	0.43
Czech Rep.	-0.25	-0.90	0.34	5.47	0.46	5.68	-0.06	-0.99	0.31	5.77	0.18	2.87
Denmark	-1.15	-4.43	-0.61	-9.51	-0.38	-4.67	0.25	4.27	-0.07	-1.20	0.02	0.24
France	-0.09	-0.21	-0.85	-7.71	-0.58	-4.13	0.30	3.07	-0.14	-1.39	0.13	1.33
Italy	1.34	3.91	-0.36	-4.35	0.02	0.20	0.37	4.98	-0.00	-0.06	0.01	0.10
Netherlands	1.13	3.60	-0.03	-0.39	0.07	0.75	0.02	0.32	0.22	3.50	-0.00	-0.00
Poland	0.05	0.14	-0.09	-1.20	0.03	0.31	0.25	3.57	-0.33	-4.26	-0.11	-1.47
Spain	1.63	4.37	-0.25	-2.77	-0.18	-1.54	0.13	1.54	-0.24	-2.57	-0.00	-0.00
Sweden	-1.14	-3.30	-0.21	-2.63	-0.13	-1.26	0.15	1.83	-0.09	-1.14	-0.36	-4.11
Male	-0.72	-4.76	-0.09	-2.60	-0.18	-3.88	-0.08	-2.24	-0.00	-0.07	-0.04	-1.01
Age ¡ 50	-1.25	-2.84	-0.18	-1.63	-0.05	-0.37	0.20	2.16	-0.17	-1.68	-0.22	-1.92
Age 50 to 55	-0.36	-1.63	-0.02	-0.38	-0.03	-0.38	0.00	0.04	-0.04	-0.86	-0.02	-0.42
Age 60 to 65	0.49	2.22	0.00	0.08	0.03	0.55	0.01	0.12	0.06	1.41	0.02	0.41
Age 65 to 70	0.57	2.37	0.06	1.12	0.09	1.23	0.00	0.04	0.03	0.63	0.03	0.53
Age ¿ 70	1.68	6.41	0.10	1.69	0.10	1.32	-0.06	-0.97	0.12	2.37	0.11	1.77
Education low	0.16	0.60	0.07	1.08	0.06	0.73	0.01	0.19	-0.02	-0.46	-0.01	-0.08
Education high	-0.59	-3.33	-0.19	-4.50	-0.29	-5.26	-0.09	-2.15	-0.00	-0.04	-0.03	-0.64
Phys. act. never	1.14	6.34	0.23	5.36	0.29	5.48	0.08	1.93	-0.06	-1.53	0.06	1.47
Phys. act. sometimes	0.58	3.03	0.02	0.39	0.03	0.47	-0.00	-0.03	0.03	0.75	0.05	0.97
Alone	0.15	0.80	-0.02	-0.42	0.01	0.20	0.04	1.01	-0.03	-0.83	-0.08	-1.72
Illness long	0.80	5.21	0.80	22.06	0.81	17.07	-0.03	-0.78	-0.00	-0.08	0.10	2.67

	Model A		Model B	
Threshold parameters	Coeff	T-values	Coeff	T-values
γ_s^1	1.00	-		
γ_s^2	-0.08	-3.27		
γ_s^3	-0.04	-1.11		
γ_v^1	1.00	-		
γ_v^2	0.32	16.07		
γ_v^3	0.11	5.49		
Variances				
σ_s^2	1.00	-	1.00	-
σ_u^2	0.00	-	0.18	4.82
σ_v^2	1.00	-	0.68	34.11
σ_o^2	5.12	97.83	5.12	181.88
ρ	0.13	8.10	0.13	7.37
Vignette dummy	Coeff	Std.error	Coeff	Std.error
ϑ	2.46	0.03	2.16	0.06

Model A assumes there is no heterogeneity in reporting behavior and no response consistency (No DIF, No RC). Model B assumes there is heterogeneity in reporting behavior and response consistency holds (DIF, RC). The estimates for β_o are the same in model A and B. Since the objective measure is a continuous variable there are no thresholds in the objective part of the model. All estimates are based on a sample that is selected on the self-assessment,the three vignettes, the four objective measures, as well as on the covariates. $N = 4788$.

Chapter 3

Anchoring vignettes and response consistency assumption

3.1 Introduction

Subjective self-assessments are a convenient and widespread method of comparing many aspects of well-being. They are a commonly used summary tool in many socio-economic surveys, avoiding the need for large batteries of detailed and very specific questions. They are often used for international comparisons or comparisons between population groups.

A potential problem with subjective self-assessments is that people in different countries or in different socio-economic groups within a country may use different response scales. Consider, for example, the question: "Overall in the last 30 days, how much of a problem did you have with concentrating or remembering things?" with answers "none", "mild", "moderate", "severe", and "extreme". Earlier research has shown that the distributions of the answers to a question like this vary much more across countries than can plausibly be explained by genuine cognitive differences. Differences in response scales may contribute to explaining the observed cross-country differences, but with self-assessment data only, response scale differences and genuine differences are not separately identified.

Anchoring vignettes can be used as a tool to identify response scale differences and correct the self-assessments for such differences, enhancing comparability of subjective measures between countries or socio-economic groups (King *et al.* (2004)). Anchoring vignettes are short descriptions of aspects of hypothetical people's lives relevant to the domain of interest. For example, in the "concentration and remembering things" example used above, a vignette would describe how well a hypothetical person remembers the names of people to whom he/she

is introduced, how well she remembers what was on the TV news, how often she has to look for her keys because she forgot where he/she put them, or how often he/she has to go back home to collect an item she forgot to take with her. If respondents in different countries assess the concentration/remembering skills of the same vignette person in systematically different ways, this has to be because they use different response scales.

Anchoring vignettes have been applied in various domains of well-being, including various aspects of health, Salomon *et al.* (2004), Bago d'Uva *et al.* (2008b), work disability, Kapteyn *et al.* (2007), job satisfaction, Kristensen and Johansson (2008), political efficacy, King *et al.* (2004), satisfaction with the health care system, Murray *et al.* (2003), Sirven *et al.* (2008), and satisfaction with life in general, Kapteyn *et al.* (2010). However, using anchoring vignettes to correct for response scale differences requires identifying assumptions. Two key assumptions are "vignette equivalence" - different respondents interpret the same vignette in the same way - and "response consistency" - respondents use the same scales when evaluating themselves and when evaluating the vignette persons. A number of papers have analyzed the validity of these assumptions using alternative measures on an objective scale. For instance Van Soest *et al.* (2011) consider drinking behavior of Irish students and analyze response scale differences in their answers to questions about the extent to which they consider their drinking behavior problematic (on a subjective scale). They use self-reports on how much respondents drink (on an objective, numerical scale) to calibrate the subjective response scales of respondents in an alternative way. Comparing models with and without response scale differences, they find that the model using anchoring vignettes to correct for response scale differences provides the best description of the data and brings subjective and objective measure closer to each other. A somewhat similar approach is followed by Bago d'Uva *et al.* (2009) who consider cognitive functioning and mobility in the English Longitudinal Study of Aging. They find that in most cases response consistency and vignette equivalence are rejected by the data.

The purpose of the current study is to collect new data in order to test the response consistency assumption on several aspects of individual health in a more direct way. Essentially this is done by giving respondents vignettes describing their own health.

The basic idea of our experiment is as follows. The response consistency assumption is that there are no systematic differences between response scales for self-reports and vignette ratings for the same respondent. We can test this with vignettes that reflect a respondent's own situation. Under the null, there

should be no systematic differences between the respondent's self-reported health and the respondent's evaluation of a vignette mimicking the health of the same respondent.

We do this for various health domains. Consider the example of mobility. We first ask if the respondent had problems with moving around over the last thirty days. We then ask two specific questions on difficulties with walking and climbing stairs. The answers to these questions are used to construct vignettes that are administered in a new interview several months later. In that second interview, we present the replica vignette as well as a number of different vignettes so that the respondents are unlikely to notice that we are giving them a description of their own health.

We use the American Life Panel, a high frequency Internet panel representative of the adult US population. This Internet panel is particularly useful for our research because 1) it allows for interviewing the same people twice in the course of a few months, and 2) exploiting the Internet survey programming flexibility, answers to the first interview about own health can be preloaded in constructing vignettes for the second interview.

The health domains we considered in the experiment are sleep, mobility, memory and concentration, feeling down or depressed, breathing, and pain. These health domains were selected because they are the health domains used in vignette experiments in SHARE, the Survey of Health, Ageing and Retirement in Europe. In this chapter we focus on all of these domains, except pain. The analysis for pain is more complicated and left for another paper[1] (see, e.g., Dourgnon and Lardjane (2007)).

For each of the domains, we first perform nonparametric tests comparing the self-assessments in the first interview and the replica vignette assessments in the second interview. If the replica vignette describes the respondent's health in the given domain correctly, and response scales are stable between the two interviews, then response consistency corresponds to the null hypothesis that the two distributions should be the same[2]. We then also estimate parametric models that explain the respondent's self-assessments and the replica vignette evaluations from covariates such as age, gender, education, etc., and test for parameter

[1] We collected self-reports and vignettes for four pain domains: Headaches; Back Pain; Joints Pain; Neck Pain. However,when constructing the replica vignettes we left out all observations for which the self-reports in the first wave indicated no pain whatsoever. Administering a vignette where a vignette person has no pain and then asking the respondent to evaluate whether the vignette person has pain seems artificial. This procedure leads to selectivity and more complicated tests and models than the ones used in this chapter.

[2] Or at least similar. We discuss below under what circumstances the distributions should be equal and how even when response consistency holds deviations are possible.

restrictions implied by response consistency and other, auxiliary, assumptions.

The remainder of this chapter is organized as follows. Section 3.2 describes the data that we use and the actual construction of vignettes in our experiments. Section 3.3 contains descriptive statistics and the results of the nonparametric tests. In Section 3.4 we present the results for parametric models, giving insight in why the nonparametric tests lead to rejection in most cases. Section 3.5 concludes.

3.2 Data and Construction of Vignettes in Our Experiment

In this research, we use the RAND American Life Panel (ALP). The ALP is an ongoing Internet panel of approximately 2500 respondents 18 and over. Respondents in the panel either use their own computer to log on to the Internet or, if they do not have a computer, a Web TV (http://www.webtv.com/pc/), which allows them to access the Internet, using their television and a telephone line. This technology allows respondents who did not have previous Internet access to participate in the panel and to use the Web TVs for browsing the Internet or using email. Currently, about 10% of the panel members use a Web TV.[3]

About twice a month, respondents receive an email with a request to visit the ALP web site and fill out one or more questionnaires. Typically a single interview will not take more than 30 minutes. Respondents are paid an incentive of about $20 per thirty minutes of interviewing (and proportionately less if an interview is shorter). Most respondents respond within one week and the majority within three weeks. To further increase response rates, reminders are sent after this period.

We implemented our questions and vignettes to test response consistency in two separate waves of the ALP. In wave 1 (December 2008) we asked self-assessments by specific health domain and a set of detailed "objective" questions on health in each of the domains. The purpose of the latter was to obtain information about the actual health of the respondent in that health domain. Then, in wave 2 (March 2009), we again asked the self-assessments by health domain, and then asked three vignette questions for each domain. One of the vignettes in each domain described the vignette person as having the domain-specific health of the respondent as reported in wave 1. We call these the replica vignettes.

[3]This describes the situation at the time of the data collection. Currently, new panel members without Internet receive a laptop and a high speed Internet connection.

The other vignettes in a given domain are constructed in such a way that they always describe a situation that is different from the respondent's situation (as reported in wave 1). The order of the three vignettes was randomized. The main reason for adding the two vignettes not describing the respondents' health was to reduce the likelihood that respondents would notice that we presented vignettes describing their own health. This was also the reason for not asking the replica vignettes in the wave 1 interview.

In the survey, we distinguish six domains: sleep, mobility, concentration and memory, breathing, affect (depression and mood swings), and pain. The analysis for the pain domain is more complicated due to selection and the fact that different types of pain are distinguished, and is left for future research. Full details of the questionnaires in both waves are presented in Appendix 3.C. Here we present two examples: sleep and concentration.

Sleep

In wave 1, we first asked the usual self-assessment question on sleep related problems, also used in, for example, the World Health Survey (WHS) and the Survey of Health, Ageing and Retirement (SHARE):

$Sleep_{SA}$ Overall during the last 30 days, How much difficulty have you had with sleeping, such as falling asleep, waking up frequently during the night or waking up too early in the morning? None, mild, moderate, severe, or extreme?

Then we asked three questions on different aspects of sleep: falling asleep, waking up during the night, and feeling well rested in the morning, with the idea that these three should give a complete picture of sleep related health problems:

$Sleep_1$ Please indicate which of the following best describes your own situation during the last 30 days:

1. When I go to bed at night I always immediately fall asleep
2. When I go to bed at night I usually fall asleep immediately but sometimes, at most once a week, it takes me more than an hour.
3. It usually takes me some time to fall asleep, like half an hour or more
4. It almost always takes me an hour or more to fall asleep
5. It usually takes me a few hours to fall asleep

6. I hardly sleep at all

$Sleep_2$ Please indicate which of the following best describes your own situation during the last 30 days:

1. Once I am asleep I don't wake up until it is time to get out of bed.
2. I occasionally wake up during the night but then easily fall asleep again.
3. I often wake up during the night and then it is sometimes hard to fall asleep again.
4. I often wake up in the middle of the night and then usually do not fall asleep again until the morning
5. I never sleep more than three or four hours and remain awake the rest of the night

$Sleep_3$ Please indicate which of the following best describes your own situation during the last 30 days:

1. I always sleep well enough to feel completely well-rested in the morning
2. I sometimes do not feel well-rested in the morning but this is because I have to wake up early or go to bed too late, not because I cannot sleep
3. I usually feel well-rested in the morning but once a month or so, I cannot sleep well and do not feel well rested when I get up
4. I often feel well-rested in the morning but once or twice a week, I cannot sleep well and do not feel well rested when I get up
5. I usually do not feel well-rested in the morning, since I do not sleep well enough
6. I never feel well-rested in the morning, since I never sleep well

In the wave 2 interview, we again asked the self-assessment question $Sleep_{SA}$, now followed by three vignette questions, with vignettes on sleep problems. One of the three is the replica vignette, combining the answers given in wave 1 to

questions $Sleep_1$, $Sleep_2$ and $Sleep_3$. For example, a respondent whose wave 1 answers were 3 to the $Sleep_1$, $Sleep_2$, as well as $Sleep_3$, got the following replica vignette:

$Sleep_{RV}$ It usually takes John some time to fall asleep, like half an hour or more. He often wakes up during the night and then it is sometimes hard to fall asleep again. He usually feels well-rested in the morning but once a month or so, he cannot sleep well and does not feel well rested when he gets up.
Overall in the last 30 days, how much difficulty does John have with sleeping?

None, mild, moderate, severe, or extreme?

The hypothetical person in this vignette ("John") exactly has the same sleep related health problems as those reported by the respondent in wave 1. If the three aspects of sleep considered (falling asleep, waking up during the night, well-rested in the morning) completely characterize sleep related health, if response scales do not vary from one wave to the next, if no reporting errors are made, and if answers to vignette and self-assessment questions use the same response scales (response consistency), then the evaluations of the replica vignette in wave 2 should be identical to the respondent's self-assessment in wave 1. This is the intuition behind the test that we will perform: maintaining the other, auxiliary, assumptions, respondents should evaluate a vignette person's health in the same way as their own health if the vignette describes exactly their own health. Under the somewhat weaker assumption that reporting errors in the form of misclassifying sleep related health status is possible but misclassification probabilities are the same for self-assessments and vignettes, self-assessments and replica vignette evaluation no longer need to be identical for each respondent, but their marginal distributions of self-assessments and replica vignette evaluations should be the same. The latter is the basis of our nonparametric tests in Section 3.3.

The other two (non-replica) vignettes are constructed using different combinations of the possible answers to questions $Sleep_1$, $Sleep_2$ and $Sleep_3$ than the combination used for the replica vignette. Some randomization is involved but implausible combinations of the three answers are avoided. Since these vignettes will not be used in the analysis in this chapter, details are not discussed.

Concentration and memory

In principle, the other domains are treated in the same way, but details differ because the challenge of describing health in a given domain by a small

number of aspects varies across domains. We therefore provide details of one additional domain, concentration and memory, where selecting the descriptors seems less straightforward than for sleep. The first question is again the usual self-assessment:

$Conc_{SA}$ Overall in the last 30 days, how much of a problem did you have with concentrating or remembering things? None, mild, moderate, severe, or extreme?

We then asked six questions in which respondents could describe their own memory and concentration problems as completely as possible on an objective scale:

$Conc_1$ When a friend introduces you to five people you never met before and you have a polite conversation with these people for just a few minutes, how many of their names would you still remember the next day? 0, 1, 2, 3, 4 or 5?

$Conc_2$ And a week later? 0, 1, 2, 3, 4 or 5?

$Conc_3$ When you watch the news with full concentration, and ten news items are presented, how many of these do you think would you still remember an hour later? 0, 1, 2, . . . , or 10?

$Conc_4$ And the next day? 0, 1, 2, . . . , or 10?

$Conc_5$ How often do you have to look for your keys, wallet, glasses, or similar things you daily use, since you don't know where you last put them?

1. Never
2. At most once a month
3. Between one and four times a month
4. Once or twice a week
5. More than twice a week but not every day
6. About once a day
7. More than once a day

$Conc_6$ How often do you go out and then realize later that you did not take everything you needed with you, like your wallet, your keys, the letter you wanted to post, the coupons you wanted to exchange at the supermarket, etc.?

1. Never
2. At most once a month
3. Between one and four times a month
4. Once or twice a week
5. More than twice a week but not every day
6. At least once a day, if I go out
7. If I go out, I almost always forget something

In wave 2, the self-assessment question is repeated, followed by three vignette questions, one of which is the replica vignette, combining the wave 1 answers to the questions $Conc_1, \ldots Conc_6$. For example, for a respondent with wave 1 answers $Conc_1 = 3$, $Conc_2 = 2$, $Conc_3 = 6$, $Conc_4 = 4$, $Conc_5 = 3$ and $Conc_6 = 3$, the replica vignette question is as follows (where the parts in brackets indicate what is taken from the wave 1 answers):

$Conc_{RV}$ When a friend introduces Jane to five people she has never met before and Jane has a polite conversation with these people for just a few minutes, Jane still remembers [three] of the five names the next day. One week later, she still remembers [two] of them. When Jane watches the news with full concentration, and ten news items are presented, Jane still remembers [six] of them an hour later. The next day, she still remembers [four] of them. [Between one and four times a month], Jane has to look for her keys, wallet, glasses, or similar things she uses daily, since she doesn't know where she last put them. [Between one and four times a month] Jane goes out and then realizes later that she did not take everything she needed with her, like her wallet, her keys, or the letter she wanted to post.
How much of a problem does Jane have with concentrating or remembering things?

The other two vignettes combine different possible answers to the questions

$Conc_1, \ldots, Conc_6$ into similar vignette descriptions, involving some randomization but avoiding implausible combinations.

We first ask all the self-assessments and then the vignette questions.[4] Details on the other three domains (mobility, breathing, and affect) are provided in Appendix 3.C.

3.3 Descriptive Statistics and Nonparametric Tests

Table 3.1 presents the frequency distribution of the self-assessments and the replica vignette evaluations in wave 1. The self-assessments (columns "self") show that respondents express the most personal difficulty with sleep, followed by affect and concentration. The other columns ("vign") refer to the evaluations of the replica vignettes. In some cases, the distributions of answers to the self-assessments and vignettes are close. These domains would include sleep, mobility, and affect. The largest differences are found for concentration and breathing. In both of these cases, the evaluations would suggest that the problems of the persons described in the replica vignettes are, on average, substantially more serious than the respondents' own problems.

Table 3.2 displays joint distributions of self-assessed difficulty (the columns) and responses to the replica vignette question (the rows) for the five domains. The fact that the majority of the observations is on the diagonal or only one category off the diagonal is reassuring. The diagonals in each panel represent cases in which responses for self-assessments and replica vignettes are identical. The fact that non-diagonal frequencies are not zeros may be due to several causes, including reporting errors in the self-assessments, in the vignette evaluations, or in the answers to the objective health questions used to construct the replica vignettes. This in itself does not provide evidence against response consistency in the sense that models such as the chopit model (King *et al.* (2004)) allow for random errors in the self-reports and the thresholds translating "true" health in

[4]Hopkins and King (2010) report experiments showing that placing vignettes before self-reports substantially improves the fit of models explaining the self-reports. We have not followed that practice for three reasons. First of all, until now typically self-reports are asked first and hence our test seems most relevant for current practice. Secondly, in principle one can use one sample to estimate vignette models and then use the result to correct self-reports in a different sample. That approach becomes infeasible if corrections are done based on models where vignettes have to be placed before self-reports. Third, order can play a role and presenting vignettes before self-reports may lead to systematic biases in the self-assessments. Put differently, the vignettes will anchor the meaning of the question about the self-report, so that the self-report now becomes incomparable with data from other surveys that do not precede the self-report by the same anchoring vignettes.

a finite scale.

One way to gauge how much responses may change over time due to idiosyncratic reporting errors is to also consider the distribution of self assessments in wave 2. Table 3.3 summarizes the correspondence between the various measures by means of correlation coefficients for the five domains (treating the responses as cardinal). For sleep, the correlation between wave 1 self-assessment and replica vignette evaluation is higher than the correlation of either of these with the wave 2 self-assessment. For the other domains, however, the correlation between the two self-assessments is higher. This suggests that the replica vignette does a better job in describing actual problems in the sleep domain than in the other domains. Particularly for concentration, the relation between replica vignette evaluations and wave 1 (or wave 2) self-assessments is low.

The results of various tests of the null hypothesis that the population distributions of self-assessments and replica vignette evaluations are the same are presented in Table 3.4. The first test is a Wilcoxon signed rank test, which compares the marginal distributions in Table 1, accounting for the matched nature of the observations (see, e.g., Siegel and Castellan Jr. (1988)). The second test is the sign test that tests the weaker hypothesis that the median of the difference between self-assessments and replica vignette evaluations is equal to zero. Both tests lead to the same conclusions: the null hypothesis is not rejected for sleep (p-value 0.23), but is clearly rejected for the other four domains (p-values 0.00, except for the sign test for mobility which gives p-value 0.02). These results are in line with what we saw in Table 1: the frequencies of self-assessments and replica vignette evaluations are much more similar for sleep than for the other domains.

It is important to note that the null hypotheses tested by these tests are much more stringent than mere response consistency. Consider the folowing simple example. Let the true health condition in a domain be distributed as $Y_s^* \sim N(0,1)$ where $N(0,1)$ is the standard normal distribution. We observe self-reports Y_s, which are generated by the following observation scheme: $Y_s = j \Leftrightarrow \tau^{j-1} < Y_s^\star \leq \tau^j \quad j = 1,2,3,4$, with $\tau^0 = -\infty$ and $\tau^4 = \infty$. Assume that the true evaluations of replica vignettes are generated by $Y_v^* \sim N(0,\sigma^2)$ and that the reported evaluations of the vignettes Y_v are generated by exactly the same observation scheme as Y_s. It is obvious that response consistency holds, since the thresholds τ^j are the same for the self-reports and the vignette evaluations (moreover they don't vary across respondents, so there is no DIF, but that is not the point of the example). The only difference is that the vignette evaluations are possibly noisier ($\sigma > 1$) or less noisy ($\sigma < 1$). The case $\sigma > 1$

is probably the most relevant case since the vignette descriptions are likely to be less complete than a respondent's knowledge of her own health condition. In view of the skewed distribution of the observed self-reports an increase in noise will shift the empirical distribution to the right, which is what we see in four out of five domains (affect being the exception). As a matter of fact we can use this simple model to see what value of σ would generate the variance that we see in the empirical distribution of the vignettes, assuming that the thresholds are indeed the same for the self-reports and the vignette evaluations[5]. We find the following values of σ: sleep: 1; mobility: 1.19; concentration and remembering things: 1.26; breathing: 1.18; affect: 1.01. For these values of σ we do indeed see a shift of the empirical distribution to the right, although typically not as much as in the actual data.

Apart from this, a potential explanation for the rejection could be order effects in vignette evaluations, in the sense that a vignette evaluation would be affected by the nature of the previous vignette. To investigate that explanation, we repeated the test for the subsamples of those who got the replica vignette before they got the other two vignettes on the same domain, exploiting the fact that the order was randomized. The results are in the second panel of Table 3.4. For this subsample, the null hypothesis is not rejected for sleep nor for mobility. For the other domains, however, the null once again gets rejected.[6] Since we did not randomize the order of vignettes across domains, we cannot check whether anchoring effects caused by vignettes in another domain play a role. We always presented sleep first, followed by pain, mobility, concentration, breathing, and affect. This might be one reason why we find the best results for sleep.

What do these results imply for the validity of the response consistency assumption? As explained above, a number of auxiliary assumptions needs to be made to interpret the tests as tests for response consistency only. Order effects in replica vignette evaluations were taken into account in Table 3.4 – they clearly cannot explain all rejections.

The most important maintained assumption is probably that the objective questions indeed give an adequate and complete description of health problems in the given domain. This assumption is more likely to hold for sleep than for domains like concentration and memory, where it seems much more difficult

[5]Let the cumulative frequencies of the self-reports be denoted by p_1, p_2, p_3 ($p_4 = 1$). Then the corresponding cumulative frequencies for the vignettes are generated as $q_i = N[\{N^{-1}(p_i, 1)/\sigma\}, 1]$

[6]For completeness, we also performed the tests for the subsamples of observations where the replica vignette was *not* presented first. Here we found that the null was not rejected for sleep (p-values for the two tests are 0.08 and 0.13).

to describe potential problems with a few objective questions (notice that in the illustrative exercise above, this domain generates the highest value of σ). An alternative interpretation of the results for concentration could therefore be that our objective questions $Conc_1, \ldots, Conc_6$ do not adequately describe the concentration and memory problems respondents have in mind when answering the concentration and memory self-assessment question. Perhaps the vignette descriptions on concentration and memory are also simply too long for the respondents to read them carefully.

Moreover, several types of reporting errors may play a role. As noted, if evaluations of vignettes are noisier than self-assessments, then this is captured by different error variances in models such as the chopit model, and the null hypothesis of equal marginal distributions no longer holds. Reporting errors in the objective questions could play a role, since they will not affect the self-assessments but they will influence the nature of the replica vignettes and their evaluations. Since most respondents report to be quite healthy, response errors will tend to shift reported health conditions in the direction of worse health. This would shift the constructed vignettes in the direction of worse health.

