Tilburg University

# Measuring Subjective Survival Expectations

de Bresser, Jochem

*Publication date:*
2019

Link to publication in Tilburg University Research Portal

*Citation for published version (APA):*
de Bresser, J. (2019). *Measuring Subjective Survival Expectations: Do Response Scales Matter?*.

**General rights**
Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

- Users may download and print one copy of any publication from the public portal for the purpose of private study or research.
- You may not further distribute the material or use it for any profit-making activity or commercial gain
- You may freely distribute the URL identifying the publication in the public portal

**Take down policy**
If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

Download date: 06. Oct. 2022

# Measuring Subjective Survival Expectations – Do Response Scales Matter?*

Jochem de Bresser
Tilburg University†
Netspar‡

March 18, 2019

**Abstract**

This paper analyzes the test-retest reliability of subjective survival expectations that are elicited on two widely used response scales: an 11-point scale from 0 to 10 and a full percentage scale from 0 to 100. We compare responses of the same individuals in two surveys fielded in the same month. Reliability is evaluated both at the level of reported probabilities and through a model that relates expectations to socio-demographic variables. Test-retest correlations of survival probabilities are between 0.5 and 0.7, similar to subjective well-being. Only 20% of probabilities are equal across surveys, but up to 61-77% are consistent once we account for rounding. Both scales perform similarly in terms on response rates, internal consistency and fifty-fifty answers. Models that analyze all probabilities jointly reveal similar associations between most covariates and the hazard of death in test and retest datasets. Moreover, expectations are persistent at the level of the individual and this unobserved heterogeneity is strongly correlated across surveys ($r \approx$ 0.8-0.9). Finally, we use a calibrated life cycle model to map survival expectations into wealth and labor supply. Though wealth accumulation is sensitive to expectations, correcting for rounding substantially improves reliability of simulated wealth profiles. Taken together this evidence suggests that the two elicitation scales yield reliable measures of expectations.

**Key words:** Subjective expectations, test-retest reliability, life cycle model, rounding

**JEL-codes:** D84; J14; C34

*This work is part of the research programme Innovational Research Incentives Scheme Veni with project number 451-15-018, which is (partly) financed by the Netherlands Organisation for Scientific Research (NWO). I gratefully acknowledge Netspar for financially supporting data collection. I thank Michael Hurd, Bas Donkers, an anonymous referee and participants at the 2016 IAAE conference, the 2016 Netspar Pension Day and the 2017 Netspar International Pension Workshop for their valuable feedback. All errors are my own.

†Tilburg University, P.O. Box 90153, 5000 LE Tilburg the Netherlands. Email: j.r.debresser@uvt.nl.

‡Netspar, P.O. Box 90153, 5000 LE Tilburg, The Netherlands

# 1 Introduction

Expectations play an important role in economic models of inter-temporal decision making, such as life cycle models of labor supply and saving (e.g. French, 2005; De Nardi et al., 2010; French and Jones, 2011). Over the past decades researchers have started to recognize the potential of data that measure subjective expectations held by survey respondents, especially when elicited in terms of probabilities (see Manski, 2004, for a review). However, the validity of such intrinsically subjective data remains controversial. One issue is that expectations are sometimes elicited on a full percentage scale ranging from 0 to 100, but often on a coarser 11-point scale from 0 to 10. Examples of such coarse scales can be found in many large surveys, such as the Rand version of the Health and Retirement Study (HRS) in the US; the Survey of Health, Ageing and Retirement in Europe (SHARE); and the DNB Household Survey (DHS) and Longitudinal Internet Studies for the Social sciences (LISS) for the Netherlands. This paper evaluates the test-retest reliability of expectations reported on different answer scales during the same month. Moreover, we link this reliability to economic behavior through a structural model. We focus on expectations regarding one's own survival and compare the responses of the same individuals between two surveys, both of which aim to measure a number of points on the subjective survival curve.

Our data have been collected in the CentERpanel, a large household panel that is representative for the Dutch population. The two surveys analyzed are the Pension Barometer (PB, full probability scale) and the DNB Household Survey (DHS, 11-point scale). Given that each elicits beliefs by means of multiple items, reliability can be gauged at two levels. Firstly, we check whether the reponses are consistent with each other one-by-one. We compare probabilities reported by the same individuals for the same target ages, taking into account differences between answer scales. Secondly, we formulate a model in which we use all reported probabilities simultaneously to look at the relationships between subjective survival and background variables. We assess whether the two sets of probabilities yield similar associations between the hazard of death and socio-economic covariates.

