WORKING
PAPER

Anchoring Vignettes and Response Consistency

ARIE KAPTEYN, JAMES P. SMITH,
ARTHUR VAN SOEST, AND HANA VOŇKOVÁ

WR-840

February 2011

This paper series made possible by the NIA funded RAND Center for the Study of Aging (P30AG012815) and the NICHD funded RAND Population Research Center (R24HD050906).

LABOR AND POPULATION

Anchoring Vignettes and Response Consistency

Arie Kapteyn, James P. Smith, RAND
Arthur van Soest, Netspar, Tilburg University
Hana Voňková, Charles University in Prague

February 16, 2011

Abstract

The use of anchoring vignettes to correct for differential item functioning rests upon two identifying assumptions: vignette equivalence and response consistency. To test the second assumption we conduct an experiment in which respondents in an Internet panel are asked to both describe their health in a number of domains and rate their health in these 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 with regard to the validity of the experiment, there should be no systematic differences between the evaluation of these vignettes in the second interview and the self-evaluations in the first interview. We analyze data for five health domains: sleep, mobility, concentration, breathing and affect. Although descriptively the vignettes and the self-evaluations are similar for a number of domains, our nonparametric analysis suggests that response consistency is satisfied for the domain of sleep, while it indicates rejection of either the auxiliary assumptions or response consistency for the other domains of health. Parametric analysis suggests that the auxiliary assumptions may be most problematic. The analysis points at the need for a systematic experimental approach to the design of anchoring vignettes before using them in practice.

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". The distributions of the answers to a question like this vary much more across countries than can plausibly be explained by genuine health 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 he/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. (2008), 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. We will discuss their approach more fully at the end of Section 5.

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 (which we will call a replica vignette).

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 analyzed in this paper are sleep, mobility, memory and concentration, feeling down or depressed and breathing. 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.

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[1]. 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 restrictions implied by response consistency and other, auxiliary, assumptions.

This paper is divided into six sections. Section 2 outlines the theory underlying vignettes. Section 3 describes the data that we use and the actual construction of vignettes in our experiments. Section 4 contains descriptive statistics and the results of the nonparametric tests. In Section 5 we present the results for parametric models, giving insight in why the nonparametric tests lead to rejection in most cases. Section 6 concludes.

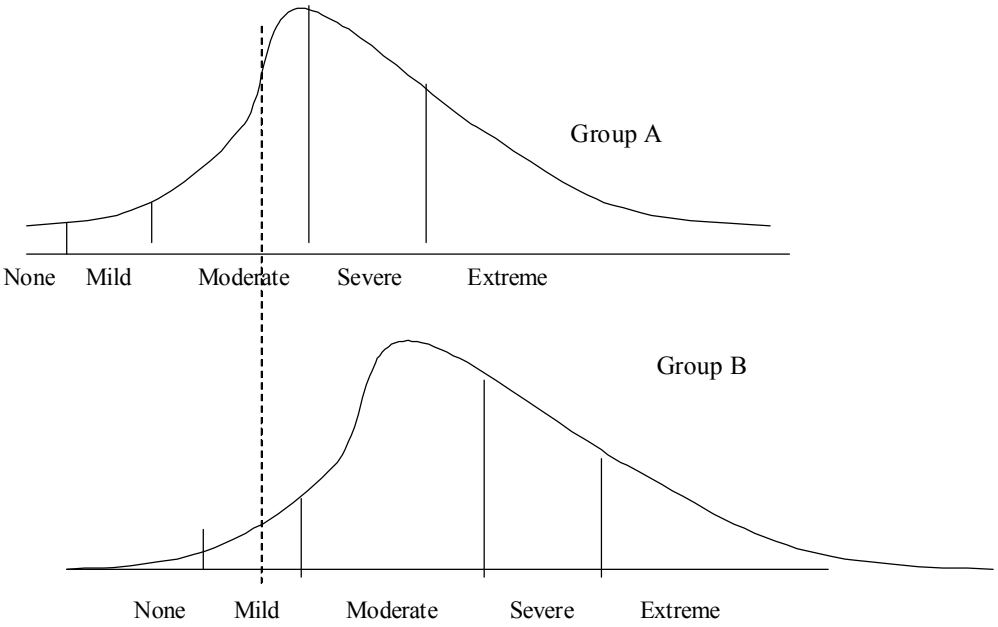
2 Anchoring Vignettes

In this section, we provide an intuitive description of the use of vignettes for identifying response scale differences and discuss the identifying assumptions behind this approach. Suppose one wants to characterize health problems in a given domain (such as the extent to which someone has problems with moving around) of two groups of individuals. Figure 1 presents the densities of the true but unobserved continuous health problems in both groups. The fact that the density of group A is to the left of that in group B, implies that on average, people in group B have more of a health problem in this domain than in people in group A.