Finally, response consistency means that respondents use the same thresholds for the evaluations of their own health and the replica vignette. If response consistency is rejected, the question can be raised which thresholds cause the problem. To analyze this, we redid the tests after grouping the outcomes in binary categories. For example, to test whether the thresholds between "none" and "mild" (the two most prevalent outcomes) are different, we combined outcomes mild and worse into one category and repeated the tests. In this case, the null hypothesis was not rejected for sleep or mobility, but it was rejected for the other three domains (details available upon request).

3.4 Parametric models

All in all, there are several additional assumptions underlying the tests and as many alternative reasons why the tests so often reject. More insight in some of these can be obtained by considering parametric models, which, for example, can capture different noise levels in self-assessments and replica vignettes.

In this section we present a formal statistical model explaining both subjective qualitative self-assessments as well as vignette evaluations of hypothetical people with possible health problems that generalizes the chopit model and its extensions that are typical for the sort of models that have been used in this context (King *et al.* (2004); Kapteyn *et al.* (2007)).

3.4.1 Self-assessments

The subjective self-assessment (Y_{si} for respondent i) in a given domain is assumed to be driven by an underlying latent index reflecting actual health in that domain, and individual specific thresholds:[7]

$$
\begin{aligned}
Y_{si}^{\star} &= \beta_s T_{si} + \delta_s X_i + \epsilon_{si} && (3.1)\\
Y_{si} = j &\Leftrightarrow \tau_{si}^{j-1} < Y_{si}^{\star} \leq \tau_{si}^{j} \quad j = 1,2,3,4 && (3.2)\\
\tau_{si}^{0} &= -\infty && (3.3)\\
\tau_{si}^{1} &= \gamma_s^1 X_i + \lambda_s^1 T_{si} + u_i && (3.4)\\
\tau_{si}^{2} &= \tau_{si}^{1} + \exp\left(\gamma_s^2 X_i + \lambda_s^2 T_{si}\right) && (3.5)\\
\tau_{si}^{3} &= \tau_{si}^{2} + \exp\left(\gamma_s^3 X_i + \lambda_s^3 T_{si}\right) && (3.6)\\
\tau_{si}^{4} &= \infty, && (3.7)
\end{aligned}
$$

Y_{si}^{*} is a latent variable describing "true" health problems in the given domain; T_{si} is a vector describing the same health problems of individual i in terms of the objective questions (like $sleep_1, \ldots, sleep_3$), and X_i contains a set of other observed respondent characteristics. X_i should not play a role (i.e. δ_s should be zero) if the given health domain is adequately captured by the objective questions in T_{si}, but, in general, the variables in X_i may be interpreted as proxies for unobserved heterogeneity in health problems not covered by T_{si}. The idiosyncratic error term ϵ_{si} is assumed to affect the subjective self-report but nothing else. We assume that $\epsilon_{si} \sim N(0, \sigma_s^2)$, independent of T_{si} and X_i. Equation (3.2) describes the usual observation function that translates values of the latent variable Y_{si}^{*} into categorical values Y_{si}, using the cut-off points (or thresholds) τ_{si}^{j}, $j = 0, ..., 4$. Equations (3.3)-(3.7) parameterize the cut-off points τ_{si}^{j} as a function of observables and of an unobserved heterogeneity term u_i. The exponentials guarantee that cut-off points are in the right order.

The fact that different respondents i use different response scales (different cut-off points) represents DIF. Using subjective self-reports on own health problems only, parameters β_s, δ_s, γ_s^1, λ_s^1 are not separately identified; only their difference is identified. On the other hand, the γ_s^j, λ_s^j for $j > 1$ will still be identified. Below we will discuss identification of the parameters in this model in more detail.

[7] As before, the answers "severe" and "extreme" are merged into one category, so that we work with four possible outcomes for all vignettes and self-assessments.

3.4.2 Replica Vignette Evaluations

The evaluation of the replica vignette is modeled using an ordered response equations similar to (3.1)-(3.7):

$$
\begin{aligned}
Y^{\star}_{vi} &= \beta_v T_{si} + \delta_v X_i + \epsilon_{vi} && (3.8)\\
Y_{vi} = j &\Leftrightarrow \tau^{j-1}_{vi} < Y^{\star}_{vi} \leq \tau^{j}_{vi} \quad j = 1,2,3,4 && (3.9)\\
\tau^{0}_{vi} &= -\infty && (3.10)\\
\tau^{1}_{vi} &= \gamma^1_v X_i + \lambda^1_v T_{si} + u_i && (3.11)\\
\tau^{2}_{vi} &= \tau^{1}_{vi} + \exp\left(\gamma^2_v X_i + \lambda^2_v T_{si}\right) && (3.12)\\
\tau^{3}_{vi} &= \tau^{2}_{vi} + \exp\left(\gamma^3_v X_i + \lambda^3_v T_{si}\right) && (3.13)\\
\tau^{4}_{vi} &= \infty, && (3.14)
\end{aligned}
$$

Because of the design of the replica vignette, the health variables are the T_{si} reported in wave 1. Respondent characteristics X_i should not play any role under *vignette equivalence*, the assumption that all respondents interpret the genuine health of a given hypothetical person in the same way. In the context of the current model, vignette equivalence can therefore be formulated as:

$$\delta_v = 0 \tag{3.15}$$

In the standard setting, a few fixed vignettes are shown to all respondents, and accordingly the chopit model has a dummy for each vignette, without any restrictions on coefficients of these dummies and the coefficients β_s that drive how genuine health depends on the objective conditions. In the current setting, however, we aim at vignettes replicating the respondent's health. In the model, the assumption that the answers to our objective questions T_{si} indeed perfectly capture health in the given domain implies:

$$\beta_s = \beta_v, \delta_s = 0 \tag{3.16}$$

This is an additional assumption that is not required in the standard chopit model correcting for DIF, simply because there does not have to be connection between a respondent's own health and the health of a vignette person. If satisfied, it leads to the over-identification that makes it possible to test response consistency.

The assumption we want to test is *response consistency*:

$$\text{RC: } \gamma^j_s = \gamma^j_v, \lambda^j_s = \lambda^j_v,\ j = 1,2,3 \tag{3.17}$$

Without imposing either (3.15) or (3.16), we cannot test (3.17), since the parameters in (3.17) are not identified. The reason is that in this unrestricted model we can identify $\lambda_s^1 - \beta_s$, $\lambda_v^1 - \beta_v$, $\gamma_s^1 - \delta_s$ and $\gamma_v^1 - \delta_v$, but not the individual parameters λ_s^1, β_s, λ_v^1, β_v, γ_s^1, δ_s, γ_v^1 and δ_v. On the other hand, the parameters γ_s^2, γ_v^2, λ_s^2, λ_v^2, γ_s^3, γ_v^3, λ_s^3 and λ_v^3 are always identified. See Appendix 3.B.

The equalities $\gamma_s^1 = \gamma_v^1$ and $\lambda_s^1 = \lambda_v^1$ can therefore not be tested without additional assumptions on β_s, β_v, δ_s, and δ_v.

Put differently, the following equalities can be tested without additional assumptions:

$$\text{RC1} \quad : \quad \lambda_s^1 - \beta_s = \lambda_v^1 - \beta_v,\ \gamma_s^1 - \delta_s = \gamma_v^1 - \delta_v \tag{3.18}$$

$$\text{RC2} \quad : \quad \gamma_s^2 = \gamma_v^2,\ \lambda_s^2 = \lambda_v^2 \tag{3.19}$$

$$\text{RC3} \quad : \quad \gamma_s^3 = \gamma_v^3,\ \lambda_s^3 = \lambda_v^3 \tag{3.20}$$

Under the maintained additional assumptions (3.15) and (3.16), we have $\beta_s = \beta_v$ and $\delta_s = \delta_v (= 0)$, so that (3.18) is equivalent to $\lambda_s^1 = \lambda_v^1$ and $\gamma_s^1 = \gamma_v^1$ and (3.18), (3.19) and (3.20) together are equivalent to the response consistency assumption (3.17) we want to test.

We will test (3.18), (3.19) and (3.20) jointly, but will also test (3.19) and (3.20) jointly, without imposing (3.18). The discussion above implies that the first test requires the maintained assumptions (3.15) and (3.16), and rejecting the null hypothesis may imply that response consistency is not satisfied, but may also imply that vignette equivalence is not satisfied or that our objective questions are insufficient to capture the health problems in the given domain. On the other hand, rejecting (3.19) and (3.20) with the second test certainly would mean response consistency is not satisfied. But the second has the drawback that it only has power for certain violations of response consistency, and not against violations of (3.18).

3.4.3 Test Results

Table 3.5 presents log likelihoods and likelihood ratio tests for restricted and unrestricted versions of the model (3.1)-(3.7) and (3.8)-(3.14).[8] We present tests of three hypotheses: (1) all equalities in (3.18)-(3.20) hold (denoted by $\forall j$); (2) equalities (3.19)-(3.20) hold (denoted by $j > 1$); (3) equation (3.18) holds

[8]Since the unrestricted model is not identified, we need some normalizations. These normalizations do not affect the value of the log-likelihood. We have chosen $\sigma_s = \sigma_v = 1$ and otherwise taken $\lambda_s^1 - \beta_s$, $\lambda_v^1 - \beta_v$, $\gamma_s^1 - \delta_s$ and $\gamma_v^1 - \delta_v$ as reduced form parameters in the estimation.

(denoted by $j = 1$). We present the value of the log-likelihood (first line) of the unrestricted model and the restricted models corresponding to each of the tests, the number of parameters estimated in each of these models, (second line) and the p-value of the test (third line).

Tables 3.5 shows that sleeping is the only domain for which all three equalities are accepted for the generic model (the line "all"). For the other domains, the joint hypotheses (3.18)-(3.20) is rejected (both the columns $\forall j$ and $j = 1$). We note however that for the case $j > 1$, the null gets accepted for all five domains.

3.5 Conclusions

Showing respondents "their own vignette" seems a natural approach to testing for response consistency. Potentially it avoids some pitfalls of other approaches, like relying on "objective" measures, as in Kapteyn *et al.* (2007). The test relies on fewer assumptions and is more direct. Having done the experiment however, a number of potential improvements to our approach have presented themselves. First of all, as the discussion of order effects has suggested, a proper test would seem to require that the replica vignette is always placed first in the vignette question sequence. Secondly, to further avoid spill-overs and context effects it is probably advisable to test vignette equivalence and response consistency one domain at a time. Third, in our experiment we have measured the vector of health conditions at baseline, but not in the second wave. Thus we have had to insert the baseline values for T_{si} in the equations for the threshold values in (3.11)-(3.13). To the extent that health has changed between waves, this would introduce measurement error in the health vector. Fourth, as may be clear from Appendix 3.C, construction of the replica vignettes in an automated fashion is not entirely straightforward and further improvements may add to the accuracy of the replica vignettes as descriptions of respondents' health.

We started out to test response consistency, but the results so far suggest that possibly vignette equivalence ($\delta_v = 0$) is a much more fragile assumption than response consistency. Similarly, the test of response consistency requires $\delta_s = 0$. Both $\delta_v = 0$ and $\delta_s = 0$, are more likely to hold true if the description of the vignette person's condition T_v is complete. It seems therefore that future efforts should be directed at improving vignette descriptions and extensive testing before they are used in practice.

It is of interest to compare our approach with the approach adopted by Bago d'Uva *et al.* (2009), using the parametric framework developed in Section 3.4. They carry out two main tests. The first test assumes that in their data

from the English Longitudinal Study of Aging the vector T_s in (3.1) provides a complete description of the respondent's own health. This allows them to assume $\delta_s = 0$. Since, moreover they impose $\lambda_s^j = 0$ for all j, both β_s and γ_s^j (for all j) are identified. They then compare the estimates of γ_s^j and γ_v^j and reject the null that the two vectors are identical for the domains considered (cognition and mobility). Rather than our exponential specification in (3.5)-(3.6) and (3.12)-(3.13) they adopt a linear specification. They also perform a weaker test of response consistency similar to our test for $j > 1$ and find that response consistency is rejected for mobility, but not for cognition. As they note, rejection of response consistency may indicate that it is a false assumption, but it is also possible that the restriction $\delta_s = 0$ does not hold. The authors also perform a test of vignette equivalence, exploiting within person comparisons of vignette evaluations. The model they consider is (largely in our notation):

$$Y_{vi}^{1*} = \theta_1 + \varepsilon_{vi}^1 \tag{3.21}$$

$$Y_{vi}^{k*} = \theta_k + \delta_v^k X_i + \varepsilon_{vi}^k \quad k = 2, ..., K \tag{3.22}$$

where K is the number of vignettes in a domain shown to respondent i. The latent variables Y_{vi}^{k*} $(k = 1, ...K)$ are translated into observable responses using (3.9). This model is identified due to the absence of X_i in equation (3.21) and obviously assumes that respondents use the same thresholds for different vignettes. Under vignette equivalence $\delta_v^k = 0$ for $k = 2, ...K$. The idea is that the difference in evaluations of different vignettes should not vary systematically across individuals. This can be tested by testing $\delta_v^k = 0$, $k = 2, ...K$. Vignette equivalence gets rejected for both cognition and mobility.

Both our results and the results of Bago d'Uva *et al.* (2009) suggest the need for further work on the design of vignettes. For vignette equivalence to hold, a description has to be complete, minimizing room for different interpretations by different respondents. On the other hand, descriptions have to be concise, as otherwise it is likely that a respondent will carefully read the description. Designing concise and yet complete vignette descriptions is clearly challenging and one needs an experimental environment, such as used in this chapter, to determine whether one has been successful.

3.A Tables

Table 3.1: Frequency Distributions % of Wave 1 Self-assessments and Wave 2 Replica Vignette Evaluations

	1		2		3		4/5		
domain	self	vign	self	vign	self	vign	self	vign	Obs.
sleep	25.0	26.6	39.3	37.6	27.1	28.0	8.6	7.7	1615
mobility	57.5	58.6	27.4	21.6	11.6	13.9	3.5	6.0	1613
concentration	41.2	30.2	44.0	39.9	12.5	23.9	2.3	6.0	1610
breathing	69.5	50.0	21.8	34.7	6.8	11.8	1.9	3.5	1609
affect	39.5	44.8	41.1	36.9	14.2	13.2	5.3	5.0	1610

Notes:
Frequencies in % of total number of observations (Obs.).
The self-assessments were formulated as: "Overall in the last 30 days, how much of a problem did you have with concentrating or remembering things?" with answers "none"(1), "mild"(2), "moderate"(3), "severe"(4), and "extreme"(5).
The replica vignette questions are the same, but with "you" replaced by a hypothetical name.
Frequencies for severe and extreme are combined because of the small numbers reporting these outcomes.

Table 3.2: Cross Tables of Wave 1 Self-assessments and Wave 2 Replica Vignettes

	sleep					mobility				
	self1					self1				
vign	1	2	3	4/5	vign	1	2	3	4/5
1	60.4	23.6	7.1	3.6	1	74.2	46.4	25.1	7.1
2	31.4	51.0	32.7	10.1	2	17.3	30.1	24.1	16.1
3	7.2	23.8	49.7	40.3	3	7.8	19.5	26.7	28.6
4/5	1.0	1.6	10.5	46.0	4/5	0.6	4.1	24.1	48.2

	concentration					breathing				
	self1					self1				
vign	1	2	3	4/5	vign	1	2	3	4/5
1	43.8	24.4	9.5	8.1	1	60.8	32.0	10.0	6.7
2	36.1	46.8	32.3	18.9	2	32.4	41.1	40.9	20.0
3	16.3	24.0	47.3	32.4	3	5.8	22.9	31.8	33.3
4/5	3.8	4.8	10.9	40.5	4/5	1.0	4.0	17.3	40.0

	affect			
	self1			
vign	1	2	3	4/5
1	66.8	38.3	18.9	1.2
2	28.1	46.7	36.4	27.1
3	4.4	13.2	32.5	28.2
4/5	0.6	1.8	12.3	43.5

Note: Columns present relative frequencies in %.

Table 3.3: Correlations between Wave 1 Self-assessments, Replica Vignettes and Wave 2 Self-assessments

	sleep	mobility	breathing	concentration	affect
self1, vign	0.59	0.51	0.46	0.33	0.53
self1, self2	0.58	0.62	0.64	0.61	0.59
self2, vign	0.48	0.47	0.41	0.31	0.41

Table 3.4: Nonparametric Tests of Response Consistency

	all		replica vign first	
	Wilcoxon test	sign test	Wilcoxon test	sign test
sleep	0.23	0.26	0.68	0.9
mobility	0	0.02	0.37	0.73
concentration	0	0	0	0
breathing	0	0	0	0
affect	0	0	0	0.03

Note: The null of Wilcoxon sign rank test is that the difference between wave 1 self-assessments and replica vignette evaluations is symmetric about zero. The null of sign test is that the true median of the difference between self-assessments and replica vignette evaluations is equal to zero. The p-values of the tests are presented for the whole sample (columns "all ") and for the subsample who got replica vignette before two other vignettes (columns "replica vign first ").

Table 3.5: Summary of Estimated Parametric Models and Tests of Response Consistency

	unrestricted	threshold pars equal for		
		$\forall j$	$j > 1$	$j = 1$
sleep	-3035.63	-3058.75	-3046.87	-3040.254
	91	47	62	77
		0.38	0.80	0.81
mobility	-2543.42	-2627.81	-2561.75	-2574.21
	97	50	66	82
		0	0.22	0
concentration	-3170.39	-3291.86	-3192.49	-3198.66
	109	56	74	92
		0	0.14	0
breathing	-2374.27	-2471.42	-2393.30	-2444.28
	97	50	66	82
		0	0.18	0
affect	-2876.66	-2929.07	-2886.50	-2899.37
	85	44	58	72
		0	0.84	0

Note: Tests of three hypotheses are presented: (1) all equalities in (3.18)-(3.20) hold (denoted by $\forall j$); (2) equalities (3.19)-(3.20) hold (denoted by $j > 1$); (3) equation (3.18) holds (denoted by $j = 1$). We present the value of the log-likelihood (first line) of the unrestricted model and the restricted models corresponding to each of the tests, the number of parameters estimated in each of these models, (second line) and the p-value of the likelihood ratio test (third line).

3.B Identification

When conditions (3.15) and(3.17) are not imposed, the models are no longer identified. It is worth considering this in more detail.

Define

$$\Omega_s^1 \equiv (\lambda_s^1 - \beta_s)T_{si} + (\gamma_s^1 - \delta_s)X_i$$

Then we have

$$\Pr(Y_{si} = 1) = \Pr\left[\epsilon_{si} - u_i \leq \Omega_s^1\right] \tag{3.23}$$

$$\Pr(Y_{si} = 2) = \Pr\left[\Omega_s^1 < \epsilon_{si} - u_i \leq \Omega_s^1 + \exp\left(\gamma_s^2 X_{si} + \lambda_s^2 T_{si}\right)\right] \tag{3.24}$$

$$\Pr(Y_{si} = 3) = \Pr\left[\begin{array}{c} \Omega_s^1 + \exp\left(\gamma_s^2 X_i + \lambda_s^2 T_i\right) < \epsilon_{si} - u_i \leq \\ \Omega_s^1 + \exp\left(\gamma_s^2 X_i + \lambda_s^2 T_i\right) + \exp\left(\gamma_s^3 X_i + \lambda_s^3 T_{si}\right) \end{array}\right] \tag{3.25}$$

$$\Pr(Y_{si} = 4) = \Pr\left[\epsilon_{si} - u_i > \Omega_s^1 + \exp\left(\gamma_s^2 X_i + \lambda_s^2 T_i\right) + \exp\left(\gamma_s^3 X_i + \lambda_s^3 T_{si}\right)\right] \tag{3.26}$$

And similarly for the replica vignettes (with Ω_v^1 defined similarly to Ω_s^1):

$$\Pr(Y_{vi} = 1) = \Pr\left[\epsilon_{vi} - u_i \leq \Omega_v^1\right] \tag{3.27}$$

$$\Pr(Y_{vi} = 2) = \Pr\left[\Omega_v^1 < \epsilon_{vi} - u_i \leq \Omega_v^1 + \exp\left(\gamma_v^2 X_i^2 + \lambda_v^2 T_i\right)\right] \tag{3.28}$$

$$\Pr(Y_{vi} = 3) = \Pr\left[\begin{array}{c} \Omega_v^1 + \exp\left(\gamma_v^2 X_i^2 + \lambda_v^2 T_i\right) < \epsilon_{vi} - u_i \leq \\ \Omega_v^1 + \exp\left(\gamma_v^2 X_i^2 + \lambda_v^2 T_i\right) + \exp\left(\gamma_v^3 X_i^2 + \lambda_v^3 T_i\right) \end{array}\right] \tag{3.29}$$

$$\Pr(Y_{vi} = 4) = \Pr\left[\epsilon_{vi} - u_i > \Omega_v^1 + \exp\left(\gamma_v^2 X_i^2 + \lambda_v^2 T_i\right) + \exp\left(\gamma_v^3 X_i^2 + \lambda_v^3 T_i\right)\right] \tag{3.30}$$

Subject to some minor normalizations, we can thus estimate $\lambda_s^1 - \beta_s$, $\lambda_v^1 - \beta_v$, $\gamma_s^1 - \delta_s$, $\gamma_v^1 - \delta_v$, γ_s^2, γ_v^2, λ_s^2, λ_v^2, γ_s^3, γ_v^3, λ_s^3, λ_v^3.

3.C More details about mobility, breathing and affect

Self-assessment questions, descriptions of the health and replica vignette for mobility, breathing and affect are presented in this appendix.

Mobility

The self-assessment question on mobility related problems:

Mob_{SA} Overall in the last 30 days, how much of a problem did you have with moving around? None, mild, moderate, severe, or extreme?

Two questions on different aspects of mobility:

Mob_1 Please indicate which of the following best describes your own situation:

1. I have no problems walking four miles and I actually sometimes go for a long walk
2. I would have no problems with walking three or four miles if I had to
3. I can walk one or two miles but I would have problems going farther than that without taking a rest
4. I can walk about half a mile without any problems but after that I feel tired and need to rest
5. I can walk two blocks without problems but feel tired when I walk farther than that
6. Moving around at home is OK for me but my health prevents me from going for more than a very short walk outside
7. I have to make an effort to move around my home
8. My health prevents me from moving around my home.

Mob_2 Please indicate which of the following best describes your own situation:

1. I can climb five sets of stairs in a row without getting tired
2. I can climb two or three flights of stairs in a row but then I need a little rest to recover
3. I can climb one flight of stairs but then I need some time to recover
4. I can climb one flight of stairs but I have to stop and take a little rest once or twice
5. Climbing one flight of stairs is a large effort for me and I have to take several breaks
6. I am not able to climb one flight of stairs

In wave 2, the replica vignette is asked. For example, for a respondent with wave 1 answers $Mob_1 = 3$ and $Mob_1 = 2$ the replica vignette is as follows:

Mob_{RV} Ruth can walk one or two miles but she would have problems going farther than that without taking a rest. She can climb two or three flights of stairs in a row but then she needs a little rest to recover.
Overall in the last 30 days, how much of a problem did she have with moving around?

Breathing

The self-assessment question on breathing related problems:

$Breath_{SA}$ Overall in the last 30 days, how much of a problem did you have because of shortness of breath? None, mild, moderate, severe, or extreme?

Three questions on different aspects of breathing:

$Breath_1$ Please indicate which of the following best describes your own situation:

1. I can jog for at least 15 minutes without getting short of breath.
2. I get out of breath when jogging, but I have no trouble walking at a brisk pace.

3. As long as I don't walk too fast, I don't get out of breath.

4. I get out of breath easily and can only walk slowly.

$Breath_2$ Please indicate which of the following best describes your own situation:

1. I never have respiratory infections, like pneumonia, bronchitis, or the flu (influenza).

2. Once every couple of years I have a respiratory infection.

3. About once a year I have a respiratory infection.

4. I have a respiratory infection more than once a year.

$Breath_3$ Please indicate which of the following best describes your own situation:

1. I cough a lot and am short of breath 3 or 4 times a week.

2. I cough a lot and am short of breath about once a week.

3. Sometimes I cough a lot and am short of breath about once a month.

4. Sometimes I cough a lot, but I am rarely short of breath (not more than once a year).

5. I rarely cough and am never out of breath.

In wave 2, the replica vignette is asked. For example, for a respondent with wave 1 answers $Breath_1 = 3$, $Breath_2 = 2$ and $Breath_3 = 4$ the replica vignette is as follows:

$Breath_{RV}$ As long as John doesn't walk too fast, he doesn't get out of breath. Once every couple of years he has a respiratory infection. Sometimes he coughs a lot, but he is rarely short of breath (not more than once a year).
Overall in the last 30 days, how much of a problem did he have because of shortness of breath?

Affect

The self-assessment question for affect:

$Affect_{SA}$ Overall in the last 30 days, how much of a problem have you had with feeling sad, low, or depressed? None, mild, moderate, severe, or extreme?

Two questions on different aspects of affect:

$Affect_1$ Please indicate which of the following best describes your own situation:

1. I love life and am happy all the time. I never worry or get upset about anything and deal with things as they come.
2. I am usually happy and positive, even when things go wrong in my life. I never get depressed, although I sometimes worry about my health or personal relations.
3. I am happy most of the time, but often worry about things in general, such as health, work, family, or relationships.
4. I am generally happy, but about once a month I feel sad and try to avoid meeting other people.
5. I have mood swings. When I get depressed, everything I do is an effort for me.
6. I feel depressed most of the time. I cry frequently and feel hopeless about the future. I feel that I have become a burden on others.

$Affect_2$ Please indicate which of the following best describes your own situation:

1. I feel nervous and anxious. I worry and think negatively about the future, but I feel better in the company of people or when doing something that really interests me. When I am alone I tend to feel useless and empty.
2. I worry all the time. I get depressed about once a week or so, thinking about what could go wrong.
3. I generally don't worry, but about once every three months I worry about what could go wrong and I get depressed.

4. I generally don't worry, but sometimes (not more than once a year or so) I worry about what could go wrong and I get depressed.

5. I never worry about a thing.

In wave 2, the replica vignette is asked. For example, for a respondent with wave 1 answers $Affect_1 = 3$ and $Affect_1 = 4$ the replica vignette is as follows:

$Affect_{RV}$ Ruth is happy most of the time, but often worries about things in general, such as health, work, family, or relationships. She generally doesn't worry, but sometimes (not more than once a year or so) she worries about what could go wrong and she gets depressed.
Overall in the last 30 days, how much of a problem has she had with feeling sad, low, or depressed?

Chapter 4

Testing Parametric Models Using Anchoring Vignettes against Nonparamteric Alternatives

4.1 Introduction

In many studies in the social sciences using survey data on individuals or households, the data available to describe the respondents' behaviors, attitudes, and well-being are inherently qualitative and subjective. In such data, people are typically asked to provide ratings on some subjective ordinal scale. An example that is commonly used in general socioeconomic surveys is the self-assessed health question on a five-point scale (from excellent to poor, for example). Among many other examples are evaluations of responsiveness of the health care system (Rice *et al.* (2010)) or political efficacy (King *et al.* (2004)).