Having quantified the correspondence between subjective beliefs elicited on different scales, we evaluate whether they are sufficiently reliable to be used as inputs for life cycle models of saving and labor supply. A life cycle model maps survival curves into behavior. We use a calibrated model to check whether such simulated behavior is sensitive to variation in survival curves of the magnitude observed between the surveys.

This paper draws on the literatures on subjective expectations and empirical life cycle models.[1] A rich body of research has established the covariates and predictive validity of survival expectations at the level of the individual (Hurd and McGarry, 1995, 2002; Smith et al., 2001; Bissonnette et al., 2017; Kutlu-Koc and Kalwij, 2017). To date, plausible associations between subjective survival and background variables provide the most important evidence in support of the validity of this type of data. However, the way questions are framed does affect reported expectations: a "die by" frame yields lower life expectancy than does a "live to" frame (Payne et al., 2012; Teppa et al., 2015). The fact that average reported probabilities are affected by framing begs the question whether responses are stable across response scales for a given frame. Lack of stability would pose a serious challenge to the use of subjective probabilities as they have been elicited so far. Reliability within frames would suggest that different question frames measure different stable concepts. Subsequent research could then try to establish which concept is most relevant for a given application.

Subjective probabilities have been used successfully in the estimation of structural models for different types of behavior, for instance in the context of anti-conception choice and the decision to pursue a certain college major (Delavande, 2008; Zafar, 2011; Van der Klaauw, 2012; Stinebrickner and Stinebrickner, 2013). However, while survival expectations play an important role in structural models of saving decisions, relatively few papers incorporate

[1]The literature on subjective expectations is summarized in Hurd (2009) and Manski (2004). Overviews of previous work on life cycle models can be found in De Nardi et al. (2016) (saving) and Blundell et al. (2016) (retirement).

subjective beliefs in life cycle models of labor supply and saving. Notable exceptions include the models of saving presented in Heimer et al. (2018), Wu et al. (2015) and Gan et al. (2015) and that of saving and labor supply in Van der Klaauw and Wolpin (2008). Researchers typically equate subjective longevity to actuarial forecasts, despite the robust finding that such figures are poor proxies even for average expectations (Bissonnette et al., 2017; Perozek, 2008). Moreover, life tables miss much of the heterogeneity found in subjective data. In order for subjective survival to be used in a life cycle model instead of actuarial figures, the data have to be interpreted as probabilities. This paper establishes whether such interpretation is valid even if the response scale consists of only 11 points.

Our analysis contributes to the literature in different ways. Test-retest analysis has been applied to survey data of various types, such as well-being (Krueger and Schkade, 2008). Hence, it allows one to compare the reliability of elicited beliefs to that of other, more commonly used types of data. Moreover, we analyze reliability at different levels of aggregation and investigate whether discrepancies between reported beliefs cancel out when probabilities are combined to fit survival curves. Furthermore, the data are clustered at two levels: the individual and the individual-survey. This enables one to disentangle the reliability of variation in beliefs for a given individual over time (within-variation) from the reliability of variation across individuals (between-variation). We take into account the specific measurement error that comes from rounding, either survey-induced or not. Finally, we simulate saving and labor supply to give economic meaning to our analysis.

Our overall finding is that reported probabilities are reliable across response scales. Test-retest correlations between individual probabilities are between 0.5 and 0.7, which is similar to the reliability of subjective well-being documented by Krueger and Schkade (2008). While only around 20% of reported probabilities are exactly equal, 25-37% are consistent when we account for the different resolutions of response scales. Rounding further increases the rate of consistent responses to 32-46% if we assume all probabilities reported by a given respondent are rounded similarly and 61-77% if we allow for the maximum degree of rounding for each reported probability. We cannot give a definitive answer as to which scale performs better, because we do not have an objective benchmark at the level of the individual. Nonetheless, both scales perform similarly well on the measures of of non-response, internal consistency and the incidence of 50s. Rounding is not related to education for either scale, but data quality as measured by consistency across scales is higher for those with education beyond secondary school. Models in which all reported probabilities are analyzed jointly show that associations between the hazard of death and most socio-demographic covariates are similar for both datasets, the exception being differences between cohorts. The oldest cohorts report higher probabilities of survival on the 11-point scale compared to the full percentage scale and this gap cannot be closed by rounding. Individual effects account for 90% of variation that cannot be explained by demographic covariates and are strongly correlated between surveys (correlations are around 0.8-0.9). The correlation between survey-effects is lower, suggesting that variation in beliefs across individuals is more reliable than longitudinal variation for a given individual. Wealth accumulation in a life cycle model is sensitive to survival expectations: the difference in median wealth between the surveys is about 30% after age 65 and actuarial tables generate wealth that is close to the DHS. However, modeling rounding substantially reduces the difference between simulated wealth profiles based on test and retest datasets. This difference is small compared to variation induced by unobserved heterogeneity. Hence, we conclude that expectations are sufficiently reliable to be used in structural models and that heterogeneity in survival expectations is an important determinant of saving in the life cycle framework. The fact that different response scales lead to similar estimates of survival curves and that this similarity improves when rounding is taken into account supports the interpretation of coarse scales as eliciting rounded probabilities.