The figure also shows that people in these two groups use very different response scales if asked to evaluate their health problem on a five-point scale (None, Mild, Moderate, Severe, Extreme). In this example, people in group B much less easily call something a serious health problem than those in

[1]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.

Figure 1: Comparing self-reported health problem in two groups in case of DIF



group A. For example, someone whose true health problem is given by the vertical dashed line has a moderate problem according to respondents in group A, but only a mild problem according to the group B respondents. The frequency distribution of self-reports in the two groups would suggest that people in group A have more problems than those in group B – the opposite of the true health distribution. Correcting for the differences in the response scales (DIF, "differential item functioning," in the terminology of King et al. (2004)) is essential to compare the actual health in the two groups.

Vignettes can be used to do the correction. The hypothetical persons described in identical vignettes given to both groups have the same health problems by construction. For example, respondents can be asked to evaluate a vignette person's health problem given by the dashed line. In group A, this will be evaluated as "moderate" and in group B as "mild". Since the actual health problem of the vignette person is the same for both groups, the difference in the evaluations by the two groups must be due to DIF. Vignette evaluations thus help to identify differences between the response scales. Using the scales in one of the two groups as the benchmark, the distribution of evaluations in the other group can be adjusted by evaluating

them on the benchmark scale. The corrected distribution of the evaluations can then be compared since they are now on the same scale.

The underlying assumptions necessary to make this adjustment are vignette equivalence – the condition represented by the dashed line is interpreted the same in both countries – and response consistency – a given respondent uses the same scale for the self-reports and the vignette evaluations.

3 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. About 10% of the panel members use a Web TV.[2]

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. The other vignettes in a given domain are constructed in such a way that they always describe a situation that is dif-

[2]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.

ferent 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 this paper, we analyze the data from five health domains: sleep, mobility, concentration and memory, breathing and affect (depression and mood swings). Here we present two examples: sleep and concentration. Details about the other three domains are presented in Appendix B.

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 4.

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 the current paper, 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.[3] Details on the other three domains (mobility, breathing, and affect) are provided in Appendix B.

4 Descriptive Statistics and Nonparametric Tests

Table 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, more serious than the respondents' own problems.

Table 2 displays distributions of responses to the replica vignette question (the rows) for given levels of self-assessments (the columns). 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 a finite scale.

One way to gauge how much responses may change over time due to

[3]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.

idiosyncratic reporting errors is to also consider the distribution of self assessments in wave 2. Table 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 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-values 0.23 and 0.26), but is clearly rejected for the other four domains (p-values 0.00, except for the sign test for mobility which yields a p-value equal to 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[4]. 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 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.[5] 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. 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 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 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

[4]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 $q_i = N[\{N^{-1}(p_i, 1)/\sigma\}, 1]$

[5]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).

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. Concentration and memory may also confound two distinct concepts.

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).

5 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)).

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:[6]

$$\begin{aligned} Y_{si}^{\star} &= \beta_s T_{si} + \delta_s X_i + \epsilon_{si} && (1)\\ Y_{si} = j &\Leftrightarrow \tau_{si}^{j-1} < Y_{si}^{\star} \leq \tau_{si}^{j} \quad j = 1,2,3,4 && (2)\\ \tau_{si}^{0} &= -\infty && (3)\\ \tau_{si}^{1} &= \gamma_s^1 X_i + \lambda_s^1 T_{si} + u_i && (4)\\ \tau_{si}^{2} &= \tau_{si}^{1} + \exp\left(\gamma_s^2 X_i + \lambda_s^2 T_{si}\right) && (5)\\ \tau_{si}^{3} &= \tau_{si}^{2} + \exp\left(\gamma_s^3 X_i + \lambda_s^3 T_{si}\right) && (6)\\ \tau_{si}^{4} &= \infty, && (7) \end{aligned}$$

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 (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)-(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.