Answers to questions with a subjective scale may depend on both the objective reality and the way in which respondents interpret the subjective answers, that is, the respondents' reporting behavior. The latter is often referred to as differential item functioning (DIF; see Holland and Wainer (1993)). Usually, we are interested in comparing the objective reality across socioeconomic groups or countries, and we therefore need to correct for differences in reporting behavior. To identify differences in reporting behavior, King *et al.* (2004) have proposed to use the tool of anchoring vignettes. These are short descriptions of hypothetical persons or situations. Respondents are asked to evaluate one or more vignettes on the same subjective scale they used to rate their own situation. Because the objective situation of the person described in the vignette(s) is the same for all respondents, systematic differences in vignette evaluations across respondents

must reflect differences in reporting behavior.

King *et al.* (2004) propose a parametric model as well as a nonparametric approach to use the anchoring vignettes in order to compare the distribution of the underlying reality of the phenomenon of interest corrected for DIF. The parametric model is referred to as compound hierarchical ordered probit model (CHOPIT). Research using anchoring vignettes has grown rapidly in recent years, and virtually all applications use the CHOPIT model or parametric extensions of this model. This includes studies on comparing several aspects of health (Bago d'Uva *et al.* (2008a), Bago d'Uva *et al.* (2008b), Lardjane and Dourgnon (2007)), health care responsiveness (Rice *et al.* (2010)), work disability (Kapteyn *et al.* (2007)), job satisfaction (Kristensen and Johansson (2008)), and life satisfaction (Kapteyn *et al.* (2010)). The CHOPIT model consists of ordered probit equations (for vignettes and own situation) with thresholds (common for all equations) that depend on the respondents' socioeconomic characteristics to account for DIF; different reporting behavior translates into different thresholds.

The nonparametric approach has been used much less often; the only applications we know of are King *et al.* (2004), King and Wand (2007) and Hopkins and King (2010). The non-parametric approach is essentially based upon comparing across different socioeconomic groups (or countries) the distributions of the rank of the respondent's evaluation of his/her own situation amongst the same respondent's vignette evaluations. It is not based upon any model and does not use other covariates than those used to distinguish the socioeconomic groups. For each socioeconomic group, the method partitions the self-reports and vignette evaluations data into (non-overlapping) cells characterized by different rankings and interprets the differences between the socioeconomic groups in the distributions over the cells.

The non-parametric approach relies on two assumptions: reporting behavior of the respondents is the same in the self-assessment questions and the vignettes ("response consistency") and the level of the variable represented in a vignette is perceived by all respondents in the same way ("vignette equivalence"). In addition to these two assumptions, the parametric model requires much more. For example, it assumes that the latent variable driving the own situation is a linear function of observed characteristics and an unobserved component; moreover, it assumes a specific functional form of the thresholds and joint normality of the unobservable (error) terms. In this chapter, we compare the parametric model with the nonparametric approach. We use the chi-squared diagnostic tests introduced in Andrews (1988), to test the specification of the parametric model against nonparametric alternatives that lead to different rankings of the

self-reports and vignette evaluations. While many alternative specification tests for the parametric model can be considered, our tests are motivated by the fact that they have power in a direction that matters: they reject the parametric model if misspecification is such that it would lead to biased conclusions concerning comparisons across socioeconomic groups - the sort of comparisons that anchoring vignettes are designed for. This makes our tests particularly useful from an applied point of view.

We run the tests for six health domains (breathing, cognition, depression, mobility, sleeping and bodily pains) on data on the population of ages 50 and older in eight European countries, from the 2004 wave of the Survey of Health, Ageing and Retirement in Europe (SHARE). For each of the six domains, self-assessments and evaluations of three vignettes describing different health level of a hypothetical person are available for each respondent.

The remainder of this chapter is organized as follows. Section 4.2 explains the parametric and nonparametric approaches for using anchoring vignettes to correct for DIF. In Section 4.3, we introduce our diagnostic tests. Section 4.4 describes the data and Section 4.5 presents the results of the tests. Section 4.6 concludes.

4.2 Parametric models and nonparametric approach

4.2.1 Parametric models

We first describe the parametric model of anchoring vignettes, also known as the CHOPIT model (King *et al.* (2004)). This is the model typically used in studies exploiting vignettes to adjust the self-assessments for heterogeneity in reporting behavior. In our exposition we assume that self-assessments and vignettes concern some aspect of health, which corresponds to our empirical application, but the method applies to any context where anchoring vignettes are used. The CHOPIT model consists of a self-assessment equation explaining the respondents' evaluation of their own health, and a vignette equation explaining the respondents' evaluation(s) of the health of (one or more) hypothetical vignette persons. The self-assessment of respondent i is modeled as follows:

$$Y_{si}^{\star} = X_i^{'}\beta_s + \epsilon_{si} \quad i = 1, \ldots, I \tag{4.1}$$

$$Y_{si} = j \Leftrightarrow \tau_i^{j-1} \leq Y_{si}^{\star} < \tau_i^{j} \quad j = 1, \ldots, J \tag{4.2}$$

$$\tau_i^1 = X_i'\gamma^1 + u_i \tag{4.3}$$
$$\tau_i^j = \tau_i^{j-1} + \exp\left(X_i'\gamma^j\right) \quad j = 2, \ldots, J-1 \tag{4.4}$$
$$\tau_i^0 = -\infty; \quad \tau_i^J = +\infty, \tag{4.5}$$

Here $Y_{si}^\star$ is the latent health of respondent i, modeled as the sum of a linear combination of explanatory variables X_i and an unobserved component ϵ_{si}, which may reflect unobserved heterogeneity, reporting error, or both. The observed value of self-assessed health Y_{si} is equal to $j(\in \{1, \ldots, J\})$ if the latent health is between thresholds τ_i^{j-1} and τ_i^j. In our application respondents rate their health on a 5-point scale (see Appendix 4.B) so that $J = 5$. The thresholds are allowed to vary across different groups of respondents characterized by observed variables X_i. Moreover, thresholds can vary with unobserved characteristics u_i. This unobserved heterogeneity term was not included in the original CHOPIT model of King *et al.* (2004) but was introduced in later extensions of the model, starting with Kapteyn *et al.* (2007).

The evaluations of vignette v_k of respondent i are modeled as follows:

$$Y_{v_k i}^\star = \theta_{v_k} + \epsilon_{v_k i} \quad i = 1, \ldots, I \quad k = 1, \ldots, K \tag{4.6}$$

$$Y_{v_k i} = j \quad \Leftrightarrow \quad \tau_i^{j-1} \leq Y_{v_k i}^\star < \tau_i^j \quad j = 1, \ldots, J \tag{4.7}$$

where $Y_{v_k i}^\star$ is the latent health of the hypothetical person described in vignette v_k, modeled as a sum of a vignette specific constant θ_{v_k} and an unobserved component $\epsilon_{v_k i}$. θ_{v_k} does not vary over respondents since we assume that each vignette is interpreted in the same way by all respondents ("vignette equivalence"). In our application we use $K = 3$ vignettes. $Y_{v_k i}$ is the reported evaluation of the health of the person described in vignette v_k by respondent i on the same J-point scale used for the self-assessments. $Y_{v_k i}$ is equal to $j = 1, \ldots, J$ if the latent health $Y_{v_k i}^\star$ is between thresholds τ_i^{j-1} and τ_i^j. The assumption of response consistency implies that the thresholds are the same as for the self-assessments.

The error terms $\epsilon_{si}, \epsilon_{v_k i}, k = 1, \ldots, K$ and the random effect u_i are assumed to be independent of each other and of the covariates X_i, with normal distributions that have mean zero and variances σ_s^2, σ_v^2 and σ_u^2, respectively:

$$\epsilon_{si} \sim \mathrm{N}(0, \sigma_s^2) \tag{4.8}$$
$$u_i \sim \mathrm{N}(0, \sigma_u^2) \tag{4.9}$$
$$\epsilon_{v_k i} \sim \mathrm{N}(0, \sigma_v^2), \quad k = 1, \ldots, K \tag{4.10}$$

By means of normalization, we will impose $\beta_{s,1} = 0$ and $\sigma_s = 1$.

4.2.2 Nonparametric Approach

The nonparametric approach is explained by King *et al.* (2004) and King and Wand (2007). Essentially, it implies re-scaling the self-assessments on the scale fixed by the vignette evaluations by socioeconomic group or country. A stylized numerical example with only one vignette is as follows.

Distribution (in %) of Self-assessments and Vignette Evaluations in Countries A and B

Country A

	vignette					
	1	2	3	4	5	all
self-assessment	(none)	(mild)	(moderate)	(severe)	(extr.)	
1 (no problem)	4	4	4	4	4	20
2 (mild problem)	4	4	4	4	4	20
3 (moderate problem)	4	4	4	4	4	20
4 (serious problem)	4	4	4	4	4	20
5 (extreme problem)	4	4	4	4	4	20
all	20	20	20	20	20	100

Country B

	vignette					
	1	2	3	4	5	all
self-assessment	(none)	(mild)	(moderate)	(severe)	(extr.)	
1 (no problem)	16	4	4	4	0	28
2 (mild problem)	8	4	4	4	0	20
3 (moderate problem)	8	4	4	4	0	20
4 (serious problem)	8	4	4	4	0	20
5 (extreme problem)	0	4	4	4	0	12
all	40	20	20	20	0	100

The cross-tabulations above give the joint distributions of self-assessments and vignette evaluations of health problems in a given domain in two hypothetical countries, A and B. Looking at the (marginal) distribution of the self-assessments only (the final column) would lead to the conclusion that respondents in country B face fewer problems in this health domain than respondents in country B –

under the assumption that they use the same response scales. The difference in the marginal distribution of the vignette evaluations (the final row), however, shows that this assumption is incorrect: respondents in country A evaluate a given health problem as more problematic, on average, and this may be an alternative explanation for the cross-country difference in the self-assessments.

The nonparametric approach simply entails comparing the relative distributions (how do the self-assessments rank compared to the vignette evaluations?) in two countries. The relative rankings are as follows:

```
Relative Ranking (RR) of Self-assessments
and Vignette Evaluations (in %) by Country

                                  Country A  Country B
RR=1: Self-ass < Vignette eval    40         24
RR=2: Self-ass = Vignette eval    20         28
RR=3: Self-ass > Vignette eval    40         48
```

The distribution of RR in country A is stochastically dominated by that in country B. The relative ranking therefore shows that, once differences in response styles are accounted for, it is clear that the health problems in country B are more serious than in country A. This is the reverse of the conclusion based upon the self-assessments only.

The example above concerns the case of only one vignette. King *et al.* (2004) consider the case with more than one vignette, say K vignettes, assuming that the evaluations of the vignettes are ranked in the same way by each respondent and assuming that each respondent evaluates different vignettes differently. In that case the self-assessment can fit in any of $2K+1$ positions in the ranking of the vignette evaluations (better or worse than any of the vignettes (2 possibilities), in between two vignettes ($K-1$ possibilities), or equal to one of the K vignettes (K possibilities). The nonparametric approach then boils down to comparing the distributions over the $2K+1$ positions in the two countries (or groups). See King *et al.* (2004, pp. 195-196) for an empirical illustration.

King and Wand (2007) discuss the more realistic case with *ties*, that is, situations where a respondent assigns the same evaluation to several vignettes, or situations where a respondent rates the vignettes in a way that does not respect the ranking of the vignette evaluations used by the majority (which is often the only natural ranking, given the wordings of the vignette descriptions). In our empirical examples with $K = 3$ vignettes for each domain, a complete listing of all possible situations, that is, all possible rankings of vignettes and self-assessments, is given in Table 4.1. The natural ordering of the vignette ratings

is assumed to be $Y_{v_1} < Y_{v_2} < Y_{v_3}$. The seven situations in the left upper block respect this ordering; the remainder of the table looks at ties. Some of these are non-problematic since all that matters is the position of the self-assessment Y_s. For example, the situation $Y_s < Y_{v_2} < Y_{v_1} < Y_{v_3}$ puts Y_s in the same position as $Y_s < Y_{v_1} < Y_{v_2} < Y_{v_3}$. For the nonparametric comparison, the two will be merged. This is indicated in the table by assigning the label 1 to both of them (column C). Similarly, $Y_{v_2} < Y_{v_1} < Y_s < Y_{v_3}$ puts Y_s in the same place as $Y_{v_1} < Y_{v_2} < Y_s < Y_{v_3}$ (and both get label 5). But in other situations, the position of Y_s is more ambiguous. For example, if $Y_{v_1} < Y_s = Y_{v_3} < Y_{v_2}$, we can plausibly conclude that we are not in situations 1 or 2 (since $Y_{v_1} < Y_s$) or in situation 7 (since $Y_s \leq \min(Y_{v_2}, Y_{v_3})$), but the "wrong" ordering of Y_{v_2} and Y_{v_3} makes it unclear which of the situations 3, 4, 5 or 6 is more plausible. This situation is therefore coded as 3-6.

This nonparametric method boils down to categorizing observations into specific cells. The 19 labels in Table 4.1 define a partition of the set Y of possible realizations of the observed dependent variables (self-assessments and vignette evaluations) into 19 cells. If the population consists of the countries A and B only, then the two countries form a partition of the set of all possible values of the regressors X into two cells. The nonparametric analysis is then based upon the partition of $Y \times X$ into 2*19=38 cells. In practice, it will often be necessary to reduce the number of cells because some of them, and particularly the ones which do not respect the natural ordering of the vignette evaluations, will typically contain very few observations. Ideally, the number of observations in the twelve cells other than those labeled $1, 2, ..., 7$ should be so small that these cells can be discarded. This is helpful for the comparison because the position of Y_s in the remaining cells respects a clear ordering, and we can say that the distribution of the health domain considered is better in country A than in country B if the distribution over the cells $\{1, ..., 7\}$ in country A stochastically dominates that in country B.

To interpret the nonparametric results, we need the assumptions of response consistency and vignette equivalence, the two assumptions underlying the use of anchoring vignettes to correct for response scale differences. These are the identifying assumptions in this framework and we will consider them as maintained hypotheses; tests of response consistency (using additional information) and vignette equivalence are discussed elsewhere and are not the topic of this chapter.[1] No additional assumptions are needed for the nonparametric approach;

[1]For tests on response consistency, using additional information in the form of a measure on an objective scale, see Van Soest *et al.* (2011), or Datta Gupta *et al.* (2009); for an analysis of

we do not even need additional regressors. The parametric model also assumes response consistency and vignette equivalence, but also makes a number of additional assumptions. Thus the nonparametric approach is less restrictive than the parametric approach. (Note that the nonparametric approach does not really formulate a model, so it seems inappropriate to say that the parametric model is nested in the nonparametric model.)

4.3 Misspecification Tests for the Parametric Model

There are many ways to test the specification of the parametric model. For example, Lagrange multiplier tests can be performed against specific parametric extensions, such as models with heteroskedastic errors or errors with a nonnormal distribution, in the spirit of, for example, Chesher and Irish (1987). Such tests will be powerful in the directions of the specific alternatives but may be less powerful in other directions. Since one of the main goals of the parametric model is to compare health or well-being across countries or socio-economic groups after purging self-assessments for response scale differences, it seems natural to look for tests with power in the directions of misspecification that lead to different conclusions concerning such comparisons.

A general category of misspecification tests are the goodness of fit tests of Andrews (1988). These tests partition the product space of outcomes and regressors $Y \times X$ into cells, and then compare the distribution over these cells in the data with the distribution generated by the estimated parametric model (taking the sample of values of the regressors as given) and generating error terms, unobserved heterogeneity terms, and values of the dependent variables. Under the null hypothesis that the parametric model is correctly specified, the two distributions should be similar. Andrews shows that this idea can be formalized by constructing a quadratic form which has a chi-squared distribution under the null. If the parametric model is estimated by maximum likelihood (as in our case), the test statistics can easily be computed by performing an auxiliary OLS regression. The dependent variable in this regression is an n-dimensional vector $(1, 1, \ldots, 1)'$, where n is the number of observations; the regressors are the deviations between realizations (1 if the observation is in the given cell, 0 otherwise) and predicted probabilities for each cell, as well as the scores of the

vignette equivalence, exploiting the overidentifying information if respondents get more than one vignette, see Corrado and Weeks (2010) or Peracchi and Rossetti (2010).

likelihood function, see Andrews (1988, p. 154) for details. Under the null of no misspecification of the parametric model, n times the R^2 of this regression is asymptotically chi-squared distributed with degrees of freedom equal to the number of cells minus one.

Each partition of $Y \times X$ gives a different test, with power in different directions (directions that will lead to a different distribution over the cells in the chosen partition). We will use the cells that are also the basis for the nonparametric approach described above. The idea behind this is that if we would use the parametric model to do the cross country comparison in the same way as using the nonparametric approach (but generating the dependent variables with the estimated model), the test will have power against alternatives that make the comparison on the basis of the parametric model different from that using the nonparametric approach. Since the test is asymptotic, the number of observations must be large enough to guarantee that the size of the test is approximately equal to the asymptotic size of 5%. In practice, this means that we will have to merge cells to guarantee that the number of observations in each cell is reasonably large. We have performed some simulations to compute the actual size of the test for various partitions and our choice of cells is based upon these simulation outcomes.

4.4 Data

We use data from Survey of Health, Ageing and Retirement in Europe (SHARE) collected in 2004. SHARE is a broad socioeconomic survey among the population of ages 50 and older and their spouses in eleven European countries, modeled after the US Health and Retirement Study (HRS) and the English Longitudinal Study on Ageing (ELSA).[2] All respondents got a personal interview and, in addition, were asked to complete a short drop off paper and pencil questionnaire. In eight countries, Belgium, France, Germany, Greece, Italy, Netherlands, Spain, and Sweden, random subsamples were given a drop off questionnaire with self-assessments and vignettes on several aspects of health (not in the context of work) and on work disability. Here we focus on the health questions, which were also used by, for example, Lardjane and Dourgnon (2007) and Bago d'Uva *et al.* (2008a). Self-assessments and three vignettes were collected for six health domains – breathing, concentration, depression, mobility, sleeping and bodily pains. The wordings of the self-assessment questions and the vignettes can be

[2]See Börsch-Supan and Jürges (2005) for details on the design and set up of the 2004 wave of SHARE.

found in Appendix 4.B. All questions could be answered on the same five-point scale: none, mild, moderate, severe, or extreme. The three vignettes in each domain are ordered, with one vignette describing a mild health problem, a second vignette describing a worse problem, and the third vignette describing the most severe health problem. This order is used as the natural order in the nonparametric approach.[3] For each domain, we denote the vignette describing the mildest of the three problems by v_1, the intermediate one by v_2, and the vignette describing the worst problem by v_3.

The total size of the vignette subsample is 4544 respondents. Due to missing observations, we work with around 4360 respondents for each domain. The precise sample sizes for and descriptive statistics for self-assessment questions and vignettes in each of the six domains are presented in Table 4.2. Most respondents rate their own health problems as none or mild for all domains. Severe or extreme health problems were reported by about 6.5 percent of the respondents, on average across the domains (3.58 percent for breathing to 8.82 percent for pain and 9.23 percent for sleep). The majority of the respondents reports to have no problem with mobility or breathing, but for the other domains, the "none" answers are a minority. In particular, pain problems are quite prevalent, with more than two thirds of all respondents reporting that they have a mild problem with bodily pains or worse.

Vignette evaluations reflect the level of the health problems of the hypothetical persons described in vignettes. As expected, the person in the third vignette (v_3) in each domain was, on average, evaluated as least healthy, followed by the person in v_2. The person described in the first vignette v_1 was, on average for each of the six domains, the person with the smallest health problem.

SHARE is quite a rich survey, with many background variables collected for all respondents. For the parametric model in Section 4.2, we use the following background variables: country, sex, age, education, frequency of physical activity, marital status[4] and suffering from a chronic disease. Descriptive statistics for the background variables for the complete vignette sample of 4544 observations are given in Table 4.3. The average age of the respondents is 63 years, 45 percent of them obtained an intermediate level of education and 20 percent obtained higher education. Most of the respondents are females (56 percent), are married and do

[3]About 50 percent of the respondents got all their vignette questions in the order from mild to severe; the other 50 percent got them in the reverse order. Vignette questions always came after the corresponding self-assessment.

[4]Based on marital status we distinguish two categories: a) not alone - married and living together with spouse, and registered partnership, b) alone - married but living separated from spouse, never married, divorced, and widowed.

not live alone or live in registered partnership (74 percent), do not suffer from a chronic illness (54 percent), and at least sometimes do physical exercise (60 percent).

4.5 Results

For each health domain, we estimated the parametric model described in Section 4.2. As an example, the estimated parameters of the parametric model for concentration are presented in Table 4.4.[5] The first column shows how, according to the parametric model, concentration and memory skills are associated with individual characteristics (including country dummies), keeping response scales constant. Most results here are plausible and confirm findings in the literature. For example, concentration and memory increase with education and fall with age or with chronic illness. There are substantial differences across countries. In particular, concentration and memory in Sweden are much better than in other countries.

The other columns present the estimates of the parameters determining the thresholds. Many variables are significant, implying that accounting for DIF is needed and not accounting for DIF would lead to biased estimates of the parameters of main interest in the first column. Particularly the estimates of γ^1 are important since $X_i'\gamma^1$ affects all thresholds in the same way (see equation (4.3)). They imply that, for example, Swedish respondents use lower thresholds than others, so that they tend to evaluate a given concentration and memory problem as more serious than respondents in other countries. Correcting their self-assessments for this makes them even better off than their self-assessment data would suggest. These findings are not new to this chapter. Using the same data, similar models have been estimated and compared to models not incorporating DIF by, for example, Bago d'Uva *et al.* (2008b). The three vignette dummies in the bottom panel of the table have the expected ranking, corresponding to the fact that the first vignette describes the mildest problem, etc. Finally, note that the standard deviation of the unobserved heterogeneity term u_i is quite precisely estimated, with a 95 percent confidence interval [0.382, 0.430]. This suggests that extending the standard CHOPIT model with this unobserved heterogeneity term is useful, even though the role of this unobserved heterogeneity term is smaller than the roles of the noise terms ϵ_{si} and $\epsilon_{v_k i}, k = 1, 2, 3$.

In this chapter, we do not focus on the parameter estimates but on the test comparing the parametric model versus the nonparametric approach. For this

[5]Parameter estimates for the other models are available upon request.

purpose, the estimated parameters of the parametric model were used to generate the dependent variables from the regressors observed in the data and use them to estimate the cell probabilities needed for the test. Specifically, for each observation in the data set, we first predict values $Y_s^\star, Y_{v_1}^\star, Y_{v_2}^\star, Y_{v_3}^\star, \tau_s^j, \tau_v^j, j = 1, 2, 3, 4$ using the given values of the regressors and the estimated coefficients. This gives one predicted value for the Y_s and the thresholds for each respondent. Then we generate values of error terms $\epsilon_{v_1}, \epsilon_{v_2}, \epsilon_{v_3}$ and the heterogeneity term u. For each respondent, we randomly chose 30 values of u from normal distribution with zero mean and the estimated σ_u^2 variance. Each randomly chosen u was then paired with 300 randomly chosen triples $\epsilon_{v_1}, \epsilon_{v_2}, \epsilon_{v_3}$. The values of ϵ_{v_k} were chosen from a normal distribution with zero mean and estimated variance σ_v^2. The randomly generated values were added to the predicted values $Y_s^\star, Y_{v_1}^\star, Y_{v_2}^\star, Y_{v_3}^\star, \tau_s^j, \tau_v^j, j = 1, 2, 3, 4$. Finally, we simulated 9000 vectors of $(Y_s, Y_{v_1}, Y_{v_2}, Y_{v_3})'$ for each respondent.[6] This simulation technique produces probabilities for each of the 19 cells of Y for each observation, that is, for given values of X. Combining these probabilities for all respondents or by groups of respondents with specific values of X_i gives the probability distribution over the partition of $Y \times X$ corresponding to the nonparametric approach. The test is based upon comparing this distribution generated by the parametric model with the distribution over the same partition in the raw data.

The number of observed and predicted observations in some cells is very low (lower than 5 %). These cells are mainly the cells corresponding to tied categories in which the natural order of the evaluations of the three vignettes is not respected (which supports the quality of data). An example of the complete observed and predicted distributions for concentration over the 19 cells in the partition of Y (not partitioning X) is given in Table 4.5. The number of observations in all cells with ties except one (the cell 2-4) is lower than 5 %. In some of the cells where the natural order of the vignette evaluations is respected, the number of observations are also rather small. In particular, the cells 3, 5, 6 and 7 in the table also have less than 5% of the observations. This is due to the fact that in general, the vignettes describe people who are in worse health than the average respondent, so that cells in which the position of the self-assessment compared to the vignettes is relatively poor have low frequencies. Similar patterns were found for other domains as well.[7]

Our tests rely on asymptotics keeping the number of cells fixed, with the

[6] Using more replications of the error terms and the heterogeneity term led to virtually identical results.

[7] Results are available upon request from the authors.

number of observations going to infinity (see Andrews (1988)). As a consequence, the finite sample properties of the test may be poor if the number of observations in certain cells is small. To prevent this, we have merged cells so that small cell sizes are avoided. Table 4.6 shows how the 19 cells are merged into 4 cells in our case. First, ties 1-4, 1-5, 1-6, 1-7, 2-4, 2-5, 2-6 are all merged with cells 3 and 4; second, ties 2-7, 3-6, 3-7, 4-6, 4-7 are merged with the low-frequency cells 5,6 and 7. This means that we map each tie into one number corresponding to (rounded) midpoint of the interval (which is a way how to map an interval into one number when there is no other information about which value in the interval is more probable). The two new cells are labeled 3 and 4, respectively; from now on, we will use this new labeling, with cells 1 and 2 as in Table 1, but with cells 3 and 4 referring to the new cells that stem from the merging procedure.

The first version of our test is based upon a partition of Y only into the four cells described above, not using X. The p-values for all domains are presented in the first row of Table 4.7. We do not reject the null for concentration (p-value=0.19) and pain (p-value=0.32). For other domains we reject the null on a 5 % significance level – depression (p-value=0.02), mobility, breathing and sleeping (p-value=0.00). These results suggest that the parametric model may be correctly specified for concentration and pain while for other domains it is not. To compare the observed and predicted probabilities on which the tests are based, both distributions are presented in Table 4.8. For pain, concentration and depression the maximum difference between observed and predicted probabilities for each of the four cells is around 1 percentage point. For mobility and sleep the maximum difference increases to 1.5 percentage points and for breathing to 2.4 percentage points. Comparing the two distributions suggests that the differences are not that big even for the domains where the null is firmly rejected.