The rest of the paper is structured as follows. Section 2 describes our data and section 3 evaluates the reliability of the reported probabilities one-by-one. Section 4 presents the model used to analyze all probabilities jointly, after which section 5 presents estimation results. We evaluate the economic significance of test-retest reliability by means of a life cycle model in section 6 and section 7 concludes.

# 2 Survival questions in the Pension Barometer and in the DNB Household Survey

Both the Pension Barometer (PB) and the DNB Household Survey (DHS) were administered to the CentERpanel. The CentERpanel is a household panel that is representative for the Dutch population and that is managed by CentERdata at Tilburg University. In both surveys respondents are offered multiple survival questions asking for the likelihood of surviving to different target ages based on their current age. Figure 1 shows graphically which ages are eligible for each question in both questionnaires. As can be seen in that figure, the PB elicits expectations for five equally spaced target ages between 70 and 90, while the DHS asks questions about age 65 and six ages between 75 and 100. We can directly compare probabilities corresponding to the target ages 75, 80, 85 and 90. The PB offers survival questions to respondents of age 25 and older who are at least 2 years younger than the target age for which expectations are elicited. Consequently, the potential sample for the PB is larger for questions referring to older ages and respondents of age 68 and younger are offered all five survival questions included in the survey. The DHS, on the other hand, asks one, two or three questions depending on the age of the respondent.

Other than the response format, the questions are phrased similarly in the PB and the DHS. The PB asks (emphasis added):

> "Please indicate on a **scale from 0 to 100 percent** how likely you think it is that you will live to age 70."
>
> ... percent

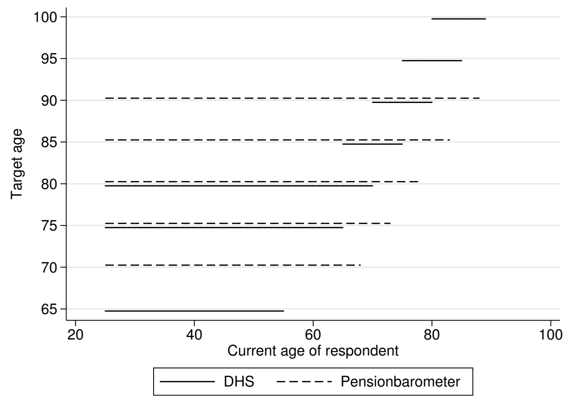



**Figure 1:** Age eligibility for survival questions in the DHS and in the Pensionbarometer

The items in the DHS are phrased as follows (emphasis added):

> "Please indicate your answer on a **scale of 0 thru 10**, where 0 means 'no chance at all' and 10 means 'absolutely certain'.
>
> How likely is it that you will attain (at least) the age of 65?"
>
> 0 1 2 3 4 5 6 7 8 9 10

While the formulation of the questions is similar, the answer formats differ substantially. Respondents have many more answer options in the PB compared to the DHS, so they can report their beliefs more precisely. Moreover, they do not face a response scale in the PB and have to type their answer. In the DHS, on the other hand, they select a number on a scale that is shown on screen.

# 3 Reliability of reported probabilities

## 3.1 Non-response, focal answers and sample construction

Rates of non-response and logically consistent answers are similar for the two surveys. 95% of age-eligible respondents answer all relevant PB survival questions compared with 91%

for the DHS. Moreover, 98% of complete responses to the PB and 99% of DHS responses decrease weakly with target age and are thus logically consistent. These rates of non-response and internally inconsistent answers are calculated separately for the two questionnaires and do not condition on answering both sets of items. Nonetheless, the fraction of logically consistent answers is much higher than for other sets of probabilities reported by the same respondents. For instance, only around 80% of the panel gave consistent answers when asked about their expectations of the replacement rate of income at retirement (De Bresser and Van Soest, 2013). Such variation across domains within the same panel suggests that respondents find it easier to report probabilities for some topics than for others. This is consistent with the finding that the propensity to give focal answers varies across domains within the HRS (Kleinjans and Van Soest, 2014).