Replica Vignette Evaluations

The evaluation of the replica vignette is modeled using an ordered response equations similar to (1)-(7):

[6] 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.

$$
\begin{aligned}
Y^{\star}_{vi} &= \beta_v T_{si} + \delta_v X_i + \epsilon_{vi} && (8)\\
Y_{vi} = j &\Leftrightarrow \tau^{j-1}_{vi} < Y^{\star}_{vi} \leq \tau^{j}_{vi} \quad j = 1,2,3,4 && (9)\\
\tau^{0}_{vi} &= -\infty && (10)\\
\tau^{1}_{vi} &= \gamma^{1}_{v} X_i + \lambda^{1}_{v} T_{si} + u_i && (11)\\
\tau^{2}_{vi} &= \tau^{1}_{vi} + \exp\left(\gamma^{2}_{v} X_i + \lambda^{2}_{v} T_{si}\right) && (12)\\
\tau^{3}_{vi} &= \tau^{2}_{vi} + \exp\left(\gamma^{3}_{v} X_i + \lambda^{3}_{v} T_{si}\right) && (13)\\
\tau^{4}_{vi} &= \infty, && (14)
\end{aligned}
$$

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:

$$\delta_v = 0 \tag{15}$$

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:

$$\beta_s = \beta_v, \delta_s = 0 \tag{16}$$

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*:

$$\text{RC: } \gamma^{j}_{s} = \gamma^{j}_{v}, \lambda^{j}_{s} = \lambda^{j}_{v},\ j = 1,2,3 \tag{17}$$

Without imposing either (15) or (16), we cannot test (17), since the parameters in (17) are not identified. The reason is that in this unrestricted model we can identify $\lambda^{1}_{s} - \beta_s$, $\lambda^{1}_{v} - \beta_v$, $\gamma^{1}_{s} - \delta_s$ and $\gamma^{1}_{v} - \delta_v$, but not the individual parameters λ^{1}_{s}, β_s, λ^{1}_{v}, β_v, γ^{1}_{s}, δ_s, γ^{1}_{v} 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 A.

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:

$$\text{RC1} \quad : \quad \lambda_s^1 - \beta_s = \lambda_v^1 - \beta_v,\ \gamma_s^1 - \delta_s = \gamma_v^1 - \delta_v \tag{18}$$

$$\text{RC2} \quad : \quad \gamma_s^2 = \gamma_v^2,\ \lambda_s^2 = \lambda_v^2 \tag{19}$$

$$\text{RC3} \quad : \quad \gamma_s^3 = \gamma_v^3,\ \lambda_s^3 = \lambda_v^3 \tag{20}$$

Under the maintained additional assumptions (15) and (16), we have $\beta_s = \beta_v$ and $\delta_s = \delta_v (= 0)$, so that (18) is equivalent to $\lambda_s^1 = \lambda_v^1$ and $\gamma_s^1 = \gamma_v^1$ and (18), (19) and (20) together are equivalent to the response consistency assumption (17) we want to test.

We will test (18), (19) and (20) jointly, but will also test (19) and (20) jointly, without imposing (18). The discussion above implies that the first test requires the maintained assumptions (15) and (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 (19) and (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 (18).

Test Results

Table 5 presents log likelihoods and likelihood ratio tests for restricted and unrestricted versions of the model (1)-(7) and (8)-(14).[7] We present tests of three hypotheses: (1) all equalities in (18)-(20) hold (denoted by $\forall j$); (2) equalities (19)-(20) hold (denoted by $j > 1$); (3) equation (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 test (third line).

Table 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,

[7]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.

the joint hypotheses (18)-(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.[8]

6 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 (11)-(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 B, 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.

Descriptively the vignettes and the self-evaluations are similar for sleep, mobility, and affect, suggesting that even now these vignettes will do a good job in separating out differences in subjective thresholds. In the domains of concentration and breathing, in particular, more conceptual and experimental work is needed in designing the vignette descriptions so that self-

[8]It is of interest to compare our approach with the approach adopted by Bago d'Uva et al. (2009). 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 (1) provides a complete description of the respondent's own health (implying $\delta_s = 0$) and that $\lambda_s^j = 0$ for all j. 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). They 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. The authors also perform a test of vignette equivalence. Their idea is that the difference in evaluations of different vignettes should not vary systematically across individuals. They assume that respondents use the same thresholds for different vignettes. Vignette equivalence gets rejected for both cognition and mobility.

assessments and replica vignettes are more similar.

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.

Our results thus 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 unlikely 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 paper, to determine whether one has been successful.

References

Bago d'Uva, T., M. Lindeboom, O. O'Donnell, and E. D. Van Doorslaer (2009). Slipping Anchor? Testing the Vignettes Approach to Identification and Correction of Reporting Heterogeneity. *Tinbergen Institute Discussion Paper* No. TI 2009-091/3.

Bago d'Uva, T., E. D. Van Doorslaer, M. Lindeboom, O. O'Donnell, and S. Chatterji (2008). Does Reporting Heterogeneity Bias the Measurement of Health Disparities? *Health Economics*, **17**(3), 351–375.

Hopkins, D. J. and G. King (2010). Improving Anchoring Vignettes. Designing Surveys to Correct for Interpersonal Incomparability. *Public Opinion Quarterly*, 1–22.