The test using a partition of Y only essentially tests whether the parametric model is able to reproduce the ranking of vignette evaluations and self-assessments, a feature of the marginal distribution of the dependent variables, not involving any regressors. This does not yet correspond to the nonparametric approach – the nonparametric approach compares two such rankings, distinguished on the basis of X (for example, two countries or groups of countries; men and women; high and low educated respondents; etc.). But if the model is already not able to reproduce the marginal ranking, there is little hope that it adequately reproduces the ranking within subsamples characterized by specific values of the regressors; this is why we also present the tests in the first row.

The remaining rows in Table 4.7 present the p-values of tests using various partitions $Y \times X$, where X is partitioned into different socio-economic groups

or countries (using regressors that are also included in the parametric model). Each of these tests has power in a specific direction corresponding to the non-parametric approach for comparing specific socioeconomic groups. To guarantee that sample sizes are large enough, we only consider partitions of X into two or three groups, leading to a partition of $Y \times X$ into $4 \times 2 = 8$ or $4 \times 3 = 12$ cells. First we consider a partition $Y \times$ country, where countries are divided into two groups - southern Europe (Greece, Spain and Italy) and the remaining countries (Belgium, France, Germany, Netherlands and Sweden). This north-south division of countries corresponds to the systematic differences found in many SHARE studies; see, for example, many chapters in Börsch-Supan *et al.* (2005). In addition, we perform the test for partitions $Y \times$ sex, $Y \times$ age, where age is categorized into three groups - younger than 56, 56-65, and older than 65; $Y \times$ education, where education is categorized into three groups - low, middle and high; $Y \times$ sport, where doing physical exercise or sports is categorized into three groups - never, sometimes and often; $Y \times$ alone, where we distinguish between marital status of respondents - not alone (married and living together with spouse and registered partnership) and alone (married, but living separately from spouse, never married, divorced, and widowed); and finally $Y \times$ illness, where a distinction is made between respondents having and not having a chronic illness.

The p-values of all these tests are presented in the second panel of Table 4.7. The tests were again performed for all domains. As expected (based on the results for the partition of Y only), the null hypothesis is rejected at the 5% significance level for all considered partitions of $Y \times X$ for breathing, sleep and mobility. For the other domains, pain, concentration and depression, the results are mixed. For the partitions $Y \times$ sport and $Y \times$ illness, the p-values are always higher than 0.05 (and in some cases even higher than 0.5) so that the null is not rejected. Similarly, p-values exceeding 5% are also found for partitions $Y \times$ sex and $Y \times$ age for pain and concentration. On the other hand, for all domains, the null hypothesis is rejected for partitions $Y \times$ country and $Y \times$ education. For pain, concentration and depression, these results can be interpreted as supporting the use of the parametric model for comparison of DIF adjusted self-assessments between groups with different levels of physical activity or sports, and between groups suffering and not suffering from chronic illness. The same conclusion can be drawn for comparing men and women or comparing various age groups when considering the domains pain or concentration and memory. On the other hand however, the test results do not support the use of the parametric model for comparison across (southern versus northern) countries or different education

groups – the types of comparisons that have been the focus in existing studies like Lardjane and Dourgnon (2007) and Bago d'Uva *et al.* (2008b).

Tables 4.9 and 4.10, present the observed and predicted distributions over the cells constructed using partitions $Y \times$ country and $Y \times$ age, respectively. For the partition $Y \times$ country, the maximum difference between the predicted and observed probability for a cell across all domains is about two percentage points. For $Y \times$ age, the maximum discrepancy across all domains is about 1.5 percentage points. For pain and concentration the maximum difference is about 1 percentage point and many observed and predicted probabilities are very similar (with differences below 0.1 percentage points).

We therefore encounter several cases where the null is firmly rejected whereas the predicted and observed probabilities seem to be very close. The Andrews' chi-square test statistic not simply compares predicted and observed probabilities but accounts for the fact that parameters are estimated using the same sample of observations, by incorporating the likelihood scores for each observation in the auxiliary regression used to obtain the test statistic (cf. Section 3). One of the possible explanations for rejecting the null in spite of the similarity of predicted and actual frequencies might be that the fact that parameters are estimated plays an important role, and this would imply that the likelihood scores lead to a substantial increase in the values of the test statistic.

To see to which extent this matters, we also computed the test statistic without the likelihood scores.[8] Table 4.11 presents the p-values of all the tests presented in Table 4.7 discarding the likelihood scores. These p-values are, for example, much higher than the correct p-values in Table 4.7 for the sleep domain. Apparently, for this domain, the fact that parameters are estimated makes it likely that under the null, predicted and observed frequencies are very similar. Adjusting for this implies that the null hypothesis is already rejected for quite modest differences, much smaller than the differences that were needed to reject the null if parameters were given instead of estimated (or if parameters would be estimated using an independent sample from the same population). For breathing, on the other hand, the p-values remain virtually zero for all partitions. Here the discrepancies between predicted and actual frequencies are so large that the null would also be firmly rejected if parameter values were given instead of estimated on the same sample. This is in line with the fact that the observed and predicted probabilities are more similar for sleep than for breathing (see, for example, the comparison of observed and predicted probabilities for the partition $Y \times$ age in Table 4.10).

[8]This means that Matrix B in Andrews' test statistic (see Andrews (1988, p. 154)) is replaced by zeros.

Results of the tests presented in Table 4.7 suggest that the parametric model specified in Section 4.2 does not work properly for all presented partitions. This holds mainly for breathing and sleep domains. To study how the results are sensitive to wrongly ordered vignettes or functional form and the stochastic specification of the parametric model we performed the following additional checks. To analyze the impact of wrongly ordered vignettes, we re-estimated the parametric model using the subsample of $Y_{v_1} \leq Y_{v_2} \leq Y_{v_3}$ respondents only, and re-computed p-values for all partitions presented in Table 4.7. The p-values for this subsample generate similar rejection regions as the p-values for the whole sample (see Table 4.12). It therefore seems unlikely that wrongly ordered vignettes are the source of the misspecification of the parametric model. Second, we re-estimated the parametric model with a linear form of the variable age instead of the age dummies used in the original model (see Table 4.4). This did not lead to substantial changes of p-values (see Table 4.13). Third, concerning the stochastic specification, we also re-estimated the model without the random effects in u_i included in thresholds (see equation (4.3)). This more restrictive version of our model corresponds exactly to the CHOPIT model originally proposed by King *et al.* (2004). Dropping the u_i leads to a substantial reduction of the p-values: the p-values of all the tests in Table 4.7 are reduced to 0.000. This shows that the parametric model without the random effects in thresholds is misspecified for all domains. Adding the random effect is part of the solution of the misspecification of the CHOPIT model, but is only enough to let the model pass the test for a subset of partitions and domains.

4.6 Conclusion

Comparing self-reported survey measures of well-being, health, or other aspects of perceived quality of life or society often suffers from the fact that different groups use different reporting scales (DIF). Anchoring vignettes are an increasingly popular tool to identify and correct for these differences in response scales. In the literature, there are two ways to use anchoring vignettes: parametric models (the CHOPIT model and its extensions) or a nonparametric approach based upon ranking vignette evaluations and self-assessments in subsamples characterized by values of control variables (such as country, age, gender, or education level). In this chapter, we consider specification tests of the parametric model that have power against alternatives that make using the parametric model inconsistent with the nonparametric approach.

We apply the tests using data on the 50+ population in eight European coun-

tries, with self-assessments and vignettes on six domains of health. Our results are mixed. The specification of the standard CHOPIT model is always rejected, but the CHOPIT model extended with unobserved heterogeneity in the reporting scales performs better. For some socioeconomic characteristics and some health domains, we cannot reject the hypothesis that the parametric model generates the same distributions of the rankings of vignettes and self-assessments as the distributions in the raw data used for the nonparametric approach. But in other cases, even the marginal distribution of the rankings of vignettes and self-assessments is not captured well enough by the parametric model in order not to reject the null of a correct specification.

What does this imply for studies using anchoring vignettes? Even though the parametric model is statistically rejected in most cases, the distributions generated by the parametric model do not differ that much from the distributions in the raw data, implying that the misspecification will not always lead to large biases in the conclusions. On the other hand, taking the parametric model for granted may not always be the best thing to do. The nonparametric approach has not been used quite often and seems a viable alternative in cases where 1) the focus is exclusively on comparisons across a few socioeconomic groups or (groups of) countries and 2) the comparison is not hampered by ties that make it very difficult to interpret the nonparametric results. The latter problem seems a major reason for considering parametric models: if vignettes are not ordered consistently by all respondents, the ties that arise make it hard or even impossible to draw conclusions from the nonparametric comparisons. On the other hand, this is no problem for the parametric model, in which idiosyncratic errors can explain any violation of the natural ordering of the vignette evaluations.

All in all, this suggests that future research should focus on 1) testing the existing parametric models for a variety of comparisons and different data sets, and 2) extending the parametric models in directions that help to improve the goodness of fit, without sacrificing the essential nature of the parametric models, making it possible to interpret parameter estimates, generate counterfactual distributions of hypothetical evaluations of one group with response scales of another group (see, e.g., Kapteyn *et al.* (2007)), etc. This chapter gives one example of what we have in mind here: adding unobserved heterogeneity in the thresholds to the standard CHOPIT model already helps. Whether it helps to use further extensions involving, for example, semiparametric specifications with more flexible distributions of systematic parts or error terms (or both) remains a topic of further research.

4.A Tables

Table 4.1: Ranking self-assessment Y_s and three vignette evaluations Y_{v_1}, Y_{v_2}, Y_{v_3} and the corresponding cells (C) according to the nonparametric approach

ranking	C	ranking	C
$Y_s < Y_{v_1} < Y_{v_2} < Y_{v_3}$	1	$Y_s < Y_{v_1} < Y_{v_3} < Y_{v_2}$	1
$Y_s = Y_{v_1} < Y_{v_2} < Y_{v_3}$	2	$Y_s = Y_{v_1} < Y_{v_3} < Y_{v_2}$	2
$Y_{v_1} < Y_s < Y_{v_2} < Y_{v_3}$	3	$Y_{v_1} < Y_s < Y_{v_3} < Y_{v_2}$	3
$Y_{v_1} < Y_s = Y_{v_2} < Y_{v_3}$	4	$Y_{v_1} < Y_s = Y_{v_3} < Y_{v_2}$	3-6
$Y_{v_1} < Y_{v_2} < Y_s < Y_{v_3}$	5	$Y_{v_1} < Y_{v_3} < Y_s < Y_{v_2}$	3-7
$Y_{v_1} < Y_{v_2} < Y_s = Y_{v_3}$	6	$Y_{v_1} < Y_{v_3} < Y_s = Y_{v_2}$	4-7
$Y_{v_1} < Y_{v_2} < Y_{v_3} < Y_s$	7	$Y_{v_1} < Y_{v_3} < Y_{v_2} < Y_s$	7
$Y_s < Y_{v_2} < Y_{v_1} < Y_{v_3}$	1	$Y_s < Y_{v_2} < Y_{v_3} < Y_{v_1}$	1
$Y_s = Y_{v_2} < Y_{v_1} < Y_{v_3}$	1-4	$Y_s = Y_{v_2} < Y_{v_3} < Y_{v_1}$	1-4
$Y_{v_2} < Y_s < Y_{v_1} < Y_{v_3}$	1-5	$Y_{v_2} < Y_s < Y_{v_3} < Y_{v_1}$	1-5
$Y_{v_2} < Y_s = Y_{v_1} < Y_{v_3}$	2-5	$Y_{v_2} < Y_s = Y_{v_3} < Y_{v_1}$	1-6
$Y_{v_2} < Y_{v_1} < Y_s < Y_{v_3}$	5	$Y_{v_2} < Y_{v_3} < Y_s < Y_{v_1}$	1-7
$Y_{v_2} < Y_{v_1} < Y_s = Y_{v_3}$	6	$Y_{v_2} < Y_{v_3} < Y_s = Y_{v_1}$	2-7
$Y_{v_2} < Y_{v_1} < Y_{v_3} < Y_s$	7	$Y_{v_2} < Y_{v_3} < Y_{v_1} < Y_s$	7
$Y_s < Y_{v_3} < Y_{v_1} < Y_{v_2}$	1	$Y_s < Y_{v_3} < Y_{v_2} < Y_{v_1}$	1
$Y_s = Y_{v_3} < Y_{v_1} < Y_{v_2}$	1-6	$Y_s = Y_{v_3} < Y_{v_2} < Y_{v_1}$	1-6
$Y_{v_3} < Y_s < Y_{v_1} < Y_{v_2}$	1-7	$Y_{v_3} < Y_s < Y_{v_2} < Y_{v_1}$	1-7
$Y_{v_3} < Y_s = Y_{v_1} < Y_{v_2}$	2-7	$Y_{v_3} < Y_s = Y_{v_2} < Y_{v_1}$	1-7
$Y_{v_3} < Y_{v_1} < Y_s < Y_{v_2}$	3-7	$Y_{v_3} < Y_{v_2} < Y_s < Y_{v_1}$	1-7
$Y_{v_3} < Y_{v_1} < Y_s = Y_{v_2}$	4-7	$Y_{v_3} < Y_{v_2} < Y_s = Y_{v_1}$	2-7
$Y_{v_3} < Y_{v_1} < Y_{v_2} < Y_s$	7	$Y_{v_3} < Y_{v_2} < Y_{v_1} < Y_s$	7
$Y_s < Y_{v_1} = Y_{v_2} < Y_{v_3}$	1	$Y_s < Y_{v_3} < Y_{v_1} = Y_{v_2}$	1
$Y_s = Y_{v_1} = Y_{v_2} < Y_{v_3}$	2-4	$Y_s = Y_{v_3} < Y_{v_1} = Y_{v_2}$	1-6
$Y_{v_1} = Y_{v_2} < Y_s < Y_{v_3}$	5	$Y_{v_3} < Y_s < Y_{v_1} = Y_{v_2}$	1-7
$Y_{v_1} = Y_{v_2} < Y_s = Y_{v_3}$	6	$Y_{v_3} < Y_s = Y_{v_1} = Y_{v_2}$	2-7
$Y_{v_1} = Y_{v_2} < Y_{v_3} < Y_s$	7	$Y_{v_3} < Y_{v_1} = Y_{v_2} < Y_s$	7
$Y_s < Y_{v_1} = Y_{v_3} < Y_{v_2}$	1	$Y_s < Y_{v_2} < Y_{v_1} = Y_{v_3}$	1
$Y_s = Y_{v_1} = Y_{v_3} < Y_{v_2}$	2-6	$Y_s = Y_{v_2} < Y_{v_1} = Y_{v_3}$	1-4
$Y_{v_1} = Y_{v_3} < Y_s < Y_{v_2}$	3-7	$Y_{v_2} < Y_s < Y_{v_1} = Y_{v_3}$	1-5
$Y_{v_1} = Y_{v_3} < Y_s = Y_{v_2}$	4-7	$Y_{v_2} < Y_s = Y_{v_1} = Y_{v_3}$	2-6
$Y_{v_1} = Y_{v_3} < Y_{v_2} < Y_s$	7	$Y_{v_2} < Y_{v_1} = Y_{v_3} < Y_s$	7
$Y_s < Y_{v_1} < Y_{v_2} = Y_{v_3}$	1	$Y_s < Y_{v_2} = Y_{v_3} < Y_{v_1}$	1
$Y_s = Y_{v_1} < Y_{v_2} = Y_{v_3}$	2	$Y_s = Y_{v_2} = Y_{v_3} < Y_{v_1}$	1-6
$Y_{v_1} < Y_s < Y_{v_2} = Y_{v_3}$	3	$Y_{v_2} = Y_{v_3} < Y_s < Y_{v_1}$	1-7
$Y_{v_1} < Y_s = Y_{v_2} = Y_{v_3}$	4-6	$Y_{v_2} = Y_{v_3} < Y_s = Y_{v_1}$	2-7
$Y_{v_1} < Y_{v_2} = Y_{v_3} < Y_s$	7	$Y_{v_2} = Y_{v_3} < Y_{v_1} < Y_s$	7
$Y_s < Y_{v_1} = Y_{v_2} = Y_{v_3}$	1	$Y_s = Y_{v_1} = Y_{v_2} = Y_{v_3}$	2-6
$Y_{v_1} = Y_{v_2} = Y_{v_3} < Y_s$	7		

Table 4.2: Distributions of self-assessments and vignettes evaluations

	breathing				concentration			
	s	v_1	v_2	v_3	s	v_1	v_2	v_3
none	64.61	10.77	2.29	2.45	44.09	22.20	5.25	2.03
mild	22.22	24.12	5.15	2.22	35.16	48.65	27.15	8.89
moderate	9.60	38.02	19.74	8.57	16.21	22.73	44.35	29.77
severe	3.05	24.10	52.20	44.25	4.14	6.08	20.68	47.27
extreme	0.53	3.00	20.61	42.51	0.39	0.35	2.58	12.04
	depression				mobility			
	s	v_1	v_2	v_3	s	v_1	v_2	v_3
none	49.53	6.46	2.31	2.20	58.37	9.62	2.31	1.55
mild	28.67	44.15	13.42	2.59	22.18	34.75	11.83	5.89
moderate	14.98	36.20	45.75	10.81	13.05	42.84	38.77	27.51
severe	5.29	11.56	33.55	42.55	5.23	11.97	40.39	48.80
extreme	1.53	1.63	4.97	41.86	1.17	0.82	6.69	16.24
	pain				sleep			
	s	v_1	v_2	v_3	s	v_1	v_2	v_3
none	32.27	15.60	2.31	1.12	42.67	2.65	1.92	1.87
mild	35.80	56.94	18.07	5.31	28.07	21.49	9.71	7.31
moderate	23.11	22.10	50.73	26.09	20.02	47.98	29.01	26.99
severe	7.15	4.79	25.72	48.63	7.36	24.05	42.51	41.69
extreme	1.67	0.57	3.16	18.85	1.87	3.84	16.85	22.13

Note: The size of the vignette subsample of the SHARE sample is 4544. We work with around 4360 respondents for each domain (4366 for breathing, 4343 for concentration, 4367 for depression, 4377 for mobility, 4366 for pain and 4375 for sleep).

Table 4.3: Descriptive statistics - background variables (percentage, except for age)

Belgium	12.48	male	44.43
France	19.48	education mid	44.60
Germany	11.18	education high	19.86
Greece	15.85	not alone	74.42
Italy	9.79	long-term illness	46.49
Netherlands	11.84	phys. act. sometimes	25.08
Spain	10.21	phys. act. often	34.46
Sweden	9.18	age - mean	63.06
		age - std dev	10.01

Note: The descriptive statistics are given for the vignette subsample of the SHARE sample, i.e. for 4544 repondents. These statistics are similar to the descriptive statistics for each health domain sample we work with. Definitions of variables are given in Table 2.7.

Table 4.4: Estimates of the parametric model for concentration

	β_s		γ_1		γ_2		γ_3		γ_4	
	coef	t-val	coef	t-val	coef	t-val	coef	t-val	coef	t-val
const	0	-	-0.023	-0.367	0.131	2.853	0.111	2.422	0.158	2.361
Belgium	0.039	0.548	-0.097	-2.033	0.122	3.217	-0.143	-3.523	-0.083	-1.326
Germany	0.076	0.969	0.235	4.836	-0.150	-3.590	-0.198	-4.539	-0.103	-1.529
Greece	0.141	2.111	0.468	10.965	-0.225	-5.766	-0.356	-8.808	-0.124	-2.097
Italy	0.133	1.808	0.258	5.425	-0.104	-2.487	-0.298	-6.885	-0.078	-1.181
Netherlands	-0.192	-2.496	-0.097	-1.894	0.086	2.093	-0.300	-6.577	-0.391	-6.249
Spain	-0.150	-2.117	0.015	0.295	-0.230	-4.937	-0.153	-3.951	0.251	4.430
Sweden	-0.877	-10.895	-0.475	-6.752	-0.417	-6.399	-0.309	-5.802	0.153	2.909
male	-0.071	-1.767	0.053	2.003	-0.038	-1.610	0.016	0.669	-0.049	-1.598
age 50-	0.030	0.342	-0.034	-0.589	0.069	1.392	0.033	0.615	0.072	1.057
age 56-60	-0.029	-0.458	-0.013	-0.326	0.016	0.433	0.026	0.675	-0.026	-0.552
age 61-65	0.028	0.423	-0.102	-2.484	0.107	2.830	0.078	1.933	0.073	1.413
age 66-70	0.122	1.781	-0.119	-2.700	0.063	1.612	0.079	1.974	0.091	1.712
age 71-75	0.188	2.534	-0.142	-2.781	0.100	2.323	0.173	4.186	0.029	0.489
age 76-80	0.262	3.129	-0.116	-2.037	0.047	0.854	0.120	2.323	0.109	1.688
age 80+	0.531	5.577	-0.016	-0.258	0.070	1.197	0.009	0.163	0.111	1.584
education low	0.227	4.641	0.081	2.518	-0.101	-3.547	0.012	0.435	-0.059	-1.529
education high	-0.160	-2.863	0.001	0.020	-0.035	-1.154	-0.080	-2.313	0.003	0.084
sport sometimes	-0.131	-2.532	-0.016	-0.445	0.055	1.917	-0.010	-0.337	0.029	0.704
sport often	-0.047	-0.965	0.025	0.810	0.063	2.313	-0.048	-1.710	0.037	0.986
alone	0.143	3.217	0.028	1.026	-0.039	-1.750	-0.007	-0.268	0.070	1.961
illness	0.355	8.557	-0.035	-1.284	-0.012	-0.492	-0.015	-0.596	0.007	0.217

	coef	se
θ_{v_1}	0.652	0.069
θ_{v_2}	1.448	0.074
θ_{v_3}	2.186	0.082
σ_s	1	-
σ_u	0.406	0.012
σ_v	0.712	0.016

Table 4.5: Observed and predicted percent distribution over 19 categories for concentration

	observed	predicted
1	38.15	38.93
2	23.62	22.55
3	3.85	5.05
4	6.06	5.06
5	2.10	2.28
6	2.65	2.92
7	2.58	1.76
1-4	1.87	3.67
1-5	0.16	0.25
1-6	1.24	0.80
1-7	0.14	0.03
2-4	7.87	7.42
2-5	0.48	1.09
2-6	4.65	3.18
2-7	0.74	0.30
3-6	0.48	1.10
3-7	0.14	0.15
4-6	2.74	2.79
4-7	0.48	0.67

Table 4.6: Merging 19 cells into four larger cells (See Table 4.1 for the definitions of the original 19 cells (C))

merged cell	C
1	1
2	2
3	3,4,1-4,1-5,1-6,1-7,2-4,2-5,2-6
4	5,6,7,2-7,3-6,3-7,4-6,4-7

Table 4.7: Goodness of fit - cells constructed using partitions of Y only and of $Y \times X$

	breathing	concentration	depression	mobility	pain	sleep
Y only	0	0.186	0.018	0.001	0.318	0
$Y \times X$						
country	0	0	0	0	0	0
sex	0	0.264	0	0.002	0.145	0
age	0	0.218	0.012	0.012	0.625	0
education	0	0	0	0	0	0
sport	0	0.655	0.055	0.009	0.596	0
alone	0	0.072	0.041	0.007	0.043	0
illness	0	0.558	0.086	0.003	0.186	0

Table 4.8: Observed and predicted distributions - cells constructed using partition of Y only

	breathing		concentration		depression	
cell	observed	predicted	observed	predicted	observed	predicted
1	71.12	72.23	38.15	38.93	57.75	58.66
2	13.74	14.55	23.62	22.55	16.99	16.64
3	12.11	9.71	26.32	26.55	18.56	17.37
4	3.03	3.52	11.91	11.97	6.69	7.33

	mobility		pain		sleep	
cell	observed	predicted	observed	predicted	observed	predicted
1	63.26	63.09	34.43	34.86	61.37	62.92
2	13.23	14.69	23.80	22.79	10.40	10.42
3	15.01	13.62	25.10	25.65	17.62	16.03
4	8.50	8.59	16.68	16.71	10.60	10.61

Table 4.9: Observed and predicted distributions - cells constructed using partition $Y\times$ country

	South		North	
	observed	predicted	observed	predicted
Y cell		breathing		
1	30.55	30.25	40.56	41.99
2	2.66	3.78	11.09	10.75
3	2.45	1.96	9.67	7.75
4	0.80	0.47	2.22	3.04
		concentration		
1	13.22	12.20	24.94	26.73
2	7.71	8.87	15.91	13.68
3	10.43	10.49	15.89	16.07
4	5.48	5.28	6.42	6.68
		depression		
1	20.84	20.80	36.91	37.87
2	5.08	6.04	11.91	10.60
3	7.47	6.63	11.11	10.74
4	3.16	3.08	3.53	4.25
		mobility		
1	24.15	23.60	39.11	39.49
2	4.39	5.44	8.84	9.25
3	4.96	4.72	10.05	8.92
4	3.15	2.89	5.35	5.70
		pain		
1	12.12	10.98	22.31	23.87
2	7.54	8.58	16.26	14.21
3	9.55	10.07	15.55	15.57
4	7.47	7.03	9.21	9.68
		sleep		
1	23.52	23.56	37.85	39.36
2	2.67	3.91	7.73	6.50
3	6.54	5.50	11.09	10.54
4	3.93	3.69	6.67	6.93

Table 4.10: Observed and predicted distributions - cells constructed using partition $Y\times$ age

	age 55 min		age 56-65		age 66 plus	
	observed	predicted	observed	predicted	observed	predicted
Y cell			breathing			
1	20.20	21.01	26.11	26.39	24.81	24.83
2	4.05	3.65	4.92	4.97	4.76	5.93
3	2.54	2.09	3.37	3.03	6.21	4.60
4	0.57	0.62	0.94	0.95	1.51	1.93
			concentration			
1	11.42	11.48	14.62	15.02	12.11	12.42
2	6.61	6.61	9.21	8.19	7.81	7.75
3	6.79	6.80	8.27	8.79	11.26	10.98
4	2.53	2.46	3.22	3.32	6.15	6.18
			depression			
1	15.57	16.34	21.16	21.48	21.02	20.85
2	5.01	4.45	6.05	5.85	5.93	6.34
3	5.08	4.64	5.98	5.76	7.51	6.96
4	1.69	1.93	2.18	2.27	2.82	3.13
			mobility			
1	19.74	19.77	24.22	24.11	19.31	19.20
2	3.70	3.76	4.48	5.07	5.05	5.85
3	2.67	2.64	4.48	4.07	7.86	6.92
4	1.17	1.11	2.08	1.99	5.25	5.49
			pain			
1	10.51	10.57	13.01	13.33	10.90	10.96
2	7.01	6.58	8.47	8.20	8.31	8.01
3	6.39	6.65	8.61	8.63	10.10	10.36
4	3.53	3.64	5.13	5.07	8.02	8.01
			sleep			
1	17.78	18.59	22.40	22.85	21.19	21.48
2	3.15	2.70	3.38	3.64	3.86	4.08
3	4.00	3.85	6.03	5.37	7.59	6.82
4	2.45	2.24	3.43	3.39	4.73	5.00

Table 4.11: Goodness of fit - cells constructed using partition Y only and $Y \times X$, test statistic computed without likelihood scores

	breathing	concentration	depression	mobility	pain	sleep
Y only	0	0.398	0.070	0.005	0.450	0.038
$Y \times X$						
country	0	0	0	0.002	0	0
sex	0	0.461	0.001	0.030	0.381	0.033
age	0	0.549	0.157	0.059	0.943	0.082
education	0	0.002	0.002	0	0.002	0
sport	0	0.916	0.320	0.088	0.720	0.344
alone	0	0.371	0.254	0.037	0.077	0.119
illness	0	0.804	0.234	0.026	0.269	0.036

Note: The test statistic is computed only using observed and predicted probabilities for each respondent (matrix A in Andrews (1988)). No likelihood scores are used (matrix $B = 0$).