Such focal point answers are responses that express an inability to reason in terms of probabilities, sometimes called epistemic uncertainty, rather than true subjective uncertainty (Bruine de Bruin et al., 2000, 2002; Bruine de Bruin and Carman, 2012). Focal answers have been used to explain excess heaping of reported probabilities at 50. As for the data analyzed in this study, 15% of probabilities are equal to 50 in either survey (see Appendix A). Two observations suggest that these answers do reflect expectations. Firstly, the fraction of 50-50s is higher for target ages around which people tend to die, age 80 and 85, than for younger or older ages. This is especially clear in the DHS, with fractions ranging from 6% for age 100 to 22% for target age 85. Secondly, only 3-4% of response sequences to either survey consist entirely of 50s. If respondents answer 50 when they are not sure how to respond, they would presumably resort to that strategy repeatedly when asked similar questions.

The metrics of non-response, internal consistency and 50/50s suggest that data quality is almost identical for both response scales. Appendix B shows that this also holds true when we stratify the sample by education. The rates of complete response are slightly lower for respondents with little education and these small differences are similar for the PB and the DHS. There is no variation in the tendency to give internally consistent answers across education groups. We find little evidence of a relationship between education and proneness to focal 50s conditional on reporting a complete and consistent set of probabilities.[2]

Before setting up a formal model, we investigate the extent to which the reported probabilities are consistent with each other for the same individuals and target ages. For most individuals both surveys were conducted in June of 2011 and 2012. Both questionnaires plausibly aim to measure the same expectations, since the period between questionnaires is short. In 2187 matched individual-year records the average time between surveys is 3.3 weeks with a median of 1 week and no more than 4 weeks between questionnaires for over three quarters of observations. Both surveys took place within one week for 6% of person-year observations.[3]

The structure of the data suggests two different samples that can be analyzed. Out of all potential person-year observations that were offered at least one set of items, one can directly compare the probabilities in the intersection of valid responses to both sets of questions. This selection rule results in a sample of 2,087 individual-years that report a total of 4,062 probabilities. The intersection of the PB and DHS will be used in the remainder of this section, where the focus lies on comparing probabilities one-by-one. When aggregating probabilities into survival curves, the sample need not be limited to this *intersection* and can be extended to the *union* of the PB and DHS. This extended sample uses all probabilities in complete and internally consistent response sequences for either survey and increases the sample for the models in section 4 from around 4,000 to 16,500 probabilities. Section 5 reports estimates for both samples. Appendix C shows that the union and intersection samples are

[2]While highly educated respondents report slightly fewer 50s relative to their less educated peers in the DHS, there is no difference in the fraction that report 50 percent for all target ages in either questionnaire.

[3]In the paper we report results using all available data, regardless of the time between surveys. Robustness checks indicate that none of our findings change when we limit the sample to cases for which the two surveys were taken within a one-week or four-week period, or exclude those cases in which test and retest surveys were taken within a period of seven days.

**Table 1:** Descriptive statistics of the reported survival probabilities and life table (LT) probabilities

| | N | Current age | Mean LT | PB Mean | PB S. D. | DHS Mean | DHS S. D. | Rank corr. |
|---|---|---|---|---|---|---|---|---|
| **a. Men** | | | | | | | | |
| Target age 75 | 823 | 25-63 | 75.2 | 65.3 | 23.0 | 68.0 | 19.2 | 0.66 |
| Target age 80 | 1000 | 25-68 | 60.6 | 52.7 | 24.9 | 55.7 | 22.7 | 0.68 |
| Target age 85 | 294 | 65-73 | 45.7 | 40.9 | 25.8 | 52.5 | 22.9 | 0.58 |
| Target age 90 | 188 | 70-78 | 25.1 | 26.4 | 24.6 | 38.5 | 24.6 | 0.55 |
| **b. Women** | | | | | | | | |
| Target age 75 | 690 | 25-63 | 83.6 | 65.8 | 22.5 | 67.5 | 19.0 | 0.56 |
| Target age 80 | 796 | 25-68 | 73.7 | 55.1 | 24.7 | 57.0 | 22.0 | 0.56 |
| Target age 85 | 168 | 65-73 | 61.7 | 44.5 | 26.0 | 54.0 | 23.0 | 0.61 |
| Target age 90 | 103 | 70-78 | 40.0 | 29.7 | 25.0 | 39.5 | 24.3 | 0.53 |

similar in terms of demographic characteristics to each other and to the potential sample of all person-years that were offered at least one survey.