Kapteyn, A., J. P. Smith, and A. Van Soest (2007). Vignettes and Self-reports of Work Disability in the United States and the Netherlands. *American Economic Review*, **97**(1), 461–473.

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

King, G., C. J. L. Murray, J. A. Salomon, and A. Tandon (2004). Enhancing the Validity and Cross-cultural Comparability of Measurement in Survey Research. *American Political Science Review*, **98**(1), 567–583.

Kristensen, N. and E. Johansson (2008). New Evidence on Cross-country Differences in Job Satisfaction Using Anchoring Vignettes. *Labour Economics*, **15**, 96–117.

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

Salomon, J. A., A. Tandon, and C. J. L. Murray (2004). Comparability of Self Rated Health: Cross Sectional Multi-country Survey Using Anchoring Vignettes. *British Medical Journal*, **328**(7434), 258–260.

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

Sirven, N., B. Santos-Eggimann, and J. Spagnoli (2008). Comparability of Health Care Responsiveness in Europe Using Anchoring vignettes from SHARE. *IRDES working paper* DT No. 15, IRDES, Paris.

Van Soest, A., L. Delaney, C. P. Harmon, A. Kapteyn, and J. P. Smith (2011). Validating the Use of Vignettes for Subjective Threshold Scales. *Journal of the Royal Statistical Society A*, Forthcoming.

Table 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 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: This table presents distributions of responses to the replica vignette question for given levels of self-assessments. The sum of relative frequencies (in %) in each column is equal to 100. The diagonals in each panel represent cases in which responses for self-assessments and replica vignettes are identical

Table 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

Note: Self-assessment questions were asked in both waves of our experiment. Replica vignettes were collected only in second wave. This table summarizes the correnspondence between these three measures using correlation coefficients.

Table 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 around 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 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 (18)-(20) hold (denoted by $\forall j$); (2) equalities (19)-(20) hold (denoted by $j > 1$); (3) equation (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).

A Identification

When conditions (15) and(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

$$\Pr(Y_{si} = 1) = \Pr\left[\epsilon_{si} - u_i \leq \Omega_s^1\right] \tag{21}$$

$$\Pr(Y_{si} = 2) = \Pr\left[\Omega_s^1 < \epsilon_{si} - u_i \leq \Omega_s^1 + \exp\left(\gamma_s^2 X_{si} + \lambda_s^2 T_{si}\right)\right] \tag{22}$$

$$\Pr(Y_{si} = 3) = \Pr\left[\begin{array}{c} \Omega_s^1 + \exp\left(\gamma_s^2 X_i + \lambda_s^2 T_i\right) < \epsilon_{si} - u_i \leq \\ \Omega_s^1 + \exp\left(\gamma_s^2 X_i + \lambda_s^2 T_i\right) + \exp\left(\gamma_s^3 X_i + \lambda_s^3 T_{si}\right) \end{array}\right] \tag{23}$$

$$\Pr(Y_{si} = 4) = \Pr\left[\epsilon_{si} - u_i > \Omega_s^1 + \exp\left(\gamma_s^2 X_i + \lambda_s^2 T_i\right) + \exp\left(\gamma_s^3 X_i + \lambda_s^3 T_{si}\right)\right] \tag{24}$$

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

$$\Pr(Y_{vi} = 1) = \Pr\left[\epsilon_{vi} - u_i \leq \Omega_v^1\right] \tag{25}$$

$$\Pr(Y_{vi} = 2) = \Pr\left[\Omega_v^1 < \epsilon_{vi} - u_i \leq \Omega_v^1 + \exp\left(\gamma_v^2 X_i^2 + \lambda_v^2 T_i\right)\right] \tag{26}$$

$$\Pr(Y_{vi} = 3) = \Pr\left[\begin{array}{c} \Omega_v^1 + \exp\left(\gamma_v^2 X_i^2 + \lambda_v^2 T_i\right) < \epsilon_{vi} - u_i \leq \\ \Omega_v^1 + \exp\left(\gamma_v^2 X_i^2 + \lambda_v^2 T_i\right) + \exp\left(\gamma_v^3 X_i^2 + \lambda_v^3 T_i\right) \end{array}\right] \tag{27}$$

$$\Pr(Y_{vi} = 4) = \Pr\left[\epsilon_{vi} - u_i > \Omega_v^1 + \exp\left(\gamma_v^2 X_i^2 + \lambda_v^2 T_i\right) + \exp\left(\gamma_v^3 X_i^2 + \lambda_v^3 T_i\right)\right] \tag{28}$$

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.

B 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?