Table 4.12: Goodness of fit - cells constructed using partition Y only and $Y \times X$, without respondents with inconsistently ordered vignettes, i.e. only $Y_{v_1} \leq Y_{v_2} \leq Y_{v_3}$ respondents are included

	breathing	concentration	depression	mobility	pain	sleep
Y only	0	0.834	0.001	0	0.352	0
$Y \times X$						
country	0	0	0	0	0	0
sex	0	0.868	0	0.001	0.022	0
age	0	0.342	0.002	0.003	0.581	0
education	0	0.008	0	0	0.001	0
sport	0	0.938	0.011	0.001	0.463	0.001
alone	0	0.403	0.002	0.003	0.077	0
illness	0	0.822	0.005	0	0.211	0

Note: These tests are done only for the sample of respondents who ordered vignettes consistently. For each domain, the parametric model is estimated only using respondents with consistently ordered vignettes and observed and predicted probabilities are compared for the same subsample.

Table 4.13: Goodness of fit - cells constructed using partition Y only and $Y \times X$; parametric model using age in linear form instead of age dummies

	breathing	concentration	depression	mobility	pain	sleep
Y only	0	0.118	0.008	0.001	0.325	0
$Y \times X$						
country	0	0	0	0	0	0
sex	0	0.191	0	0.002	0.148	0
age	0	0.054	0.002	0.001	0.847	0
education	0	0	0	0	0	0
sport	0	0.531	0.034	0.006	0.585	0
alone	0	0.039	0.021	0.007	0.050	0
illness	0	0.427	0.045	0.003	0.202	0

4.B Self-assessment questions and vignettes

Here we present self-assessment questions and three vignettes for all health domains. Vignettes are ordered based on the level of the health problems. Vignette v_1 describes less health problems and vignette v_3 describes a person with serious health problems. Self-assessment questions and vignettes were rated on the five-point scale: none, mild, moderate, severe and extreme.

Self-assessment questions

breathing

In the last 30 days, how much of a problem did you have because of shortness of breath?

concentration

Overall in the last 30 days how much difficulty did you have with concentrating or remembering things?

depression

Overall in the last 30 days, how much of a problem did you have with feeling sad, low, or depressed?

mobility

Overall in the last 30 days, how much of a problem did you have with moving around?

pain

Overall in the last 30 days, how much of bodily aches or pains did you have?

sleep

In the last 30 days, how much difficulty did you have with sleeping such as falling asleep, waking up frequently during the night or waking up too early in the morning?

Vignettes

breathing

v_1
Mark has no problems with walking slowly. He gets out of breath easily when climbing uphill for 20 meters or a flight of stairs. In the last 30 days, how much of a problem did Mark have because of shortness of breath?

v_2
Paul suffers from respiratory infections about once every year. He is short of breath 3 or 4 times a week and had to be admitted in hospital twice in the past month with a bad cough that required treatment with antibiotics. In the last 30 days, how much of a problem did Paul have because of shortness of breath?

v_3
Henri has been a heavy smoker for 30 years and wakes up with a cough every morning. He gets short of breath even while resting and does not leave the house anymore. He often needs to be put on oxygen. In the last 30 days, how much of a problem did Henri have because of shortness of breath?

concentration

v_1
Lisa can concentrate while watching TV, reading a magazine or playing a game of cards or chess. Once a week she forgets where her keys or glasses are, but finds them within five minutes. Overall in the last 30 days, how much difficulty did Lisa have with concentrating or remembering things?

v_2
Sue is keen to learn new recipes but finds that she often makes mistakes and has to reread several times before she is able to do them properly. Overall in the last 30 days, how much difficulty did Sue have with concentrating and remembering things?

v_3
Eve cannot concentrate for more than 15 minutes and has difficulty paying attention to what is being said to her. Whenever she starts a task, she never manages to finish it and often forgets what she was doing. She is able to learn the names of people she meets. Overall in the last 30 days, how much difficulty did Eve

have with concentrating or remembering things?

depression

v_1
Karen enjoys her work and social activities and is generally satisfied with her life. She gets depressed every 3 weeks for a day or two and loses interest in what she usually enjoys but is able to carry on with her day-to- day activities. Overall in the last 30 days, how much of a problem did Karen have with feeling sad, low, or depressed?

v_2
Maria feels nervous and anxious. She worries and thinks negatively about the future, but feels better in the company of people or when doing something that really interests her. When she is alone she tends to feel useless and empty. Overall in the last 30 days, how much of a problem did Maria have with feeling sad, low, or depressed?

v_3
Anna feels depressed most of the time. She weeps frequently and feels hopeless about the future. She feels that she has become a burden on others and that she would be better dead. Overall in the last 30 days, how much of a problem did Anna have with feeling sad, low, or depressed?

mobility

v_1
Rob is able to walk distances of up to 200 metres without any problems but feels tired after walking one kilometre or climbing more than one flight of stairs. He has no problems with day-to-day activities, such as carrying food from the market. Overall in the last 30 days, how much of a problem did Rob have with moving around?

v_2
Kevin does not exercise. He cannot climb stairs or do other physical activities because he is obese. He is able to carry the groceries and do some light household work. Overall in the last 30 days, how much of a problem did Kevin have with moving around?

v_3
Tom has a lot of swelling in his legs due to his health condition. He has to make an effort to walk around his home as his legs feel heavy. Overall in the last 30 days, how much of a problem did Tom have with moving around?

pain

v_1
Paul has a headache once a month that is relieved after taking a pill. During the headache he can carry on with his day-to-day affairs. Overall in the last 30 days, how much of bodily aches or pains did Paul have?

v_2
Henri has pain that radiates down his right arm and wrist during his day at work. This is slightly relieved in the evenings when he is no longer working on his computer. Overall in the last 30 days, how much of bodily aches or pains did Henri have?

v_3
Charles has pain in his knees, elbows, wrists and fingers, and the pain is present almost all the time. Although medication helps, he feels uncomfortable when moving around, holding and lifting things. Overall in the last 30 days, how much of bodily aches or pains did Charles have?

sleep

v_1
Alice falls asleep easily at night, but two nights a week she wakes up in the middle of the night and cannot go back to sleep for the rest of the night. In the last 30 days, how much difficulty did Alice have with sleeping, such as falling asleep, waking up frequently during the night or waking up too early in the morning?

v_2
Karen wakes up almost once every hour during the night. When he wakes up in the night, it takes around 15 minutes for her to go back to sleep. In the morning she does not feel well-rested. In the last 30 days, how much difficulty did Karen have with sleeping such as falling asleep, waking up frequently during the night or waking up too early in the morning?

v_3

Maria takes about two hours every night to fall asleep. She wakes up once or twice a night feeling panicked and takes more than one hour to fall asleep again. In the last 30 days, how much difficulty did Maria have with sleeping, such as falling asleep, waking up frequently during the night or waking up too early in the morning?

Chapter 5

Stated preferences analysis: retirement decisions

5.1 Introduction

The population in many developed countries is ageing and individuals are living longer, leading to a permanent change in the ratio between the numbers of economically active and inactive people. In the Netherlands, for example, expectations are that without changes in pension and retirement policies, the ratio of the number of people aged 65 and above to the economically active population would double in the next 30 years, to over 40 %, see Kakes and Broeders (2006). Pension systems will become unsustainable if they do not adjust to this demographic change, see, e.g., Capretta (2007).

One of the problematic features of many pension systems is the existence of generous early retirement schemes which stimulate labour market exit long before the normal retirement age and greatly add to the total cost of the system. See, e.g., Gruber and Wise (1998, 2004) for a summary of the evidence in many countries and Kapteyn and De Vos (1998, 2004) for the Netherlands. In the Netherlands and other countries, early retirement became common in the 1970's when the social partners sought to "free up" jobs for younger workers facing a high unemployment rate. In the 1990's the government and the social partners realized that the early retirement programs imposed a prohibitive tax on continued work and a start was made to gradually phase them out. A new system of "pre-pension" with fewer employment disincentives was introduced. Pre-pension benefits are, in contrast to the old early retirement benefits, adjusted according to the retirement age, with lower benefits for early retirees. In 2005 other steps to discourage people to retire early were taken. The Dutch government passed

legislation that will phase out the tax-favoured treatment of all employer-based early retirement programs, see Capretta (2007). As a consequence, the male labour force participation rate in the age group 55-64 has risen from about 40 % in the 70's to around 60 % in 2006. For women it rose from less than 20 % to around 40 %. Raising the effective retirement age further is often seen as the most feasible way to improve sustainability of the pension system.

Employment after the normal retirement age (usually 65) is very uncommon in many European countries. In the Netherlands, a negligible percentage of employees currently remain at work after age 65. Mandatory retirement is the default, although in principle, firms can rehire workers after age 65, for example on a temporary and part-time basis. Factors that potentially hamper late retirement are the fact that not all pension funds allow for accumulating pension entitlements after age 65 and the obligation to pay wages for two years after an employee becomes ill.

The more recent debate focuses on creating more flexibility in order to optimize the use of the capabilities of older workers, accounting for heterogeneity in preferences and productivity. See, for example, Belloni *et al.* (2006) for an overview of policies towards flexible retirement in European countries and Bovenberg and Gradus (2008) for a discussion of proposed policy changes in the Netherlands. For the supply side this means, for example, making the retirement age more flexible with rewards for workers who postpone retirement, in the form of actuarially fair pension adjustments and tax arrangements that stimulate later retirement, and creating more opportunities for gradual retirement, see Kantarci and Van Soest (2008). Working after age 65 is an explicit part of the new plans of the Dutch government. For example, eligibility for the first pillar pension (AOW) that provides the minimum subsistence level currently starts at age 65 by default, but a new arrangement makes it possible to delay receiving this in exchange for 5% higher benefits for every year of delaying. Moreover, the government has launched new plans to delay eligibility to age 67 for everyone.

In order to design successful reforms of retirement policy, it is important to design financially sustainable retirement plans that are attractive for workers. This chapter aims at analyzing workers' preferences for potential retirement plans, with emphasis on plans that allow for full-time or part-time work after age 65.

In the economic literature, there are many empirical models explaining labour supply behaviour of older workers in an inter-temporal framework. They usually use data on observed actual behaviour of the individuals, i.e. revealed preference (RP) data (see, e.g., Lumsdaine and Mitchell (1999) for an overview and Kapteyn

and De Vos (2004), Heyma (2004), Euwals *et al.* (2007) or Mastrogiacomo *et al.* (2004) for applications to the Netherlands). In general, revealed preference data capture actual choices made by individuals and are well suited to short-term forecasting of the effects of small departures from the current state of affairs. To study preferences of people in settings which differ a lot from the current state, it is generally better to use stated preference (SP) data relying on the choices of people in hypothetical situations described in survey questions. This method is commonly used in marketing research and transport economics (see, e.g., Louviere *et al.* (2002)) and is gaining ground in economics (e.g., Barsky *et al.* (1997) or Revelt and Train (1998)). Respondents are provided with information on hypothetical (but potentially realistic) retirement scenarios and are asked to state their choice between several scenarios, to rank the scenarios, or to rate each of the scenarios.[1]

In our analysis we use stated preference data to analyze preferences of Dutch people for early, late and gradual retirement. The main reason for using stated rather than revealed preferences is that we want to estimate preferences for pension plans which do not exist or to which many workers do not have access, such as retirement after age 65 or gradual retirement. Moreover, stated preference data allow for a design where choice opportunities are exactly known and variation in choices is substantial and by construction exogenous to preferences. Stated preference data on retirement of Dutch workers have been collected and analyzed by Nelissen (2001), Bruinshoofd and Grob (2005), Van Soest *et al.* (2006) and Fouarge *et al.* (2008). Compared to these earlier studies, we use richer (and more recent) data from various years and focus on estimating a flexible structural model that can be used to compute income and substitution effects on retirement decisions.

Specifically, survey respondents of ages 25 and older in the CentER panel (a representative sample of the Dutch adult population) were given hypothetical retirement scenarios describing the age(s) of (partial and full) retirement and corresponding replacement rates. Several types of retirement trajectories were considered – three trajectories without gradual retirement and with retirement ages 65 (standard retirement age), 67 (late retirement), and 63 (early retirement) and a partial retirement trajectory. Retirement trajectories were evaluated in both rating and choice questions. The data were collected in 2006, 2007 and 2008, partly for the same respondents (leading to an unbalanced panel).

We use the SP data to estimate an intertemporal utility model in which the

[1]See Van Soest *et al.* (2006) for more discussion on the advantages and potential disadvantages of SP and RP data on retirement preferences.

individual's utility is the discounted sum of within period utilities that depend on employment status (working, partially retired, or (fully) retired) and income in that period. Parameters of the utility function are allowed to depend on observed and unobserved respondent characteristics and on the year of data collection. The estimated model is then used to analyze how retirement preferences differ by background characteristics such as sex, age, and education and how they evolve over the survey years. Simulating the choice of the retirement age under actuarially fair and unfair trade-offs, we then analyze how the preferred retirement age changes if pension income levels change irrespective of the retirement age (the "(pension) income effect"), or if the pension benefit accrual induced by delaying retirement changes (the "price" or "substitution" effect).

Confirming most findings in the international literature, we find large effects of financial incentives on the preferred retirement age. The effects we find are often larger still than the effects found with revealed preferences, which is in line with the fact that we allow for flexible choices without imposing restrictions like mandatory retirement at age 65. According to our simulations of a choice among actuarially fair retirement scenarios at all ages between 60 and 70, an increase in life-time pension incomes by 10% would lower the average retirement age by 3 months (the "income effect"). Changing the compensation for delaying retirement from actuarially fair to 50% of what would be actuarially fair would reduce the average retirement age by 9.7 months (the "substitution effect").

This chapter is organized as follows. Section 5.2 describes the questions and the data on stated retirement preferences. Section 5.3 introduces the model, describes the estimation procedure, and presents the parameter estimates. Section 5.4 presents the results of simulations and the implied estimates of the income and substitution effects. Section 5.5 presents some sensitivity checks. Section 5.6 concludes.

5.2 Data and Stated Preference Questions

The questionnaires were included in the Dutch CentERpanel, administrated by CentERdata at Tilburg University. The CentERpanel covers the population in the Netherlands of ages 16 and older. It is composed of over 2000 households in which one or more adults complete questionnaires at home every week through the Internet. The CentERpanel is not restricted to households with prior access to Internet: households without access are provided with access by CentERdata and are given a set-top box that can be connected to their television set and phone line if they do not have a personal computer. (And households without

a television set are also given a portable TV.) About 75% of all panel members respond to the questions in a given weekend.

The Netspar questionnaire about retirement preferences was fielded in June 2006, June 2007 and June 2008. In each wave respondents were asked to evaluate hypothetical and stylized retirement trajectories, designed to be similar to the choices people potentially face, so that they are perceived as realistic. On the other hand, many of the retirement scenarios are currently not offered by all employers, and in order to make sure we measure individual preferences and not demand side constraints, respondents are explicitly asked to assume that their employer will offer each scenario.

To describe a hypothetical situation, respondents first received an introductory text. Respondents younger than 60 were asked to assume that they would still work when turning 60, that their job at 60 would be similar to their current job and that their employer would fully cooperate with any trajectory. Respondents of age 60 and older got similar instructions with adjusted wording and were asked about the job they had just before turning 60. Before the scenario questions were asked, respondents first reported their number of working hours per week (WH), which was then used to formulate the hypothetical retirement scenarios. Respondents younger than 60 were specifically asked how many hours per week they currently worked, while respondents of age 60 and older were asked how many hours per week they worked for pay just before they turned 60.

Since the trajectories are based on the number of paid working hours WH before retirement and their reduction during gradual and full retirement, it makes little sense to interview people younger than 60 who work zero hours or people older than 60 who worked zero hours at the time they turned 60. Furthermore, some respondents report that they have paid work but also that they have no income. To avoid these problems people who work (or worked when turning 60) for pay less than 3.5 hours per week or whose monthly net income is (or was) less than 45 Euro were not given the scenario questions. Moreover, the questions were not administered to panel members younger than 25, mainly since we thought they probably had not thought much about pensions yet.

This selection leads to a sample in which men and people with high income and high education are overrepresented. The age of respondents is between 25 and 93 years, with medium age of 51 years. The composition of the sample is given in Table 5.1. In total 2978 observations on 1605 respondents are available. 429 people are interviewed in all three waves, 515 people in two waves and 661 people are interviewed just once.

Respondents got four scenarios describing standard, late, early and partial

retirement plans. The scenarios for waves 2006 and 2007 were the following:

Trajectory 1 - standard retirement
WH till age 65, full retirement at age 65. Disposable pension income is [60%/65%/70%] of last net earnings.

Trajectory 2 - late retirement
WH till age 68, full retirement at age 68. Disposable pension income is [80%/85%/90%] of last net earnings.

Trajectory 3 - early retirement
WH till age 62, full retirement at age 62. Disposable pension income is [45%/50%/55%] of last net earnings.

Trajectory 4 - gradual retirement
WH till age [60/62/64], reduced working time to 60 % of *WH* from age [60/62/64] till age [63/65/67], full retirement at age [63/65/67]. Disposable labour income from [60/62/64] till [63/65/67] is [70%/75%/80%] of earnings at age [60/62/64]; pension income after age [63/65/67] (incl. AOW) is [60%/65%/70%] of net earnings at age [60/62/64].

In each wave respondents were randomly allocated into three groups. Based on this, in all trajectories they were offered one of the three replacement rates given in brackets.[2] In the partial retirement trajectory, ages for partial and full retirement were also varied across the three groups.[3] In the 2008 wave, somewhat different trajectories were used, with different replacement rates and a small change in the age of gradual retirement. This was done in order to increase the variation across trajectories and to improve the efficiency of the estimator. The evaluated trajectories in all waves are summarized in Table 5.2.

Respondents evaluated the hypothetical trajectories of standard, late, early and partial retirement by rating each trajectory and by choosing between pairs of trajectories. In the four rating questions the attractiveness of each retirement trajectory was assessed on a ten point scale from 1 (very unattractive) to 10 (very attractive). The answers will be denoted by R1, R2, R3 and R4 for the four scenarios of benchmark, late, early and gradual retirement, respectively. In the

[2]The replacement rates are low compared to replacement rates of actual retirees (see Fouarge *et al.* (2008)) but are reasonably representative of subjective expectations of future replacement rates of current employees. For example, Van Dalen *et al.* (2008) report an average expected replacement rate of 67%.

[3]The order in which the trajectories were presented to the respondents was also randomized.

two choice questions respondents were asked to choose between two trajectories – standard and late retirement (trajectories 1 and 2) and standard and gradual retirement (trajectories 1 and 4). The reported choices are denoted by C1 and C2, respectively.

In Figure 5.1, histograms of the evaluations of the standard retirement trajectory (R1) and their comparisions with the evaluations of late retirement trajectory (R1-R2), early retirement trajectory (R1-R3) and gradual retirement trajectory (R1-R4) are given for the year 2006 and the randomization group of respondents g3 in Table 5.2.[4] In Table 5.3, means and standard errors of evaluations of the four retirement trajectories (R1, R2, R3 and R4) as well as of the choices (C1 and C2) are presented, separately for each wave and for each of the three random assignment groups. The mean of the benchmark evaluations in Figure 5.1(a) is 5.27 (see Table 5.3), with quite large dispersion. Possible reasons for this may be genuine heterogeneity in how attractive this specific scenario is to different respondents, the fact that different respondents may have different response scales, or noise in the assessments.

The histogram in Figure 5.1(b) shows that the benchmark is preferred to late retirement more often than the other way around. This corresponds to the fact that in the choice question C1 (standard versus late retirement trajectory), 73 % of people choose the standard retirement trajectory (see Table 5.3). Figure 5.1(c), where the benchmark is compared to the early retirement scenario, shows that most people give lower ratings to early retirement than to the benchmark. On the other hand, the symmetric distribution of differences R1-R4 (benchmark minus gradual retirement) in Figure 5.1(d) shows that the group preferring the benchmark to gradual retirement is about as large as the group with the opposite preference. In the choice question C2 (standard versus gradual retirement), 58 % of people chose the standard retirement trajectory (see Table 5.3).

There are some statistically significant changes in the average ratings between 2006 and 2007; in particular, many of the mean ratings in 2007 are lower than the corresponding means in 2006, suggesting that respondent evaluations have become more negative. Where comparable,[5] the means in 2008 are not significantly different from those in 2006, but there are some significant differences between 2008 and 2007.[6]

[4]Looking at the differences instead of the levels eliminates response scale differences between respondents (cf. Van Soest et al. 2006).

[5]for example for question R1, group g3 in 2006 got the same replacement rates as group g2 in 2008; see Table 5.2

[6]P-values lower than 0.05 are obtained for R2 – group g2 in 2007 and group g3 in 2008 – and for R3 group g2 in 2007 and any of the groups in 2008.

Comparing the mean evaluations of the three groups in a given year for a given question R1, R2 or R3 shows how the evaluations vary with the replacement rate. Group g1 got the lowest and group g3 the highest replacement rate, except for R3 in 2008, where the replacement rate for all groups was 50 % (see Table 5.2). The evaluation of a retirement scenario with a higher replacement rate is either significantly higher or not statistically different from that of the same retirement scenario with a lower replacement rate. The biggest difference between the groups is found for question R1 (standard retirement trajectory), where trajectories with replacement rates lower than 70 % are evaluated significantly less than the trajectories with replacement rates 70 %. This can be due to the general preference for defaults - the default retirement age in the Netherlands is 65 with an accompanying pension income equal approximately to 70 % of the last earned wage.

Gradual retirement trajectories differ in the replacement rate during partial retirement as well as after full retirement, but also in the age of partial retirement and the age of full retirement. This makes it impossible to directly interpret the differences in evaluations of R4 across groups and years. In order to understand what these evaluations imply, we will use the structural model introduced in the next section.

Table 5.4 compares responses to rating and choice questions of the same respondents in the same wave. Respondents who prefer the standard retirement trajectory to the late retirement trajectory in the choice question (C1=1) also tend to evaluate the standard retirement trajectory higher than the late retirement trajectory in the rating questions (R1>R2). Specifically, 47.4 % of the C1=1 respondents rate the standard retirement trajectory higher, 34.6 % give the same ratings for both trajectories and 18.1 % rate the standard retirement trajectory lower than late retirement. For C1=1 respondents, the mean ratings of the standard and late retirement trajectories are 4.12 and 3.26, resp. Of the other respondents who chose the late retirement trajectory over standard retirement (C1=0), 13.4 % rated the late retirement trajectory lower than the standard retirement trajectory (R1>R2). The other C1=0 respondents either gave a higher rating to the late retirement trajectory (65.5 %) or rated the two trajectories equally (21.1 %). The mean evaluation of the late retirement trajectory (mean R2 = 6.01) by C1=0 respondents is significantly higher than their mean evaluation of the standard retirement trajectory (mean R1 = 4.41).

In the second choice question C2, respondents could choose between standard and gradual retirement. Again, on average, the choices are in line with the ratings (see Table 5.4) but there are also many inconsistencies. For example, 78.4 % of

the respondents who prefer the benchmark trajectory to the gradual retirement trajectory (C2=1) rate the standard retirement trajectory higher (R1<R4) or in the same way (R1=R4), while for 21.6 % the ratings are inconsistent with the choice. The inconsistencies may be due to reporting errors in both choice and rating questions, and Table 5.4 makes clear that it is important to account for these errors in the structural model.

5.3 Model of Stated Retirement Preferences

We use a life-cycle model similar but more general than the one of Van Soest et al. (2006). We assume that the total utility of retirement trajectory q for individual $i = 1, \ldots, I$ in wave $s = 1, 2, 3$, U^q_{is}, has the following form:

$$U^q_{is} = \sum_{t=60}^{100} \rho^{t-60} U^q_{ist}, \tag{5.1}$$

where U^q_{ist} is the utility at age $t = 60, \ldots, 100$ and ρ is the discount factor. The time horizon is fixed at 100 years of age and thus each work – retirement trajectory covers ages from 60 (the earliest retirement age in the scenarios) to 100.

U^q_{ist} is modelled as follows:

$$U^q_{ist} = \alpha^0_{is} + \alpha^p_{ist} P^q_{ist} + \alpha^r_{ist} R^q_{ist} + \alpha^y_{is} y^q_{ist} + \alpha^{py} P^q_{ist} y^q_{ist} \tag{5.2}$$

$$\alpha^a_{is} = X_{is}\beta^a + \gamma^a_{is} + \delta^a_s \qquad a = 0, y \tag{5.3}$$

$$\alpha^b_{ist} = X_{is}\beta^b + \gamma^b_{is} + \delta^b_s + \eta^b t \qquad b = p, r \tag{5.4}$$

$$\gamma^c_i \overset{iid}{\sim} \mathrm{N}(0, \sigma^2_c) \qquad c = 0, p, r, y \tag{5.5}$$

$$\gamma^c_i \perp X_{is} \qquad c = 0, p, r, y \tag{5.6}$$

Here P and R are dummies for partial and full retirement, respectively, and $\perp$ denotes statistical independence. At each age t, a person can be not retired ($P = R = 0$) and working pre-retirement hours (WH), partially retired ($P = 1$, $R = 0$) and working 60% of pre-retirement hours, or fully retired ($P = 0$, $R = 1$). y_t denotes logarithm of the replacement rate, that is the log of net (pension and/or labour) income at age t as a fraction of pre-retirement net earnings. For example, if after tax pension income during full retirement is 70% of pre-retirement after tax earnings then $y = log(0.7)$ at that age. Note that the replacement rates vary by design of each scenario, independent of individual characteristics. Before (gradual) retirement, we have $y = log(1) = 0$.

As apparent from equation(5.4), α^p_{ist} is the preference parameter for partial retirement, determining the utility of partial retirement compared to the utility of not retired at age t for respondent i in wave s. The parameter is assumed to depend on a set of observed characteristics X_{is} at the time of survey s, like gender, age, and education. Moreover, α^p_{ist} can depend on unobserved characteristics of person i, γ^p_i, assumed to be normally distributed with expected value 0 and standard deviation σ^p, independent of observed characteristics X_{is}. Wave effects are captured by the parameters δ_s and the effects of age t in each period considered by η^p_t.