## 3.2 Descriptives

Table 1 shows descriptives of reported subjective probabilities and corresponding probabilities from the 2010 life tables published by Statistics Netherlands.[4] Summary statistics are presented by target age and for each target age we limit the sample to those respondent-years that reported a probability in both surveys. Looking first at the means of the probabilities reported in the PB and in the DHS, we observe that they are close together for the target ages of 75 and 80: differences are less than 3 percentage points (pp). Differences between surveys are much smaller than variation within surveys, since the standard deviations are between 20 and 25pp. For the older target ages the average probability in the DHS is around 10pp higher than that in the PB. As a result the average DHS probability is higher than the life table forecast for ages 85 and 90 for men. Women report probabilities that are substantially

[4] The life tables are matched based on gender and age at the time of the survey, so differences between the age distribution of the Dutch population and that of the subsample that answers a particular question do not affect the comparison.

below actuarial predictions for all ages, so for them the DHS yields expectations that are more in line with official forecasts. Note that these larger differences for older target ages are still less than half of the standard deviations within measures. Hence, variation within surveys is more important than variation between surveys, even when the sample is reduced to narrow age groups (respondents for target ages 85 and 90 are 65-73 and 70-78 years old respectively). Average reported probabilities and their differences across answer scales do not vary with education. The only exception is the DHS at target age 90, for which the average probabilities are 43% and 35% for the poorly and highly educated respectively. However, this education gradient is not significantly different from that found for the PB, which is itself not significant (results available on request).

The (rank) correlations between PB and DHS probabilities are between 0.53 and 0.68, which is a similar range as that found for subjective well-being (Krueger and Schkade, 2008). Hence, according to this measure the reliability of subjective survival expectations is comparable to that of another widely researched type of subjective data, even though the levels are different for older respondents and target ages. Test-retest correlations are similar across education groups (results available on request).

### 3.3 One-by-one reliability

The most intuitive way to compare PB and DHS probabilities may be to look at the distribution of the differences between the two. However, the possibility of rounding implies that the (absolute) difference between reported probabilities is not a good measure of the extent to which the data are compatible. Rounding is the tendency for survey respondents to report a certain probability, e.g. 10%, whenever the true subjective probability lies in an interval, such as 5-15%. While rounding may be enforced by the answer format, such as the 11-point scale in the DHS, bunching of reported probabilities at multiples of 5 and 10 suggests that it also affects answers on full percentage scales (Manski and Molinari, 2010; Kleinjans and Van

Soest, 2014). Rounding means that the size of the difference between probabilities does not say much about whether those probabilities may reflect the same underlying expectations. For instance, reported probabilities of 100% in the DHS and 55% in the PB are consistent if the former is rounded to a multiple of 100 (so that the true probability lies in $[50, 100]$). On the other hand, probabilities of 70% and 55% would be incompatible, since the latter is only consistent with rounding to multiples of 1 or 5 and the intervals for the true probability do not overlap.

Therefore, our approach is to determine the extent of rounding based on three different rounding schemes and to check whether the probabilities reported in the PB and the DHS can reflect the same underlying true probability under each of those rules. The first scheme assumes that each probability is reported as precisely as allowed by each survey: all probabilities in the PB are rounded to multiples of 1 and all probabilities in the DHS to multiples of 10. Hence, under this *minimal* rounding rule any two probabilities are compatible if $P^{PB} \in \left[P^{DHS} - 5, P^{DHS} + 5\right]$.[5] The second, *common*, scheme allows for more rounding, but maintains that all survival probabilities reported by an individual are rounded similarly. We distinguish between the levels of rounding proposed by Manski and Molinari (2010): multiples of 100 (all probabilities 0 or 100); multiples of 50 (all 0, 50 or 100); multiples of 10; multiples of 5; at least one probability 1-4% or 96-99%; and at least one probability that does not fall in the other categories (multiples of 1). If at least one probability is 1-4% or 96-99% and the rest are all multiples of 5, the former probabilities are assumed to be precisely reported while the latter are interpreted as rounded to multiples of 5. For example, if a respondent reports $\{100, 60, 55\}$, the common rounding scheme interprets the sequence as rounded to multiples of 5, so the reported probability of 60% yields the interval $[57.5; 62.5)$ for the true probability. The 11-point scale of the DHS only allows for rounding to multiples of 100, 50 and 10. Finally, the third *general* rounding rule interprets each reported probabil-

[5] $P^{PB} = 15$ is consistent with $P^{DHS} = 10$ *and* $P^{DHS} = 20$, since the true PB probability may be anywhere in $[14.5, 15.5)$.

**Table 2:** Fraction of consistent responses to PB and DHS survival questions

| | N | Exactly equal | Minimal rounding | Common rounding | General rounding |
|---|---|---|---|---|---|
| Age 75 | 1513 | 0.22 | 0.37 | 0.46 | 0.77 |
| Age 80 | 1796 | 0.22 | 0.31 | 0.40 | 0.75 |
| Age 85 | 462 | 0.18 | 0.26 | 0.34 | 0.68 |
| Age 90 | 291 | 0.16 | 0.24 | 0.32 | 0.61 |
| All combined | 2087 | 0.09 | 0.18 | 0.27 | 0.63 |

ity to be rounded to the maximum extent, regardless of the other answers of that respondent (Bissonnette and de Bresser, 2018). We distinguish between rounding to multiples of 100, 50, 25, 10, 5 and 1 for the PB and to multiples of 100, 50 and 10 for the DHS. Hence, a reported probability of 35% in the PB is interpreted as rounded to a multiple of 5 regardless of the other probabilities, and a probability of 50% is always rounded to a multiple of 50.

Appendix D shows the distribution of rounding in the sample according to both the common and the general rounding rule. Under common rounding we find that rounding to multiples of 5 is the most prevalent type for the PB, while rounding to multiples of 10 is most prevalent for the DHS (57% of individuals round to multiples of 5 in the PB and 94% round to multiples of 10 in the DHS). Rounding to multiples of 10 is the most frequent category for general rounding at the level of the probability (52% of PB probabilities and 76% of DHS probabilities are rounded to multiples of 10). According to both rounding schemes all three education groups round their answers similarly for both question types (results available on request).

The rates of compatible responses to PB and DHS questions by target age are shown in Table 2. Around one fifth of reported probabilities are equal across questionnaires. Under the assumption that all probabilities are rounded to the minimal extent allowed by each survey, the rates of consistent response range from 37% for target age 75 to 24% for age 90. Allowing for common rounding increases this to 32-46% and under the most conservative general rounding scheme 61-78% of responses are compatible with at least one underlying true probability. The fraction of consistent responses is higher for younger target ages regardless of the rounding rule. These differences are mostly related to the current age of the respondents, rather than the target age to which questions refer. For a given target age the rate of consistent answers to the two sets of questions is flat up to age 68 and declines sharply afterwards, which matches the age gradient in probability numeracy documented by Hudomiet et al. (2018). The rate of consistent probabilities is the same when we restrict the sample to those observations that report the same level of subjective health in both surveys or to surveys taken within a one-week or four-week period. Furthermore, time between surveys is not a significant predictor of the absolute difference between probabilities in a multivariate model. Hence, differences probably reflect measurement error rather than changes in the actual expectations held by respondents. Moreover, the fact that the fraction of consistent response is higher for younger target ages regardless of the rounding scheme shows that age-related differences in rounding cannot explain the divergence in levels observed for target ages 85 and 90.

The consistency of answers across the two scales does vary with education. Respondents with a university degree are more likely to report consistent probabilities for the PB and the DHS for all relevant target ages than are individuals with no education beyond lower secondary school. This disparity is especially pronounced at the younger target ages of 75 and 80, for which rates of consistent response are 5-8pp higher for those who finished university education (results available on request).

The upshot of the one-by-one comparison is that while the two sets of probabilities are fairly strongly correlated, it takes considerable rounding error to make the PB and DHS compatible with at least one underlying true probability for a majority of the cases. The differences between the two sets of probabilities raise the question whether reliability can be improved by modeling all probabilities jointly, or by modeling rounding error and expectations simultaneously. In the next section we set up two models to answer those questions.

# 4 Reliability of survival curves

## 4.1 Model without rounding

The model we use in this paper is closely related to that proposed by Kleinjans and Van Soest (2014) for expectations regarding binary outcomes and extended to continuous outcomes in De Bresser and Van Soest (2013). The key advantage is that it allows aggregation of all information contained in the probabilities into survival curves for an individual or socio-economic group.

Expectations follow a Gompertz distribution with the baseline hazard shifted proportionally by demographic variables. This parameterization of expectations implies that the true probability of surviving to target age $ta_k$, conditional on having survived to current age $a_{it}$, is given by:

$$\begin{aligned} S_{itk}^q|a_{it} &= \Pr\left(t \geq ta_k | t \geq a_{it}\right) = \frac{\Pr\left(t \geq a_{it} | t \geq ta_k\right) \times \Pr\left(t \geq ta_k\right)}{\Pr\left(t \geq a_{it}\right)} \\ &= \frac{1 \times \Pr\left(t \geq ta_k\right)}{\Pr\left(t \geq a_{it}\right)} = \frac{\exp\left(-\frac{\gamma_{it}^q}{\alpha^q}\left(\exp\left(\alpha^q\left(ta_k/100\right)\right)-1\right)\right)}{\exp\left(-\frac{\gamma_{it}^q}{\alpha^q}\left(\exp\left(\alpha^q\left(a_{it}/100\right)\right)-1\right)\right)} \end{aligned} \tag{1}$$