The preference parameter α^r_{ist} for full retirement has the same specification as α^p_{ist}. We expect that the parameters η^p and η^r will be positive because people's valuation of retirement increases with age, due to e.g. deteriorating health.

The coefficient α^y_{is} determines the influence of an income change in full retirement. It is assumed to depend on the observed characteristics X_{is}, an unobserved heterogeneity term γ^y and a survey wave effect δ_s. Thus α^y_{is} is not allowed to vary with age t. The reason is that, with the given design, there would be a high negative correlation between tR_t and ty_t preventing estimation of both coefficients. To solve this problem more variation in the replacement rates in the scenarios would have been needed, but this would also involve the drawback of making the scenarios less realistic.

The influence of an income change in partial retirement on utility is captured by $\alpha^y_{is} + \alpha^{py}$. The parameter α^{py} reflects the difference between the effects of income on utility in periods of partial and full retirement. Note that y^q_{ist} when not retired is always equal to $log(1) = 0$, which is why no second interaction term (between $log(y)$ and R) could be included.

The coefficient α^0_{is} determines the level of utility regardless of labour force status and income. When comparing utility of two trajectories, this coefficient does not play any role. It depends on observed and unobserved characteristics of the individual and may vary across the three waves of the survey, but it does not depend on age – age effects on α^0_{is} would not be identified (because we always consider the age range 60 – 100).

As described in section 5.2, the respondents rated four pension trajectories on a discrete scale from 1 to 10. The observed ratings $R^q_{is}, q = 1, \ldots, 4$ are modeled

as follows:

$$V_{is}^q = U_{is}^q + \epsilon_{is}^{1q} \qquad q = 1,2,3,4 \tag{5.7}$$

$$C_{is}^q = k \Leftrightarrow \mu_{k-1} < V_{is}^q \leq \mu_k \qquad k = 1,\ldots,10 \tag{5.8}$$

$$\epsilon_{is}^{1q} \overset{iid}{\sim} \mathrm{N}(0,\sigma_1^2) \tag{5.9}$$

$$\epsilon_{is}^{1q} \perp X_{is}, \gamma_i^c \qquad c = 0,p,r,y \tag{5.10}$$

ϵ^1 is the "reporting error" in the rating questions. Threshold parameters $-\infty = \mu_0 < \mu_1 < \ldots \mu_9 < \mu_{10} = \infty$ are assumed to be the same for all respondents. For identification, μ_1 is set to 1.5 and μ_9 to 9.5.

In the choice questions respondents choose between the standard retirement trajectory and late retirement (C1) or partial retirement (C2). An observed choice of the standard retirement trajectory is coded by 1, a choice of the alternative is coded by 0. Observed choices C_{is}^1 and C_{is}^2 are modelled as follows:

$$C_{is}^1 = 1 \Leftrightarrow U_{is}^1 - U_{is}^2 > \epsilon_{is}^{21} \qquad C_{is}^1 = 0 \quad \text{otherwise} \tag{5.11}$$

$$C_{is}^2 = 1 \Leftrightarrow U_{is}^1 - U_{is}^4 > \epsilon_{is}^{22} \qquad C_{is}^2 = 0 \quad \text{otherwise} \tag{5.12}$$

$$\epsilon_{is}^{2q} \overset{iid}{\sim} \mathrm{N}(0,\sigma_2^2) \tag{5.13}$$

$$\epsilon_{is}^{2q} \perp X_{is}, \gamma_i^c, \epsilon_{is}^1 \qquad c = 0,p,r,y \tag{5.14}$$

The optimization errors in choice questions $q = 1,2$ are denoted as ϵ^{2q}. Their variance can be different from that of ϵ_{is}^1 because noise levels in ratings and choices may well differ (see Louviere et al. (2002)).

5.3.1 Estimation

The estimation of our model is similar to the estimation of a mixed logit model and other random coefficient models (cf., e.g., Revelt and Train (1998)). These models are usually estimated by simulated maximum likelihood. The likelihood contribution for individual i conditional on unobserved heterogeneity parameters $\vec{\gamma}_i = (\gamma_i^0,\gamma_i^p,\gamma_i^r,\gamma_i^y)'$ can be written as a product of the probabilities of the observed outcomes $R_{is}^q, q = 1,\ldots,4$ and $C_{is}^q, q = 1,2$, the answers to the ratings and choice questions of respondent i in all waves $s = 1,2,3$.[7] Model assumptions in 5.7 and 5.11 imply that these probabilities can be written as follows:

[7] In case of item non-response (if a respondent answers "don't know" or "refuse" to a specific question) or unit nonresponse (if a respondent does not participate in a given survey wave) the corresponding probability is replaced by 1. (We work with the full unbalanced panel.)

$$P(C_{is}^q = k|\mathcal{A}_{is}, \vec{\gamma}_i) = \Phi\left(\frac{\mu_k - U_{is}^q}{\sigma_1}\right) - \Phi\left(\frac{\mu_{k-1} - U_{is}^q}{\sigma_1}\right) \quad k = 1, \ldots, 10$$
$$q = 1, 2, 3, 4 \quad (5.15)$$

$$P(C_{is}^1 = l \mid \mathcal{A}_{is}, \vec{\gamma}_i) = \Phi\left((-1)^{1-l}\frac{U_{is}^1 - U_{is}^2}{\sigma_2}\right) \quad l = 0, 1 \quad (5.16)$$

$$P(C_{is}^2 = l \mid \mathcal{A}_{is}, \vec{\gamma}_i) = \Phi\left((-1)^{1-l}\frac{U_{is}^1 - U_{is}^4}{\sigma_2}\right) \quad l = 0, 1, \quad (5.17)$$

where $\mathcal{A}_{is} = \{X_{is}, P_{ist}^q, R_{ist}^q, y_{ist}^q, \beta^c, \delta_s^c, \eta^b, \quad c = 0, p, r, y, \quad b = p, r, \quad t = 0, \ldots, 40\}$ is the set of all relevant individual and trajectory characteristics and parameters and Φ is the standard normal distribution function.

The (unconditional) likelihood contribution for individual i can be written as a four dimensional integral:

$$\int\int\int\int \prod_{s=1}^{3}\prod_{q=1}^{6} P\left(C_{is}^q = k_{is}^q \mid \mathcal{A}_{is}, \vec{\gamma}_i\right) f(\vec{\gamma}_i)\, d\vec{\gamma}_i, \quad (5.18)$$

where f denotes the density of the vector of random coefficients. The assumption in equation (5.5) implies that the density of $\vec{\gamma}_i$ can be rewritten as a product of univariate normal densities:

$$f(\vec{\gamma}_i) = \prod\nolimits_{c=0,p,r,y} \phi(\gamma_i^c/\sigma_c)/\sigma_c. \quad (5.19)$$

Since it is not feasible to compute the integral numerically we approximate the integral using simulated values of the random coefficients and use simulated maximum likelihood (see, e.g., Gourieroux and Monfort (1996)), replacing (5.18) by:

$$\frac{1}{\text{Sim}} \sum_{\text{sim}=1}^{\text{Sim}} \prod_{s=1}^{3}\prod_{q=1}^{6} P\left(C_{is}^q = k_{is}^q \mid \mathcal{A}_{is}, \tilde{\gamma}_{i,\text{sim}}^c,\ c = 0, p, r, y\right), \quad (5.20)$$

where Sim is the number of simulations and $\tilde{\gamma}_{i,\text{sim}}^c$ is a random draw from a normal distribution with mean zero and standard deviation σ_c. Usually a large number of pseudo-random draws is needed to assure a reasonably low simulation error in the estimated parameters. The number of draws and thus the time the estimation procedure takes can be substantially reduced (keeping the same simulation variance) by using quasi-random numbers of Halton sequence (see

Train (2003)). The number of draws per individual is 500.[8]

Estimates of the covariance matrix of the parameter estimates are based upon the asymptotic result from Gourieroux and Monfort (1991). One of the key assumptions is that $\sqrt{N}/\text{Sim} \to 0$ if $N, \text{Sim} \to \infty$, where N is number of observations and Sim number of simulations for each respondent (see, e.g., Gourieroux and Monfort (1996) for details on simulated maximum likelihood).

5.3.2 Estimation Results

Parameter estimates are presented in Table 5.5. The first column ("α^0") presents the coefficients β^0, which determine α^0, the utility in year t of the pre-retirement benchmark status ($y = 0$, $D = 0$, $P = 0$). Since the other parameters drive the change in utility due to a deviation from this benchmark, β^0 affects the ratings of the scenarios but not the choices. Many of the parameters in β_0 are significant, implying substantial heterogeneity in the (absolute) utility ratings. For example, the age groups 45-64 give less positive utility ratings than the younger and older age groups, and the lower income groups give more positive ratings than the middle and high income groups. Respondents with a small part-time job are more positive than those who work(ed) longer hours. Note that this may not be a causal effect – it may be due to common preference factors that drive both current working hours and desired future working hours. The same remark applies to all included employment status variables.

The large and significant estimate of γ_0 implies that there is also substantial heterogeneity that is not captured by the observed respondent characteristics. The significant estimates of δ_2 and δ_3 imply that utility ratings in 2007 and 2008 were less and more positive than those in 2006, respectively. These time effects might reflect, for example, temporary effects due to the political debate at the time of the survey.

Parameters β^p and β^r in the second column ("α^p") and third column ("α^r") determine how the differences in utility between partial retirement and pre-retirement (α^p) and between full retirement and pre-retirement α^r) vary with respondent characteristics. We do not find a significant effect of gender, education, home ownership or partnership status. The utility of partial retirement is significantly lower for the older birth cohorts, while no significant cohort effect on the utility of full retirement is found. Keeping the other variables constant, the higher income respondents attach higher utility to working part-time or not

[8] Estimated coefficients using Sim = 600 or Sim = 700; were virtually identical to those with Sim = 500. For the four random coefficients we use Halton sequences with primes $3, 5, 7$ and 11.

working at all, reflecting a life-time income effect if leisure is a normal good. Part-time workers have the largest preference for partial as well as full retirement. Full-time workers value partial retirement more than non-workers but less than part-timers.

The parameters δ_s^p and δ_s^r indicate how the evaluations of partial and full retirement vary with the time of data collection. The utility of part-time work (compared to the utility of full-time work) is significantly lower in 2008 than in 2006 or 2007, suggesting that preferences for partial retirement have decreased.

The significant estimate of η^r implies that respondents attach increasing utility to full retirement when they get older. This may reflect that expected health deterioration at older ages is seen as an impediment to full-time work. The small and insignificant estimate of η^p implies that such an impediment much less applies to part-time work and suggests that partial retirement might make it easier to keep people with a health concern in the labour market.

Although we have included many observed characteristics of respondents, we still find significant unobserved heterogeneity in α^p. On the other hand, unobserved heterogeneity in α^r is virtually zero (and insignificant).

The last column indicates the effect of the log replacement rate during full or partial retirement. A larger replacement rate is valued significantly less by the age cohort 55−64 than by the youngest and oldest age cohort. The effects of other respondent characteristics are not significant at the 5% level. Still, the large and significant estimate of the unobserved heterogeneity parameter γ^y shows that there is substantial dispersion in how respondents value a higher replacement rate.

The negative estimate of α^{py} implies that the utility from an increase in income is significantly lower during partial retirement than during full retirement. The estimated value of the discount factor ρ is equal to 0.89 and it is very accurately determined with a standard error of only 0.005. This also captures the mortality rate since mortality is not explicitly taken into account.

Finally, the estimated standard deviations of the error terms imply that the amount of noise is much larger in the ratings than in the choices: the estimate of σ_1 is more than three times larger than that of σ_2. For a given level of noise, ratings of a set of scenarios would provide more information than only the choice among these scenarios, but this difference is counteracted by the difference in noise levels.

5.4 Simulations

In this section, we discuss the implications of the model estimates. We first discuss how the preferences for early and late full and gradual retirement vary with background characteristics. Then we show how people respond to a change in pension income in partial and full retirement. We also simulate choices among actuarially neutral trajectories with retirement age varying from 60 to 70. Finally, we analyze the (pension) income and substitution effects on the preferred age of retirement. The simulations are all based on the estimated parameters in Table 5.5 of the previous section.

5.4.1 Comparing to the Benchmark

Simulated probabilities presented in Tables 5.6 and 5.7 are computed in the following way. For each respondent i in each year s, we first compute the probability of choosing the alternative scenario if the choice is between this alternative and the benchmark scenario (retirement at age 65, replacement rate 70%) only. This probability takes into account observed and unobserved individual heterogeneity and the optimization error in the choice questions (ϵ_2). These probabilities are averaged over the sub-samples of respondents with observed characteristics as indicated in the tables. For example, the number 1.36 in the first row ("Late 1") and eighth column of Table 5.6 indicates that the probability that a person of age 55-64 chooses the Late 1 trajectory rather than the benchmark retirement trajectory is 1.36 %.

The scenarios are defined in columns 2–5 of the table; they are taken from Van Soest et al. (2006). The first six – Late 1 to Early 3 – do not involve gradual retirement. Scenarios Late 1, Late 2 and Late 3 describe late retirement at age 70 with net pension incomes equal to 90 %, 100 % and 110 % of net pre-retirement earnings, respectively. Simulated probabilities show that most people prefer the benchmark to these late retirement trajectories. In particular, only 3 % of the people would prefer postponed retirement with a replacement rate of 90 % to benchmark. With increasing replacement rates, the number of people choosing postponed retirement increases, but even with a replacement rate of 110% (a compensation for late retirement that is more than actuarially fair), only 11% of all respondents would opt for late retirement.[9] The final three columns give the choice probabilities by age group. Particularly in the age groups 45-64 very few respondents would choose late retirement.

[9]The benchmark with retirement age 65 and replacement rate 70 % is actuarially equivalent to late retirement at age 70 and replacement rate 103 %; see also Table 5.7 and its discussion.

Scenarios, Early 1, 2 and 3 describe early retirement at age 62 with replacement rates equal to 50, 60 and 70 % of net pre-retirement earnings. Scenario Early 1 is preferred to the benchmark by 13 % of the respondents. An increase in the replacement rate substantially increases the attractiveness of early retirement: scenario Early 2 with replacement rate 60 % is already preferred to the benchmark by more than a quarter of the respondents, and scenario Early 3 with replacement rate 70 % is preferred to the benchmark scenario by 57 % of all respondents. The annual incomes in this scenario differ from those of the benchmark scenario only during the period from age 62 to age 65. The utility of being fully retired compared to being at work at these ages compensates the decrease in utility due to the lower income during early retirement. Particularly in the age group 45-64, many respondents would be willing to pay this rather low (and actuarially less than fair) price for early retirement.

The last six scenarios, Partial 1 to Early partial, involve gradual retirement. Partial 1, 2 and 3 have partial retirement at age 63 and full retirement at age 67, with three different replacement rates. On average, respondents appear to be indifferent between Partial 1 and the benchmark. An increase in the replacement rate during partial retirement (Partial 2, by 15%-points) or full retirement (Partial 3, by 10%-points) makes gradual retirement more attractive, but the effect is much stronger in the latter case. This is mainly a consequence of the negative estimate of α^{py} which reduces the importance of the replacement rate during partial retirement compared to that during full retirement.

In scenarios Late partial 1 and 2, the partial retirement age is 65 and the full retirement age is 70 – the same age as in Late 1, 2 and 3. Scenario Late partial 1 offers a 20% −points higher replacement rate than the benchmark in return for working 60 % of the pre-retirement working week for five years. This scenario is found more attractive than the benchmark scenario by 17 % of the sample, mainly in the youngest and oldest age cohorts. Late partial 2 increases the replacement rates by 10%-points compared to Late partial 1. Accordingly, the fraction of people preferring this scenario to the benchmark rises to 27 %. These fractions are much higher than the fractions preferring to work until age 70 without gradual retirement. Almost no-one wants to work their full pre-retirement hours till age 70, but many more people are willing to work a reduced number of hours until this age.

Finally, the scenario Early partial offers partial retirement at age 60 and full retirement at age 65. About 60 % of the respondents prefer this to the benchmark, although the corresponding pension income is lower than what would be actuarially fair. For many respondents, the early partial retirement scenario is

apparently also more attractive than scenario Early 2, which gives the same replacement rate after age 65 but has immediate full retirement at age 62. This shows that early and late gradual retirement may be attractive alternatives for early and late full retirement. Early gradual retirement is particularly attractive for the age group 54-64, while the youngest and oldest age groups often prefer late gradual retirement.

The results for the complete sample can be compared with those of Van Soest et al. (2006, Table 9, final column)[10] who used a similar methodology with older data and a less flexible model. Most results are qualitatively similar though we find a smaller tendency to choose the gradual retirement scenarios. Moreover, we find an even smaller effect of increasing the replacement rate during partial retirement, in line with our negative estimate of α^{py}.

In the second and third panel of Table 5.6, we present simulated choice probabilities for various subsamples of respondents characterized by background characteristics other than age. The differences between groups are generally smaller than the differences between age groups in the top panel. Women have somewhat lower preferences for late retirement trajectories and higher preferences for early retirement trajectories than men. They also seem to be less interested in gradual retirement. Preferences for early or late retirement hardly vary with education level, but the higher educated have a stronger preference for gradual retirement than other educational groups. Respondents living with a partner have a stronger preference for all forms of early retirement and an accordingly larger distaste for late retirement than respondents not living with a partner. The same applies to home owners versus renters. The choices of the high income groups are more sensitive to the replacement rate than those of lower income groups, particularly when it comes to early retirement. Higher income respondents are also more interested in gradual retirement. Full-time workers have the largest tendency to choose late gradual retirement, while part-timers (working 16-32 hours per week) have the strongest preference for early retirement or early gradual retirement. Comparing the simulated probabilities over the years of the data collection, we find that the attractiveness of all gradual retirement scenarios is falling over time. This can also explain why we find fewer choices of gradual retirement than Van Soest et al. (2006). In 2008, we also find a substantially smaller tendency to choose early retirement and a somewhat increased tendency to choose late retirement. These results may reflect changing social norms.

[10]Since Van Soest et al. (2006) cannot estimate the noise level in choice questions, they use either the noise level in ratings or noise level zero in their simulations. Our results are better comparable to the latter case (final column in their Table 9), since our estimates imply that the noise level in choices is much smaller than in ratings.

5.4.2 Choice of Retirement Age

Table 5.7 considers the choice between the benchmark (retirement at age 65; replacement rate 70%) and a scenario that is actuarially equivalent[11] to the benchmark but has a different retirement age. Gradual retirement is not considered here. The actuarially fair replacement rates (in the second column) are taken from Queisser and Whitehouse (2006), on the basis of a 2 % interest rate, average life expectancy for OECD countries, and price indexation of pensions.

Like the previous table, the table presents the simulated probabilities of choosing full retirement at the alternative age (third column). For example 19.8 % of all people would prefer to retire at age 62 with a replacement rate of about 57% rather than at age 65 with replacement rate 70%. The simulated probabilities show that most people prefer standard retirement at age 65 to actuarially equivalent early as well as late retirement.

In the remaining simulations we consider the choice between 11 options: retirement at age $60, 61, \ldots, 69$ or 70, without any opportunities for gradual retirement, and for a variety of (retirement age dependent) replacement rates. The baseline case is the set of 11 actuarially equivalent scenarios already presented in Table 5.7, but instead of comparing each of these scenarios with the benchmark, we now consider the choice between all 11 scenarios. Column "rr" of Table 5.8 presents the probability of each choice averaged over the complete sample, as well as the corresponding average retirement age for this baseline case.[12] The mode is 65 years and the mean desired retirement age is 65.08 years, corresponding to the symmetry we already found in Table 5.7. Still there is also substantial dispersion, with, for example, more than 20% choosing to retire at age 63 or earlier, and more than 23% opting for retirement at age 67 or later.

The other columns of the table give insight in the "(pension) income effect" on the preferred retirement age, i.e. how does the preferred retirement age change if the total value of life-time pension income changes, irrespective of the retirement age. To compute it, we increased or decreased the replacement rates in all 11 scenarios by a fixed percentage – 10, 20 or 30 % – and calculated the simulated probabilities for each new choice set. These simulated probabilities are presented in the other columns of Table 5.8, labeled "0.7 rr" (replacement rates reduced by 30%), "0.8 rr",..., "1.3 rr".

An increase in the replacement rates makes early retirement more attractive

[11] Actuarial neutrality of pension trajectories requires that the present value of accrued pension benefits for working an additional year is the same as in the year before. See Queisser and Whitehouse (2006) for a discussion of actuarial neutrality and related concepts.

[12] Both the unobserved heterogeneity terms and the optimization errors are taken into account.

and makes late retirement less attractive: in columns "1.1 rr", "1.2 rr" and "1.3 rr" we observe a gradual increase of early retirement choices and a decrease of late retirement choices. For example, the probability to retire at age 63 or earlier rises from 20.3% in the baseline case ("rr") to 29.3% when all pension incomes would be raised by 30%. At the same time, the percentage retiring at age 67 or later would fall from 23.3% to 10.7%. For the lower replacement rates (columns "0.9 rr", "0.8 rr" and "0.7 rr") we observe the opposite trend. A graphical illustration of these shifts in probability distributions for changing pension incomes is presented in Figure 5.2. These changes can be seen as pure income effects, since the accruals, i.e., the rewards for retiring earlier or later, do not change (in relative terms), implying that the substitution effects are zero. The implied results for the average retirement age show that the income effects are of the expected negative sign and substantial: a 10% increase in all replacement rates would, for example, reduce the average age of preferred retirement by three months (see the bottom rows of the table).

The income effect can be compared with the "pension wealth" effect found by Euwals *et al.* (2007) who analyze preferences for early retirement of Dutch public sector employees, using administrative data from the main public sector pension fund. They find that reducing pension wealth by 100,000 euros would induce the average worker to postpone retirement by 5 or 6 months (p.21). The lump sum of 100,000 euros corresponds to an annuity of about 25% of average pre-retirement earnings and is therefore similar to an increase of the replacement rate by somewhat more than 30%. According to our estimates, this would raise the average retirement age by more than 8 months, which is larger than the result of Euwals et al. This is not so surprising since we look at desired retirement instead of actual retirement and allow for quite flexible choices (any retirement age from age 60 to 70), while the literature provides evidence that retirement choices are often much more restricted and certainly in the Netherlands, actual opportunities for retiring after age 65 are scarce (cf., e.g, Van Solinge and Henkens (2007)).

With some additional assumptions, we can also roughly compare these income effects with the "wealth effects" found by Brown *et al.* (2006) who look at the effect of (expected and unexpected) inheritances on retirement using the US Health and Retirement Study. One of their dependent variables is the two-year (i.e., wave to wave) retirement rate, with a sample average of 19.2% (Table 5 in Brown et al.). They find that a $\$100,000$ inheritance increases this rate by about 2.1%-points. To compare this with our findings, we consider the retirement rate at age 62 or age 63, which is 17.7% in our baseline case with actuarially fair trade offs ((5.77+11.38)/(100-0.83-2.32), see Table 5.8). A $\$100,000$ lump sum transfer

at age 62 would roughly correspond to an annuity of about 15 to 20% of average annual pre-retirement earnings. The retirement rate at age 62 or 63 for this higher replacement rate can be derived from the columns "1.2 rr" and is about 22.6%, 4.9%-points higher than in the baseline case. This is much larger than the 2.1% found by Brown et al. Note, however, that in their later analysis, Brown et al. find larger effects of unexpected inheritances than of expected inheritances, a distinction not made for this particular estimate, and our estimate probably corresponds more to the effect of an unexpected inheritance[13].

In Table 5.9 we present the income effects on the mean preferred retirement age for different socioeconomic groups. The first column concerns the baseline case. The main differences across socio-economic groups here are the age differences: the age groups 45-64 prefer to retire earlier than the younger and older age groups. The other columns present the income effects in terms of changes (in months) of the average preferred retirement age, computed in the same way as in the bottom row of Table 5.8. The sign of the income effect is the same for all subgroups, but there is some variation in magnitude. For example, the income effects increase with socio-economic status (education level and income) and are relatively small for workers with a small part-time job.

Substitution effects on the retirement age are presented in Table 5.10. The baseline (column "rr") is the same as in Table 5.8. The alternatives do not change generosity of pensions when retiring at age 65, but increase or decrease the accruals, i.e., the rewards for retiring later or the penalty for retiring earlier, giving "flatter" or "steeper" relationships between the retirement age and the replacement rate. To be precise, the new replacement rates are equal to $70 + x(rr - 70)$, where rr are the replacement rates in the actuarially neutral scenarios (Table 5.7), 70 is the replacement rate in the benchmark scenario with retirement age 65 and x is a multiplication factor. For example for $x = 0.5$ the new replacement rate when retiring at age 60 is equal to $70 + 0.5(50.26 - 70) = 60.13\%$, for retirement age 61 it is $70 + 0.5 * (53.45 - 70) = 61.73\%$, etc. If x is equal to 1, the replacement rates are those of the baseline case with actuarially equivalent trajectories. If $0 \leq x < 1$, the accruals are negative and early retirement scenarios become financially more attractive. If $x > 1$, accruals are positive, implying a stronger financial incentive to retire later. In our simulation, we consider x equal to $0, 0.33, 0.5, 1, 2$ and 3. In the extreme case, $x = 0$, the replacement rate

[13]We cannot compare our estimates to these later estimates of Brown et al., since these use the dependent variable "retiring earlier than expected" which we cannot construct. Substantial negative income effects for Dutch workers are also implied by the simulation results of Mastrogiacomo et al. (2004, p.790); the magnitude of these effects is not comparable to our estimates since they look at changes in pre-retirement wages.

is equal to 70 irrespective of the retirement age.

In the baseline choice set "rr", people on average prefer to retire at age 65.1, as we saw before. With the positive accruals in column "70+2(rr-70)", the average retirement age would increase by almost one year, since later retirement is made more attractive. For example, the percentage preferring to retire at age 67 or later would increase from 23.3% to 44.9%, while the percentage wanting to retire at age 63 or earlier would drop from 20.3% to 10.3%. On the other hand, if the accruals are reduced so much that the only "penalty" for retiring early is a lower income during the years of early retirement (column "70"), the average retirement age would fall by almost 1.75 years, with about 56% wanting to retire at age 63 or earlier. The main reason why many respondents do not choose to retire even earlier according to our model estimations is the effect of age on utility when retired, which implies that, keeping income constant, for many respondents retirement is less attractive than pre-retirement at age 60 or 61.