where $q$ indexes questionnaires ($q \in \{PB, DHS\}$); $\gamma_{it}^q = \exp\left(\mathbf{x}_{it}'\boldsymbol{\beta}_1^q + \xi_i^q + \eta_{it}^q\right)$ depends on the demographics of respondent $i$ in survey-year $t$; $\alpha^q$ determines the shape of the baseline hazard; $ta_k$ is a target age in the questionnaire and $a_{it}$ is the age of $i$ in year $t$. We distinguish two types of unobserved heterogeneity: individual effects $\xi_i^q$ and question sequence effects $\eta_{it}^q$. Distributional assumptions for these error components follow. The null hypotheses of interest are that $\boldsymbol{\beta}_1^{PB} = \boldsymbol{\beta}_1^{DHS}$ and $\alpha^{PB} = \alpha^{DHS}$, which imply that the two surveys yield the same associations between covariates and survival and the same baseline hazard of death. We divide both the target age and the current age by 100 to facilitate estimation of $\alpha^q$.

However, we do not observe $S^q_{itk}$ directly. Instead, the reported probabilities are perturbed by recall error:

$$P^{*q}_{itk} = S^q_{itk} + \varepsilon^q_{itk}; \quad \varepsilon^q_{itk} \sim \mathcal{N}\left(0, \sigma^2_{it}\right) \tag{2}$$

where recall error $\varepsilon^q_{itk}$ is independent of all covariates and across thresholds, surveys, years and individuals. The assumption that recall errors are independent across probabilities reported by a given individual might seem strict, but note that positive correlation between such errors would be indistinguishable from heterogeneity in expectations. We do allow for heteroskedasticity and model the variance of recall errors as $\ln\left(\sigma_{it}\right) = \mathbf{x}'_{it}\boldsymbol{\beta}^q_2$. Such heteroskedasticity allows some groups to report expectations that differ more from the Gompertz distribution than others, for instance because their baseline hazard is not monotonic in age (i.e. the risk of death does not increase or decrease monotonically with age). We regard the fact that expectations are approximated by a functional form that restricts them to be reasonable as an advantage, especially when it comes to using such expectations in a life cycle model. Moreover, the Gompertz specification captures all heterogeneity in expectations through a single parameter, which makes it feasible to incorporate such variation in dynamic models. While it is simple, the model does capture variation in individuals' ability to reason in terms of probabilities.

In the baseline model we do not allow for rounding of the reported probabilities, but we do take into account censoring between zero and the lowest probability reported previously in the sequence. Hence, the density for a reported probability $P^q_{itk}$ conditional on covariates is given by

$$f\left(P^q_{itk}|\mathbf{x}_{it}\right) = \begin{cases} 1-\Phi\left(\frac{P^q_{it,k-1}-S^q_{itk}}{\sigma_{it}}\right) & \text{if } P^q_{itk} = P^q_{it,k-1} \text{ (censored from above)} \\ \phi\left(\frac{P^q_{itk}-S^q_{itk}}{\sigma_{it}}\right) & \text{if } 0 < P^q_{itk} < P^q_{it,k-1} \text{ (uncensored)} \\ \Phi\left(\frac{P^q_{itk}-S^q_{itk}}{\sigma_{it}}\right) & \text{if } P^q_{itk} = 0 \text{ (censored from below)} \end{cases} \tag{3}$$

where $\phi(.)$ and $\Phi(.)$ respectively denote the standard normal density and CDF and for the first threshold $k = 1$ we set $P^q_{it0} = 100\%$.[6]

The model is completed by distributions of the individual effects $\xi^q_i$ and survey effects $\eta^q_{it}$. Both are assumed to be bivariate normal with covariance matrices $\Sigma_\xi$ and $\Sigma_\eta$ and they are independent of covariates and of each other. We estimate the elements of the covariance matrices of unobserved heterogeneity, the baseline hazards $\alpha^{PB}$ and $\alpha^{DHS}$ and the vectors $\boldsymbol{\beta}^{PB}_1$, $\boldsymbol{\beta}^{PB}_2$, $\boldsymbol{\beta}^{DHS}_1$ and $\boldsymbol{\beta}^{DHS}_2$ by maximum simulated likelihood where we integrate numerically over the distributions of individual and question sequence effects.