The substitution effect can be compared with the "price effect" of Euwals et al. (2007) who find that increasing the peak value by 100,000 euros would induce a worker to postpone retirement by about 8 months. Changing from column "rr" to column "70+3(rr-70)" increases the reward for postponing retirement in terms of pension income per year at age 65 from 5%-points to 15%-points of pre-retirement earnings, corresponding to a change in peak value (defined as the increase in lifetime wealth if the worker decides to continue working for one year) of about 40,000 euros for the average worker. Our estimates would imply that this increases the average retirement age by 18 months. The substitution effect we find is therefore much larger than the effect found by Euwals et al. (2007). As for the income effect, a plausible explanation for the difference is that we look at desired retirement allowing for maximum flexibility - each age between 60 and 70 is possible, whereas Euwals et al. (2007) consider actual retirement, which may also be affected by implicit or explicit restrictions imposed by the employer like mandatory retirement at age 65.

The final column of Table 5.10 (column "90,70") shows the response to an arrangement that mimics a stylized version of the generous early retirement arrangements in the Netherlands and other countries as they existed in the nineties: a fixed replacement rate of 70% after age 65 (irrespective of the retirement age), and a replacement rate of 90% between early retirement and age 65 (irrespective of the early retirement age). As expected, this makes early retirement even more attractive than the arrangement which also gives a replacement rate of 70% in the years between early retirement and age 65. More than 75% would prefer to retire at age 63 or earlier, and almost 50% of the respondents would choose re-

tirement at age 62 or earlier. Retirement at age 60 remains uncommon, because of the estimated negative utility of retirement at this age. The average retirement age would drop by almost 30 months compared to the baseline choice set with actuarially equivalent choices. This estimate fits in the range of estimates given by Kapteyn and De Vos (2004, p. 493) who simulate a "common reform" from the actual system with generous early retirement opportunities to an approximately actuarially fair system with retirement between age 60 and age 65. Depending on their model specification, they find smaller or larger effects than we do. Again, we would expect to find larger effects than Kapteyn et al. (2004) because we also allow for retirement beyond age 65.

Table 5.11, presents the substitution effect for various socioeconomic groups. The first column is the same as in Tabls 5.9, giving the average preferred retirement age for the baseline of actuarially fair choices. The other columns show the substitution effects expressed as the number of months the average preferred retirement age by subgroup changes when the rewards for retiring later increase or decrease (as in the final row of Table 5.10). The results are comparable to those in Table 5.9: the groups with the higher income effects also have the higher substitution effects (high income, high education level). The group of respondents with a small part-time job generally seems less sensitive to financial incentives than all other groups.

5.5 Sensitivity Analysis

In this section, we investigate the sensitivity of the simulated income and substitution effects presented in Tables 5.8 and 5.10 on the preferred retirement age for some of the specification choices made in our model. We compare the results of the benchmark model, from now on referred to as $M0$, to those of five alternative models, named $M1$, $M2$, $M3$, $M4$ and $M5$. The estimated income effects are presented in Table 5.12, and Table 5.13 presents the substitution effects.

Model $M1$ extends the benchmark model by adding a quadratic term $\alpha^{y^2} y_{ist}^2$ to the right hand side of equation 5.2. Differences in the simulated income effects and substitution effects calculated using model $M1$ and the benchmark model $M0$ are small. The estimated parameter α^{y^2} is not significantly different from zero. It demonstrates that extending the benchmark model with a quadratic term of log income neither leads to a better fit nor to different conclusions.

Models $M2$ and $M3$ are simplified versions of the benchmark model $M0$. They both incorporate fewer observed characteristics X_{is} than $M0$. Model $M2$ uses just sex and age of the respondents while model $M3$ includes sex, age, education

and partnership status. Compared to Model $M0$, $M3$ drops income, number of paid working hours and home ownership, variables which might be determined by the same unobserved characteristics that drive the tastes for work versus leisure and therefore also retirement preferences, so that their effects are not necessarily causal. As shown in Tables 5.12 and 5.13, the differences in the simulated income and substitution effects of models $M0$, $M2$ and $M3$ are negligible, demonstrating the robustness of our results for including these variables that are in a sense potentially endogenous.

Finally, we consider two models in which the discount rate is fixed to a given value rather than estimated. The discount rate appeared to be numerically the hardest parameter to estimate - with a fixed discount rate, estimating the model appeared to be much faster than when also estimating the discount rate. This is why we wanted to investigate the consequences of setting the discount rate to a specific value. In the benchmark model $M0$ the estimated discount factor is $\rho = 0.89$. In models $M4$ and $M5$ we set the discount factor to 0.95 and 0.85, respectively. The results in Table 5.12 show that the income effects crucially depend on the discount rate. Setting the discount rate to a very low value (0.85, model M5) leads to much larger estimates of the income elasticities than setting it to a higher value (0.95, model M4) – in the latter model, the estimates are less than half as large as the estimates in the former model. The benchmark model with its estimated discount rate of 0.90 gives income effects in between those of the models with $\rho = 0.85$ and $\rho = 0.95$.

On the other hand, the columns in Table 5.13 except the last one show that the discount rate hardly affects the estimates of the substitution effects. The effects in the final column of this table, the simulation mimicking the generous early retirement opportunities of the nineties, are a combination of (negative) income and (negative) substitution effects. Accordingly, model $M5$ with the largest negative income effects also gives the largest negative effect of changing from actuarially fair trade-offs to this system that rewards early retirement. In Model $M4$ the negative income effect is much smaller, leading to a total effect that is also much smaller than according to the model with estimated discount rate. This leads to the conclusion that fixing the discount rate to the wrong value may bias the estimates of the effects of policy simulations.

5.6 Conclusion

This chapter analyzes retirement preferences using stated preference data. We work with unbalanced panel data on Dutch individuals, collected in 2006, 2007

and 2008. In each year, respondents evaluated four types of hypothetical retirement scenarios - standard retirement (age 65), late full retirement, early full retirement and partial retirement. To study the preferences over different retirement trajectories in detail, we use an intertemporal utility model of labour force participation and income for periods of work and retirement. The model is estimated by simulated maximum likelihood.

One of the main findings is that people prefer gradual retirement trajectories to the benchmark retirement trajectory (retirement age 65, replacement rate 70 %), although these offer actuarially less income than the benchmark trajectory. Most people do not wish to work full time to high ages even if relatively high income in retirement period is offered. The fraction of people willing to work very long can be increased if we allow for gradual retirement. Gradual retirement seems therefore to be an appropriate tool to keep older people working.

Another key finding concerns the change of preferences over time. Taking into account the results presented in both our study and in Van Soest *et al.* (2006), which uses data collected by CentER in year 2004, we can observe a decrease in preferences for early retirement and an increase in preferences for late retirement in period 2004-2008. This may reflect changes in social norms.

We study the income effect on preferred retirement age. First, we let people choose between retirement scenarios with full retirement at ages between 60-70 years which are actuarially equivalent to the benchmark scenario. Then people could choose between all actuarially neutral scenarios with higher or lower pension income levels than in the benchmark choice set. We find that the income effect is negative and substantial. The preferred retirement age for the benchmark choice set is 65.1 years. The increase of pension income by 10 % lowers the preferred retirement age by 3 months. A decrease of the income by 10 % increases the preferred retirement age by 3.2 months.

Similarly, we calculate the substitution effect by changing the accruals, keeping the replacement rate when retiring at the normal retirement age of 65 at its benchmark value of 70%. We find substantial substitution effects. For example, reducing the accruals to half their actuarially neutral values would reduce the average retirement age by almost 10 months. The results also explain the popularity of generous early retirement opportunities as they existed in the Netherlands until the nineties - according to our simulations they reduced the average retirement age of those who had access to them by almost 2.5 years.

Our model can be extended in several ways. It would be reasonable to include for example savings or joint decision making of spouses. Changing the formulation of the hypothetical retirement scenarios should be considered, to make the

hypothetical retirement options more understandable for the surveyed people.

5.A Tables and figures

Table 5.1: Sample Composition and Background Characteristics

background characteristics	percent
male	63.6
age 34-	16.0
age 35-44	18.4
age 45-54	25.4
age 55-64	20.0
age 65+	20.2
education low (basis, VMBO)	26.9
education medium (HAVO, VWO, MBO)	31.2
education high (HBO, WO)	41.9
partner	75.3
income low (net inc 1000-)	15.6
income medium (net inc 1001-2000)	55.7
income high (net inc 2001+)	28.7
work hours 15-	7.1
work hours 16-32	25.1
work hours 33+	67.8
own house	75.0
wave 1 - year 2006	34.7
wave 2 - year 2007	37.4
wave 3 - year 2008	27.9

Note: 2978 observations; 429 respondents participated in all three waves, 515 in two waves, and 661 in one wave.

Table 5.2: Description of Pension Trajectories in SP Questions

		waves 1,2			wave 3		
trajectory		g1	g2	g3	g1	g2	g3
1 - standard	age full	65	65	65	65	65	65
	rr full	60	65	70	65	70	75
2 - late	age full	68	68	68	68	68	68
	rr full	80	85	90	75	85	95
3 - early	age full	62	62	62	62	62	62
	rr full	45	50	55	50	50	50
4 - partial	age part	60	62	64	61	61	64
	age full	63	65	67	65	65	68
	rr part	70	75	80	100	75	85
	rr full	60	65	70	60	70	80

Note: In each wave, people were randomly assigned to one of three groups g1, g2 or g3, with different replacement rates. Each respondent evaluated four trajectories defined by partial retirement age (age part), replacement rate in partial retirement (rr part), full retirement age (age full) and replacement rate in full retirement period (rr full).

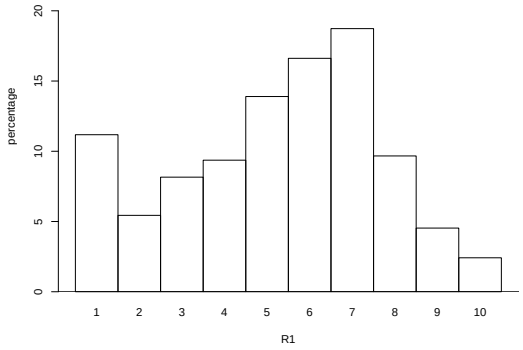
Table 5.3: Means and Standard Deviations of Ratings and Choices

		wave 1				wave 2				wave 3				waves
		all	g1	g2	g3	all	g1	g2	g3	all	g1	g2	g3	all
R1	mean	4.09	3.65	3.41	5.27	3.90	3.36	3.38	4.99	4.75	3.91	5.24	5.12	4.20
	s.d.	2.34	2.09	2.08	2.40	2.35	2.13	2.12	2.44	2.45	2.27	2.41	2.45	2.40
R2	mean	4.21	4.19	3.90	4.56	3.82	3.69	3.72	4.07	4.12	3.73	4.17	4.47	4.04
	s.d.	2.78	2.69	2.71	2.90	2.60	2.50	2.57	2.71	2.75	2.59	2.69	2.91	2.71
R3	mean	3.12	2.93	3.15	3.29	2.98	2.86	2.91	3.18	3.43	3.48	3.43	3.39	3.16
	s.d.	2.01	1.95	2.07	2.00	1.98	2.00	1.92	2.02	2.20	2.36	2.10	2.13	2.06
R4	mean	4.69	4.71	4.49	4.89	4.40	4.47	4.35	4.38	4.81	4.60	5.54	4.34	4.61
	s.d.	2.26	2.13	2.24	2.41	2.30	2.28	2.14	2.46	2.50	2.43	2.20	2.68	2.35
C1	mean	0.70	0.70	0.67	0.73	0.70	0.65	0.70	0.77	0.75	0.75	0.76	0.76	0.72
	s.d.	0.46	0.46	0.47	0.45	0.46	0.48	0.46	0.42	0.43	0.44	0.43	0.43	0.45
C2	mean	0.34	0.23	0.21	0.58	0.35	0.20	0.23	0.62	0.45	0.37	0.32	0.66	0.37
	s.d.	0.47	0.42	0.41	0.49	0.48	0.40	0.42	0.49	0.50	0.48	0.47	0.47	0.48

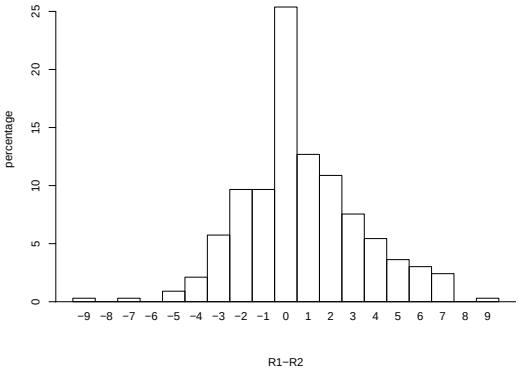
Note: choices C1 and C2 coded as 1 if benchmark trajectory (R1) is chosen; 0 otherwise.

Figure 5.1: Histograms of the evaluations of the standard retirement trajectory (benchmark) and their comparison with the evaluations of late, early and partial retirement trajectories for wave 1, group 3.

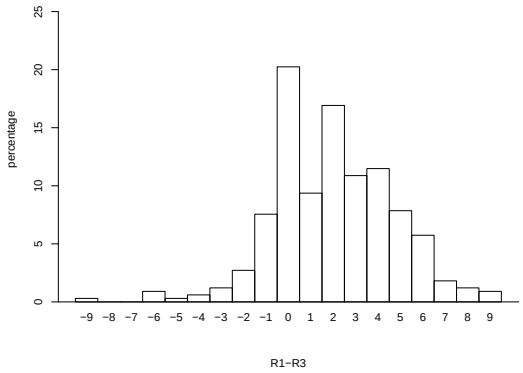
(a) benchmark



(b) benchmark - late retirement



(c) benchmark - early retirement



(d) benchmark - partial retirement

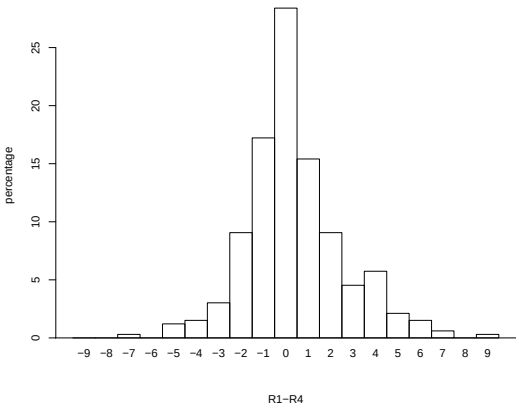


Table 5.4: Ratings and Choices

	percentage			mean	
	R1>R2	R1=R2	R1<R2	R1	R2
C1=1	47.39	34.56	18.05	4.12	3.26
C1=0	13.40	21.12	65.48	4.41	6.01
	R1>R4	R1=R4	R1<R4	R1	R4
C2=1	44.37	34.02	21.60	4.69	3.87
C2=0	16.03	25.07	58.90	3.91	5.06

Note: In the first choice question respondents could choose the standard retirement trajectory (C1=1) or the late retirement trajectory (C1=0). For each of these two choices the table shows how often the rating of the standard retirement trajectory was higher than the late retirement trajectory (R1>R2), the same (R1=R2), or lower (R1<R2). Similarly, in the second choice question respondents chose between the standard retirement trajectory (C2=1) and the gradual retirement trajectory (C2=0) and the table shows how the ratings (R1 and R4) compared to the choice.

Table 5.5: Estimation Results

	α^0		α^p		α^r		α^y	
	Coeff.	T-val.	Coeff.	T-val.	Coeff.	T-val.	Coeff.	T-val.
const	-0.471	-4.777	0.039	0.558	-0.771	-4.552	0.470	2.936
male	0.018	0.619	0.020	0.680	-0.027	-0.340	0.085	1.022
age 45-54	-0.149	-5.070	-0.006	-0.209	0.121	1.518	-0.085	-1.056
age 55-64	-0.100	-3.141	-0.054	-1.694	0.018	0.207	-0.176	-1.994
age 65+	0.091	2.895	-0.063	-2.046	-0.051	-0.616	0.102	1.172
education mid	-0.020	-0.668	0.011	0.381	0.041	0.517	0.023	0.279
education high	0.029	0.964	0.043	1.489	0.056	0.711	0.118	1.429
partner	-0.057	-2.071	0.024	0.905	0.056	0.762	-0.033	-0.440
income mid	-0.102	-2.755	0.066	1.747	0.111	1.085	0.077	0.734
income high	-0.130	-2.918	0.094	2.047	0.187	1.510	0.141	1.104
work hours 16-32	-0.135	-2.827	0.175	3.477	0.286	2.125	0.232	1.671
work hours 33+	-0.105	-2.056	0.131	2.468	0.163	1.139	0.105	0.712
own house	-0.047	-1.750	-0.017	-0.631	0.012	0.160	-0.065	-0.865
σ (s.d. of γ)	0.198	20.643	0.168	11.387	0.001	0.096	0.357	15.384
δ_2	-0.084	-3.436	0.024	0.853	0.130	1.781	0.107	1.494
δ_3	0.059	2.040	-0.154	-4.002	-0.099	-1.133	-0.077	-0.871
η			0.022	0.289	0.229	2.060		
	Coeff.	T-val.						
α^{py}	-0.556	-4.720						
		S.e.						
ρ	0.889	0.005						
σ_1	2.030	0.024						
σ_2	0.597	0.031						
Thresholds in Rating Equation								
μ_1	1.5							
μ_2	2.365	0.023						
μ_3	3.272	0.032						
μ_4	4.025	0.037						
μ_5	5.000	0.044						
μ_6	5.958	0.051						
μ_7	7.097	0.058						
μ_8	8.524	0.058						
μ_9	9.5							

Table 5.6: Probabilities of Choosing Described Trajectory rather than Benchmark

	Retirement trajectory				Probability (in %)				
	age part	rr part	age full	rr full	all	age 44-	age 45-54	age 55-64	age 65+
Late 1			70	90	3.01	3.10	1.01	1.36	7.04
Late 2			70	100	6.44	6.85	2.60	3.13	13.85
Late 3			70	110	10.92	11.87	5.13	5.70	21.77
Early 1			62	50	12.75	9.58	18.44	19.36	4.45
Early 2			62	60	25.14	20.84	35.65	33.23	11.22
Early 3			62	70	56.96	53.99	70.20	62.36	40.04
Partial 1	63	85	67	70	50.46	55.26	46.83	46.83	50.43
Partial 2	63	100	67	70	53.44	58.87	48.74	47.03	56.42
Partial 3	63	85	67	80	68.54	73.53	64.08	61.96	72.15
Late partial 1	65	90	70	90	17.05	20.85	10.04	10.41	25.96
Late partial 2	65	100	70	100	27.44	32.60	18.11	17.40	40.33
Early partial	60	75	65	60	60.79	62.71	66.17	65.55	46.05

	male	female	education law	education mid	education high	partner	no partner	house rented	house own
Late 1	3.43	2.30	3.42	2.51	3.13	2.59	4.30	4.51	2.52
Late 2	7.29	4.96	6.77	5.43	6.97	5.67	8.78	9.14	5.54
Late 3	12.26	8.59	10.93	9.36	12.08	9.80	14.34	14.84	9.62
Early 1	11.19	15.47	14.80	14.33	10.26	13.77	9.64	9.07	13.98
Early 2	23.06	28.76	25.73	27.21	23.21	26.83	19.99	19.16	27.12
Early 3	55.44	59.62	52.61	58.28	58.78	59.02	50.70	49.72	59.37
Partial 1	52.26	47.30	48.80	49.42	52.29	50.17	51.31	53.04	49.59
Partial 2	55.83	49.26	50.19	51.63	56.87	52.88	55.14	56.96	52.26
Partial 3	70.91	64.40	65.14	66.75	72.05	67.96	70.29	71.92	67.41
Late partial 1	19.11	13.45	15.85	15.12	19.25	15.90	20.53	21.90	15.43
Late partial 2	30.36	22.35	24.68	24.50	31.39	25.90	32.13	33.79	25.32
Early partial	60.51	61.28	60.22	62.08	60.21	62.11	56.79	57.79	61.79

	income low	income mid	income high	work hrs 15-	work hrs 16-32	work hrs 33+	wave 1	wave 2	wave 3
Late 1	3.50	2.89	2.98	3.81	1.80	3.38	3.19	2.06	4.08
Late 2	6.62	6.21	6.78	6.67	4.41	7.17	6.73	5.11	7.87
Late 3	10.44	10.58	11.85	9.99	8.20	12.03	11.27	9.49	12.42
Early 1	16.19	12.88	10.62	19.00	14.09	11.60	12.79	12.21	13.43
Early 2	25.59	25.60	23.98	24.92	29.99	23.36	24.31	27.75	22.67
Early 3	48.60	57.64	60.20	39.92	66.91	55.05	54.52	66.46	47.28
Partial 1	45.01	51.63	51.13	40.36	48.41	52.26	56.21	51.69	41.66
Partial 2	45.11	54.52	55.87	37.88	52.17	55.52	58.59	56.42	43.03
Partial 3	59.88	69.55	71.29	52.01	67.71	70.57	73.01	71.82	58.58
Late partial 1	13.75	17.32	18.31	11.04	13.86	18.85	20.22	16.61	13.69
Late partial 2	21.20	27.63	30.46	16.23	24.23	29.79	30.64	28.47	22.08
Early partial	57.31	62.47	59.42	53.82	62.30	60.96	66.29	63.31	50.58

Table 5.7: Probability of Choosing Actuarially Neutral Alternative rather than Benchmark Trajectory

age	rr -actuar. neutral	probability (in %)
60	50.26	5.49
61	53.45	10.93
62	56.95	19.83
63	60.80	32.10
64	65.14	44.51
65	70.00	50.00
66	75.00	44.62
67	81.16	35.05
68	87.71	23.80
69	95.17	14.52
70	103.53	7.93

Note: Acturial neutral retirement scenarios taken from Queisser and Whitehouse (2006). They are calculated for the OECD average based on a 2 % interest rate, average life expectancy for OECD countries, and price indexation of pensions.

Table 5.8: Income Effect on Preferred Retirement Age

age	prob. distribution of preferred retirement age (in %)						
	rr	0.7 rr	0.8 rr	0.9 rr	1.1 rr	1.2 rr	1.3 rr
60	0.83	0.94	0.89	0.85	0.81	0.80	0.80
61	2.32	1.87	2.00	2.15	2.54	2.78	3.08
62	5.77	3.83	4.39	5.03	6.62	7.60	8.71
63	11.38	7.12	8.39	9.81	13.10	14.95	16.90
64	17.59	11.50	13.48	15.53	19.59	21.43	23.01
65	20.97	15.78	17.80	19.57	21.90	22.34	22.32
66	17.88	17.04	17.88	18.15	17.10	15.92	14.45
67	12.84	16.49	15.73	14.46	11.02	9.19	7.44
68	6.78	12.71	10.73	8.70	5.10	3.69	2.57
69	2.78	8.26	6.00	4.18	1.77	1.08	0.62
70	0.86	4.47	2.71	1.57	0.44	0.21	0.10
mean age (years)	65.08	65.95	65.63	65.35	64.83	64.61	64.41
difference (months)	0	10.44	6.60	3.24	-3.00	-5.64	-8.04

Note: In this table we change the pension wealth in actuarially neutral trajectories and study its impact on the mean age. In column "rr" we let people choose between all eleven scenarios with full retirement at ages 60 to 70 and replacement rates in Table 5.7. In columns "1.1 rr", "1.2 rr" and "1.3 rr" we increase these replacement rates by 10 %, 20 % and 30 %, respectively. In columns "0.9 rr", "0.8 rr" and "0.7 rr" we decrease these replacement rates as indicated. In each column, we give the probability distributions of preferred retirement age, the mean retirement age measured in years and the difference between this and the mean retirement age and the baseline choice set "rr".

Figure 5.2: Income Effect on the Preferred Retirement Age - Probability distribution of preferred retirement ages for varying replacement rates (cf. Table 5.8)

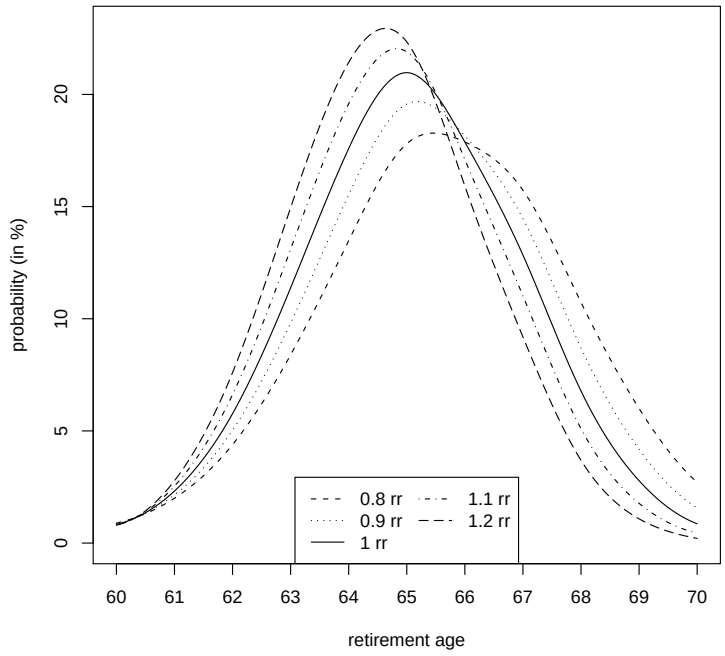


Table 5.9: Income Effects on Mean Preferred Retirement Age by Socioeconomic Group

	retirement age						
	rr	0.7 rr	0.8 rr	0.9 rr	1.1 rr	1.2 rr	1.3 rr
all	65.08	10.44	6.60	3.24	-3.00	-5.64	-8.04
age 44-	65.28	10.80	6.92	3.33	-3.06	-5.88	-8.50
age 45-54	64.52	9.96	6.31	3.01	-2.74	-5.25	-7.57
age 55-64	64.59	8.74	5.55	2.63	-2.41	-4.62	-6.66
age 65+	65.93	11.40	7.40	3.60	-3.39	-6.55	-9.47
male	65.20	10.64	6.82	3.28	-3.03	-5.82	-8.41
female	64.86	9.70	6.18	2.95	-2.71	-5.22	-7.53
education low	65.04	9.20	5.89	2.83	-2.61	-5.02	-7.25
education mid	64.94	9.89	6.31	3.02	-2.77	-5.31	-7.67
education high	65.21	11.30	7.24	3.48	-3.22	-6.19	-8.94
partner	64.98	10.18	6.50	3.12	-2.87	-5.51	-7.95
no partner	65.38	10.65	6.84	3.29	-3.05	-5.88	-8.50
house rented	65.44	10.73	6.90	3.33	-3.10	-5.95	-8.60
house own	64.96	10.15	6.48	3.10	-2.85	-5.48	-7.92
income low	65.01	8.23	5.25	2.52	-2.34	-4.48	-6.48
income mid	65.05	10.30	6.59	3.16	-2.90	-5.58	-8.06
income high	65.18	11.42	7.31	3.51	-3.25	-6.25	-9.03
working hours 15-	64.97	6.33	4.04	1.93	-1.78	-3.40	-4.95
working hours 16-32	64.83	10.98	7.01	3.35	-3.07	-5.91	-8.51
working hours 33+	65.18	10.46	6.69	3.22	-2.97	-5.72	-8.26
wave 1	65.11	10.01	6.41	3.07	-2.83	-5.46	-7.87
wave 2	64.97	11.54	7.37	3.53	-3.25	-6.24	-9.01
wave 3	65.19	8.99	5.76	2.77	-2.56	-4.92	-7.12

Note: Income effects are calculated as in Table 5.8. In column "rr", we let the given group choose between all eleven scenarios with full retirement at ages from 60 to 70 and replacement rates "rr" from Table 5.7. We present the mean preferred retirement age (in years) by group. In other columns we change the replacement rates as indicated and calculate the differences (in months) between the new mean and the mean in the baseline (column "rr"). The differences in the row "all" correspond to the differences in Table 5.8.

Table 5.10: Substitution Effect on the Preferred Retirement Age

age	Distribution of preferred retirement age (in %)						
	rr	70	70+0.33(rr-70)	70+0.5(rr-70)	70+2(rr-70)	70+3(rr-70)	90,70
60	0.83	1.69	0.97	0.85	1.35	2.39	5.09
61	2.32	8.50	4.82	3.80	1.45	0.80	17.25
62	5.77	19.72	12.98	10.44	2.54	1.30	27.48
63	11.38	26.21	21.38	18.52	5.00	2.69	25.63
64	17.59	22.62	23.61	22.69	9.51	5.75	15.34
65	20.97	13.50	18.73	20.26	15.97	11.71	5.98
66	17.88	5.67	10.69	13.08	19.23	17.47	2.39
67	12.84	1.69	4.78	6.78	20.16	23.07	0.68
68	6.78	0.35	1.57	2.61	14.37	19.27	0.14
69	2.78	0.05	0.40	0.78	7.59	11.25	0.02
70	0.86	0.00	0.08	0.18	2.82	4.31	0.00
mean age (years)	65.08	63.34	63.98	64.27	66.07	66.58	62.61
difference (months)	0.00	-20.88	-13.20	-9.72	11.88	18.00	-29.64

Note: In column "rr", we let people choose between eleven actuarially equivalent scenarios with full retirement at ages from 60 to 70 and replacement rates "rr" in Table 5.7. In other columns labeled "70+x(rr-70)", we change replacement rates as indicated (see the text for an example). In column "90,70", the replacement rates are all equal to 90 till age 65 and 70 from age 65. In each column, we give the probability distribution of the preferred retirement age, the mean retirement age (in years) and its difference (in months) with the mean retirement age for the baseline choice set "rr".

Table 5.11: Substitution Effect on the Mean Preferred Retirement Age by Socioeconomic Group

	retirement age						
	rr	70	70+0.33(rr-70)	70+0.5(rr-70)	70+2(rr-70)	70+3(rr-70)	90,70
all	65.08	-20.88	-13.20	-9.72	11.88	18.00	-29.64
age 44-	65.28	-21.86	-13.85	-9.93	12.25	18.30	-31.20
age 45-54	64.52	-20.25	-13.15	-9.54	12.77	19.55	-28.44
age 55-64	64.59	-17.37	-11.23	-8.14	10.94	16.81	-24.48
age 65+	65.93	-23.65	-14.66	-10.37	11.73	17.14	-33.72
male	65.20	-21.67	-13.75	-9.85	12.23	18.33	-30.84
female	64.86	-19.60	-12.55	-9.05	11.65	17.65	-27.60
education low	65.04	-18.46	-11.74	-8.44	10.74	16.25	-26.04
education mid	64.94	-19.90	-12.72	-9.15	11.80	17.88	-28.08
education high	65.21	-23.24	-14.76	-10.58	13.00	19.41	-33.12
partner	64.98	-20.66	-13.19	-9.49	12.06	18.21	-29.28
no partner	65.38	-21.68	-13.67	-9.77	11.88	17.69	-30.84
house rented	65.44	-21.88	-13.77	-9.83	11.87	17.66	-31.08
house own	64.96	-20.59	-13.16	-9.47	12.06	18.22	-29.16
income low	65.01	-16.29	-10.33	-7.43	9.46	14.30	-22.80
income mid	65.05	-20.85	-13.30	-9.55	12.12	18.30	-29.64
income high	65.18	-23.55	-14.96	-10.74	13.21	19.73	-33.60
work hrs 15-	64.97	-12.03	-7.58	-5.41	6.85	10.21	-16.44
work hrs 16-32	64.83	-22.54	-14.49	-10.44	13.39	20.23	-31.92
work hrs 33+	65.18	-21.23	-13.47	-9.66	12.05	18.11	-30.12
wave 1	65.11	-20.16	-12.81	-9.20	11.60	17.48	-28.68
wave 2	64.97	-23.90	-15.28	-11.00	13.76	20.65	-33.96
wave 3	65.19	-17.85	-11.29	-8.09	10.20	15.39	-25.20

Note: Substitution effects are calculated as in Table 5.10. In column "rr", the choices are between all eleven scenarios with full retirement at ages from 60 to 70 and replacement rates "rr" from Table 5.7. We present the mean preferred retirement age (in years) by group. In other columns replacement rates imply positive or negative accruals as indicated, and the differences (in months) between the new mean and the mean in the baseline (column "rr") is presented. The differences in the row "all" correspond to those in Table 5.10.

Table 5.12: Sensitivity Analysis Income Effect on Preferred Retirement Age

		retirement age						
		rr	0.7 rr	0.8 rr	0.9 rr	1.1 rr	1.2 rr	1.3 rr
M0	mean	65.08	65.95	65.63	65.35	64.83	64.61	64.41
	dif	0.00	10.44	6.60	3.24	-3.00	-5.64	-8.04
M1	mean	65.11	65.80	65.56	65.33	64.91	64.72	64.53
	dif	0.00	8.25	5.34	2.61	-2.45	-4.71	-6.99
M2	mean	65.09	65.92	65.62	65.32	64.86	64.63	64.43
	dif	0.00	10.03	6.41	2.83	-2.76	-5.54	-7.96
M3	mean	65.08	65.93	65.62	65.36	64.84	64.64	64.43
	dif	0.00	10.14	6.52	3.31	-2.85	-5.23	-7.81
M4	mean	65.29	65.72	65.55	65.42	65.16	65.07	64.95
	dif	0.00	5.16	3.12	1.50	-1.61	-2.71	-4.11
M5	mean	64.96	66.07	65.67	65.31	64.64	64.34	64.07
	dif	0.00	13.35	8.55	4.17	-3.88	-7.42	-10.74

Note: M0 - benchmark model of Section 5.3; M1 - M0 with term $\alpha^{y^2}(y_{ist})^2$ added to right hand side in eq. 5.2; M2 - M0 but observed characteristics X_{is} are just sex and age; M3 - M0 but observed characteristics X_{is} are just sex, age, education and partner; M4 - M0 with fixed discount factor $\rho = 0.95$; M5 - M0 with fixed discount factor $\rho = 0.85$.

Table 5.13: Sensitivity Analysis Substitution Effect on Preferred Retirement Age

		retirement age						
		rr	70	70+0.33(rr-70)	70+0.5(rr-70)	70+2(rr-70)	70+3(rr-70)	90,70
M0	mean	65.08	63.34	63.98	64.27	66.07	66.58	62.61
	dif	0.00	-20.88	-13.20	-9.72	11.88	18.00	-29.64
M1	mean	65.11	63.38	64.02	64.33	66.14	66.61	62.64
	dif	0.00	-20.73	-13.11	-9.40	12.39	18.05	-29.67
M2	mean	65.09	63.37	64.01	64.31	66.08	66.60	62.66
	dif	0.00	-20.59	-12.88	-9.25	12.00	18.20	-29.14
M3	mean	65.08	63.36	63.97	64.29	66.09	66.59	62.65
	dif	0.00	-20.67	-13.27	-9.46	12.06	18.08	-29.20
M4	mean	65.29	63.71	64.26	64.53	66.28	66.78	63.41
	dif	0.00	-18.96	-12.36	-9.12	11.88	17.88	-22.56
M5	mean	64.96	63.15	63.84	64.15	65.94	66.45	62.10
	dif	0.00	-21.78	-13.47	-9.70	11.74	17.86	-34.29

Note: M0 - benchmark model of Section 5.3; M1 - M0 with term $\alpha^{y^2}(y_{ist})^2$ added to right hand side in eq. 5.2; M2 - M0 but observed characteristics in X_{is} are just sex and age; M3 - M0 but observed characteristics X_{is} are just sex, age, education and partner; M4 - M0 with fixed discount factor $\rho = 0.95$; M5 - M0 with fixed discount factor $\rho = 0.85$.

Bibliography

Andrews, D. W. K. (1988) Chi-square diagnostic tests for econometric models. *Journal of Econometrics*, **37**, 135–156.

Bago d'Uva, T., Lindeboom, M., O'Donnell, O. and Van Doorslaer, E. D. (2009) Slipping anchor? Testing the vignettes approach to identification and correction of reporting heterogeneity. Tinbergen Institute Discussion Papers No. 09-091/3, TI.

Bago d'Uva, T., O'Donnell, O. and Van Doorslaer, E. D. (2008a) Differential health reporting by education level and its impact on the measurement of health inequalities among older Europeans. *International Journal of Epidemiology*, **37**, 1375–1383.

Bago d'Uva, T., Van Doorslaer, E. D., Lindeboom, M., O'Donnell, O. and Chatterji, S. (2008b) Does reporting heterogeneity bias the measurement of health disparities? *Health Economics*, **17**, 351–375.

Barsky, R. B., Juster, F. T., Kimball, M. S. and Shapiro, M. D. (1997) Preference parameters and behavioral heterogeneity: An experimental approach in the health and retirement study. *Quarterly Journal of Economics*, 537–579.

Belloni, M., Monticone, C. and Trucchi, S. (2006) Flexibility in retirement. A framework for the analysis and a survey of European countries. Research report commissioned by the European Commission, CeRP, Turin.

Börsch-Supan, A., Brugiavini, A., Jürges, H., Mackenbach, J., Siegrist, J. and Weber, G. (2005) *Health, Ageing and Retirement in Europe - First Results from the Survey of Health, Ageing and Retirement in Europe.* Mannheim Research Institute for the Economics of Aging (MEA).

Börsch-Supan, A. and Jürges, H. (2005) *The Survey of Health, Aging, and Retirement in Europe - Methodology.* Mannheim Research Institute for the Economics of Aging (MEA).

Bovenberg, A. L. and Gradus, R. H. J. M. (2008) Dutch policies towards ageing. *European View*, **7**, 265–275.

Brown, J. R., Coile, C. C. and Weisbenner, S. J. (2006) The effect of inheritance receipt on retirement. NBER working paper 12386, NBER, Cambridge MA.

Bruinshoofd, W. A. and Grob, S. (2005) Arbeidsparticipatie van ouderen: Microfinanciële motivaties en beleidsaspecten. DNB Occasional Studies, vol. 3, no. 1, De Nederlandsche Bank, Amsterdam.

Buckley, J. (2008) Survey context effects in anchoring vignettes. URL `http://polmeth.wustl.edu/media/Paper/surveyartifacts.pdf`.

Capretta, J. C. (2007) Global aging and the sustainability of public pension systems - an assessment of reform efforts in twelve developed countries. Tech. rep., Center for Strategic & International Studies, Washington.

Chesher, A. and Irish, M. (1987) Residual analysis in the grouped and censored normal linear model. *Journal of Econometrics*, **34**, 33–61.

Corrado, L. and Weeks, M. (2010) Identification strategies in survey response using vignettes. Cambridge Working Papers in Economics 1031, University of Cambridge.

Datta Gupta, N., Kristensen, N. and Pozzoli, D. (2009) External validation of the use of vignettes in cross-country health studies. IZA Discussion paper No. 3989.

Dourgnon, P. and Lardjane, S. (2007) Les comparaisons internationales d'état de santé subjectif sont-elles pertinentes? une évaluation par la méthode des vignettes-étalons. *Economie et Statistique*, **403-404**, 165–177.

Euwals, R., Van Vuuren, D. and Wolthoff, R. (2007) Early retirement behaviour in the Netherlands. Netpsar Discussion Paper 2007-013.

Fouarge, D., Kerkhofs, M. and Ester, P. (2008) The effect of financial incentives on pension age: Results from a stated preferences experiment. Paper presented at the EALE conference.

Gourieroux, C. and Monfort, A. (1991) Simulation based inference in models with heterogeneity. *Annales d'Economie et de Statistique*, **20/21**, 69–107.

Gourieroux, C. and Monfort, A. (1996) *Simulation-Based Econometric Methods*. Oxford: Oxford University Press.

Gruber, J. and Wise, D. (1998) Social security and retirement: An international comparison. *American Economic Review*, **88**, 158–163.

Gruber, J. and Wise, D. (2004) Introduction and summary. In *Social Security Programs and Retirement around the World: Micro-Estimation* (eds. J. Gruber and D. Wise), 1–40. National Bureau of Economic research, Chicago.

Herzog, A. R. and Wallace, R. B. (1997) Measures of cognitive functioning in the AHEAD study. *Journal of Gerontology Series B*, **52B**, 37–48.

Heyma, A. (2004) A structural dynamic analysis of retirement behaviour in the Netherlands. *Journal of Applied Econometrics*, **19**, 739–759.

Holland, P. W. and Wainer, H. (1993) *Differential Item Functioning*. Hillsdale, N. J.: Lawrence Erlbaum.

Hopkins, D. J. and King, G. (2010) Improving anchoring vignettes: Designing surveys to correct for interpersonal incomparability. *Public Opinion Quarterly*, **74**, 201–222.

Kakes, J. and Broeders, D. (2006) The sustainability of the Dutch pension system. DNB Occational Studies 06, Nederlands Central Bank.

Kantarci, T. and Van Soest, A. (2008) Gradual retirement: Preferences and limitations. *De Economist*, **156**, 113–144.

Kapteyn, A. and De Vos, K. (1998) Social security and labor force participation in the Netherlands. *American Economic Review*, **88**, 164–167.

Kapteyn, A. and De Vos, K. (2004) Incentives and exit routes to retirement in the Netherlands. In *Social Security Programs and Retirement around the World: Micro-Estimation* (eds. J. Gruber and D. Wise), 461–498. National Bureau of Economic research, Chicago.

Kapteyn, A., Smith, J. P. and Van Soest, A. (2007) Vignettes and self-reports of work disability in the United States and the Netherlands. *American Economic Review*, **97**, 461–473.

Kapteyn, A., Smith, J. P. and Van Soest, A. (2010) Life satisfaction. In *International Differences in Well-Being* (eds. E. Diener, J. F. Helliwell and D. Kahneman), 70–104. Oxford: Oxford University Press.

King, G., Murray, C. J. L., Salomon, J. A. and Tandon, A. (2004) Enhancing the validity and cross-cultural comparability of measurement in survey research. *American Political Science Review*, **98**, 191–207.

King, G. and Wand, J. (2007) Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis*, **15**, 46–66.

Kristensen, N. and Johansson, E. (2008) New evidence on cross-country differences in job satisfaction using anchoring vignettes. *Labour Economics*, **15**, 96–117.

Lardjane, S. and Dourgnon, P. (2007) Les comparaisons internationales d'état de santé subjectif sont-elles pertinentes? une évaluation par la méthode des vignettes-étalons. *Economie et Statistique*, **403-404**, 165–177.

Llewellyn, D. J., Lang, I. A., Langa, K. M. and Huppert, F. A. (2008) Cognitive function and psychological well-being: Findings from a population-based cohort. *Age and Ageing*, **37**, 685–689.

Louviere, J., Hensher, D. and Swait, J. (2002) *Stated Choice Methods*. Cambridge: Cambridge University Press.

Lumsdaine, R. and Mitchell, O. (1999) New developments in the economic analysis of retirement. In *Handbook of Labor Economics, Vol. 3C* (eds. O. Ashenfelter and D. Card), 3261–3307. North-Holland, Amsterdam.

Mastrogiacomo, M., Alessie, R. and Lindeboom, M. (2004) Retirement behaviour of Dutch elderly households. *Journal of Applied Econometrics*, **19**, 777–793.

Mehta, K. M., Yaffe, K., Langa, K. M., Sands, L., Whooley, M. A. and Covinsky, K. E. (2003) Additive effects of cognitive function and depressive symptoms on mortality in elderly community-living adults. *Journal of Gerontology Series A*, **58A**, 461–467.

Murray, C. J. L., Özaltin, E., Tandon, A., Salomon, J. A., Sadana, R. and Chatterji, S. (2003) *Empirical Evaluation of the Anchoring Vignette Approach in Health Surveys*, chap. 30, 369 – 399. World Health Organization.

Nelissen, J. (2001) Het effect van wijzigingen in vervroegde uittredingsregelingen op de arbeidsparticipatie van oudere werknemers. Ministry of Social Affairs and Employment, The Hague.

Nunn, A. J. and Gregg, I. (1989) New regression equations for predicting peak expiratory flow in adults. *British Medical Journal*, **298**, 1068–1070.

Peracchi, C. and Rossetti, C. (2010) The heterogeneous thresholds ordered response model: Identification and inference. Mimeo, Tor Vergata University.

Quanjer, P. H., Lebowitz, M. D., Gregg, I., Miller, M. R. and Pedersen, O. F. (1997) Peak expiratory flow: Conclusions and recommendations of a working party of the European respiratory society. *European Respiratory Journal*, **24**, 2S–8S.

Queisser, M. and Whitehouse, E. (2006) Neutral or fair? Acturial concepts and pension-system design. OECD Social, employment and migration working paper No. 40, OECD.

Revelt, D. and Train, K. (1998) Mixed logit with repeated choices: Household choices of appliance efficiency level. *Review of Economics and Statistics*, **80**, 647–657.

Rice, N., Robone, S. and Smith, P. C. (2010) International comparison of public sector performance: The use of anchoring vignettes to adjust self-reported data. *Evaluation*, **16**, 81–101.

Salomon, J. A., Tandom, A. and Murray, C. J. L. (2004) Comparability of self rated health: Cross sectional multi-country survey using anchoring vignettes. *British Medical Journal*, **328**, 258–260.

Siegel, S. and Castellan Jr., N. J. (1988) *Nonparametric Statistics for the Behavioral Sciences*. New York: McGraw-Hill.

Sirven, N., Santos-Eggimann, B. and Spagnoli, J. (2008) Comparability of health care responsiveness in Europe using anchoring vignettes from SHARE. IRDES working paper DT No. 15, IRDES, Paris.

Train, K. E. (2003) *Discrete Choice Methods with Simulation*. Cambridge: Cambridge University Press.

Van Dalen, H., Henkens, K. and Hershey, D. A. (2008) Are pension savings sufficient? Perceptions and expectations of American and Dutch workers. CentER discussion paper No. 2008-58, Tilburg.

Van Soest, A., Delaney, L., Harmon, C. P., Kapteyn, A. and Smith, J. P. (2011) Validating the use of vignettes for subjective threshold scales. *Journal of the Royal Statistical Society A*, Forthcoming.

Van Soest, A., Kapteyn, A. and Zissimopoulos, J. (2006) Using stated preferences data to analyze preferences for full and partial retirement. DNB working paper 081, Netherlands Central Bank, Research Department.

Van Solinge, H. and Henkens, K. (2007) Involuntary retirement: The role of restrictive circumstances, timing, social embeddedness and control. *Journal of Gerontology: Social Sciences*, **628**, S295–S305.

Nederlandse samenvatting (Dutch summary)

In dit proefschrift stellen wij empirische analyses voor, waarin subjectieve data worden gebruikt. Hoofdstuken 2, 3 en 4 richten zich op de anchoring vignette methode. In Hoofdstuk 5 gebruiken wij de *stated preference* methode om de invloed van financiële prikkels op de keuze van pensioenleeftijd te bestuderen.

De anchoring vignette methode wordt relatief vaak door onderzoekers gebruikt als ze met subjectieve maten werken. Een voorbeeld van zo'n maat zijn de antwoorden van respondenten op de volgende vraag: "Hoeveel moeite had u de laatste dertig dagen om u te concentreren en dingen te herinneren" op de volgende 5-puntsschaal: geen, mild, matig, ernstig en extreem. Antwoorden op dergelijke vragen kunnen afhankelijk zijn van zowel de objectieve situatie als van het responsgedrag van de respondent. Responsgedrag kan variëren tussen landen en culturen. Om de objectieve situatie en het responsgedrag uit elkaar te houden kunnen wij de anchoring vignettes gebruiken. Vignettes zijn korte beschrijvingen van gezondheid, vermogen om te werken enz.(of, zoals in ons voorbeeld, concentratie) van een hypothetisch persoon. Respondenten worden dan gevraagd om deze hypothetische persoon te beoordelen op dezelfde schaal als waarop ze zichzelf beoordelen. Met gebruik van de vignette beoordelingen kunnen wij het responsgedrag van respondent identificeren en dan deze informatie gebruiken om zelfbeoordelingen aan te passen.

In dit proefschrift valideren wij de anchoring vignette methode. In Hoofdstuk 2 analyseren wij verschillen tussen zelfbeoordelingen die door verschillende vignettes binnen een gegeven gebied aangepast worden. Onderzoekers gebruiken verschillende beschrijvingen van een hypothetisch persoon om de aanpassingen van zelfbeoordelingen te maken, maar er werd nooit bestudeerd of de verschillende vignettes op dezelfde manier werken. Wij vergelijken zelfbeoordelingen, die aangepast zijn met verschillende vignettes, met een objectieve maat op het gebied. Dat laat niet alleen maar zien of we overeenkomstige aangepaste zelfbeoordelingen krijgen maar ook of deze aangepaste zelfbeoordelingen de objectieve

maat goed benaderen. Wij gebruiken data uit de Survey on Health, Ageing and Retirement in Europe (kortweg SHARE) uit zowel 2004 als 2007 voor drie gezondheidsdomeinen - mobiliteit, ademen en concentreren. Wat betreft concentreren, zijn de angepaste zelfbeoordelingen voor de drie vignettes niet helemaal vergelijkbaar - de correlatie tussen een objectieve maat voor concentratie en de zelfbeoordelingen aangepast met 'het beste' vignette is 0.83, terwijl de andere twee vignettes tot correlaties van 0.74 en 0.52 leiden. Als we geen aanpassing gebruiken, is de correlatie 0.75. Dat betekent dat slechts een van de vignettes ons helpt om dichter bij de objectieve maat te komen. Ook bij ademen blijkt de keuze van de vignette uit te maken. Maar de correlatie tussen de objectieve maat voor ademen en de aangepaste zelfbeoordelingen is hoog. Het beste resultaat hebben wij voor mobiliteit gekregen, waar alle vignettes tot vergelijkbare aanpassingen leiden, die dichtbij de objectieve maat liggen.

Met hulp van deze resultaten krijgen wij een intuïtie of de anchoring vignette methode over het algemeen geldig is. Wij laten zien dat de methode niet altijd helpt en daarom stellen wij de volgende vraag: Zijn de twee onderliggende veronderstellingen voor de methode geldig? Met de twee onderliggende veronderstellingen bedoelen wij response consistency en vignette equivalence. Response consistency is de aanname dat respondenten dezelfde schalen gebruiken voor beoordelingen van zichzelf en hypothetische vignette personen. Vignette equivalence betekent dat respondenten het concept dat een vignette beschrijft allemaal op dezelfde manier interpreteren. En er zijn ook andere vragen: Is het parametrische anchoring vignette model dat door onderzoekers bijna uitsluitend gebruikt wordt correct ? Zijn de statistische aannames van het model juist ?

In Hoofdstuk 3 toetsen wij de response consistency aanname en bediscussiëren wij de vignette equivalence aanname. Respondenten in een Internet panel werden gevraagd om hun gezondheid in verschillende domeinen te beschrijven en hun eigen gezondheid te beoordelen. Enkele maanden later was er een ander internet interview. Respondenten worden dan gevraagd om vignettes te beoordelen die gebaseerd zijn op beschrijvingen van hun eigen gezondheid. Als de response consistency aanname geldig is zou er geen systematisch verschil tussen de zelfbeoordelingen uit het eerste interview en beoordelingen van vignettes (replica vignettes) uit het tweede interview zijn. Het resultaat van onze niet-parametrische analyse is dat response consistency voor het slaap domein geldig is. Voor andere domeinen is dat niet geldig. Wij gebruiken dan een parametrische methode om meer informatie te krijgen.

In Hoofdstuk 4 toetsen wij of het parametrische model van anchoring vignettes correct is. Wij gebruiken een Chi-kwadraattoets voor parametrische

modellen die door Andrews (1988) geïntroduceerd werd. Cellen voor de toets worden op een niet-parametrische manier gemaakt. Wij gebruiken namelijk de niet-parametrische anchoring vignette methode om deze cellen te definiëren. De metode vereist geen verklarende variabelen en heeft geen veronderstellingen over de verdeling nodig. Het is dus mogelijk om enkele diagnostische toetsen uit te voeren voor het parametrische model. Als het verworpen wordt, kunnen wij nog de niet-parametrische methode gebruiken. Voor deze analyse gebruiken wij data uit SHARE 2004, namelijk zelfbeoordelingen en beoordelingen van drie vignettes in zes gezondheidsdomeinen (ademen, concentratie, depressie, mobiliteit, pijn en slapen) en enige achtergrondvariabelen van respondenten. Wij laten zien dat de random effect in de drempels een belangrijke rol speelt. Als we geen random effect gebruiken, verwerpen wij de nulhypothese dat het parametrische model correct is voor alle gezondheidsdomeinen. Met gebruik van de random effect verwerpen wij de nulhypothese voor concentratie en pijn niet.

In het laatste hoofdstuk 5 gebruiken wij stated preference methode om voorkeuren van Nederlanders voor vervroegde, verlate en geleidelijke pensionering te analyseren. In een internet experiment beoordelen respondenten verschillende hypothetische scenario's die verschillende pensioenleeftijd en pensioenuitkering beschrijven. De data werden in jaren 2006, 2007 en 2008 verzamelt. Ons experiment laat zien dat er een groot effect van financiële prikkels op de gekozen uittredingsleeftijd is. Geleidelijke pensionering na de gewone uittredingsleeftijd (65 jaar) stimuleert mensen om langer te werken. Onze simulaties met keuzes tussen actuarieel neutrale uittredingsregelingen met leeftijden tussen de 60 en 70 jaar lieten zien dat het verhogen van het pensioeninkomen met 10 procent tot het verlagen van de gemiddelde uittredingsleeftijd met drie manden zou leiden ("het inkomenseffect"). Als we het pensioeninkomen voor verlate pensionering veranderen naar 50 procent van wat actuarieel eerlijk zou zijn, verlaagt dat de uittredingsleeftijd met 9.7 maanden ("substitutie-effect").