## 4.2 Model with rounding

The basic setup is the same as for the baseline model, but now $P^{*q}_{itk}$ is not only censored but also rounded prior to being reported. We allow for rounding to multiples of 100, 50, 25, 10, 5 and 1 for the PB and to multiples of 100, 50 and 10 for the DHS. The rounding model is ordinal:

$$R^q_{itk} = r \iff \mu^q_{r-1} \leq y^{*q}_{it} = \mathbf{x}'_{it}\boldsymbol{\beta}^q_3 + \xi^{r,q}_i + \eta^{r,q}_{it} + \varepsilon^r_{itk} < \mu^q_r \tag{4}$$

where $r \in \{1, 5, 10, ..., 100\}$ for the PB and $r \in \{10, 50, 100\}$ for the DHS. The rounding equation includes individual and question sequence effects, allowing rounding to be corre-

[6] When estimating the model we also condition on individual and survey effects, but we omit them here for ease of exposition.

lated across repeated observations for a given individual and to be more strongly correlated within survey waves than between them. Moreover, both types of unobserved heterogeneity may be correlated across surveys (PB and DHS) and with their respective counterparts in the equation that shifts survival curves ($\xi_i^{PB}$, $\xi_i^{DHS}$, $\xi_i^{r,PB}$ and $\xi_i^{r,DHS}$ follow a fourdimensional normal distribution and so do the survey effects $\boldsymbol{\eta}_{it}$). We assume that the idiosyncratic rounding shocks $\varepsilon_{itk}^{r}$ follow a standard normal distribution and are independent from covariates and all other errors, so the conditional probabilities of each category of rounding $\Pr\left(R_{itk}^{q} = r|\mathbf{x}_{it}, \boldsymbol{\xi}_i, \boldsymbol{\eta}_{it}\right)$ take the shape of an ordered probit.

A reported probability in combination with a particular level of rounding implies an interval for the perturbed probability $P_{itk}^{*q} \in [LB_{itk}^{r}, UB_{itk}^{r})$. For instance, a reported probability of 25% that is rounded to a multiple of 5 yields the interval $P_{itk}^{*q} \in [22.5, 27.5)$. The probability of that event is easy to calculate, since $P_{itk}^{*q} \sim \mathcal{N}\left(S_{itk}^{q}, \sigma_{it}^{2}\right)$. However, rounding is a latent construct, because a given reported probability may result from different degrees of rounding. A reported probability of 25% may be rounded to a multiple of 25 (interval: 12.5-37.5%); to a multiple of 5 (interval: 22.5-27.5%); or to a multiple of 1 (interval: 24.5-25.5%). We therefore average across the different degrees of rounding to obtain the likelihood contribution. In particular, define for each reported probability the set $\Omega_{itk}$ that consists of all types of rounding that are consistent with that probability. We obtain the conditional density as:

$$f\left(P_{itk}^{q}|\mathbf{x}_{it}\right) = \sum_{r \in \Omega_{itk}} \Pr\left(R_{itk}^{q} = r|\mathbf{x}_{it}\right) \times \Pr\left(LB_{itk}^{r} \leq P_{itk}^{*q} < UB_{itk}^{r}|\mathbf{x}_{it}\right) \tag{5}$$

where $\Pr\left(LB^r_{itk} \le P^{*q}_{itk} < UB^r_{itk}|\mathbf{x}_{it}\right)$ is given by

$$\Pr\left(LB^r_{itk} \le P^{*q}_{itk} < UB^r_{itk}|\mathbf{x}_{it}\right) = \begin{cases} \Pr\left(LB^r_{itk} \le P^{*q}_{itk}|\mathbf{x}_{it}\right); & \text{if } P^q_{itk} \ge P^q_{it,k-1} - 0.5r \\ \quad LB^r_{itk} = P^q_{it,k-1} - 0.5r & \\ \Pr\left(LB^r_{itk} \le P^{*q}_{itk} < UB^r_{itk}|\mathbf{x}_{it}\right); & \text{if } 0.5r \le P^q_{itk} < P^q_{it,k-1} - 0.5r \\ \quad LB^r_{itk} = P^q_{itk} - 0.5r & \\ \quad UB^r_{itk} = P^q_{itk} + 0.5r & \\ \Pr\left(P^{*q}_{itk} < UB^r_{itk}|\mathbf{x}_{it}\right); & \text{if } P^q_{itk} < 0.5r \\ \quad UB^r_{itk} = 0.5r & \end{cases} \tag{6}$$

Whether or not a given probability is censored depends on the degree of rounding and on the preceding probability. An example: a reported 25% is consistent with rounding to multiples of 25, 5 and 1. This means $\Omega = \{25, 5, 1\}$. If the preceding probability is 35%, the density is given by

$$\begin{aligned} f(25) = {} & \Pr(R = 25) \times \Pr(P^* \ge 12.5) + \Pr(R = 5) \times \Pr(22.5 \le P^* < 27.5) \\ & + \Pr(R = 1) \times \Pr(24.5 \le P^* < 25.5) \end{aligned} \tag{7}$$

All probabilities in the equation above are calculated from univariate normal distributions and are therefore easy to obtain.

Figure 2 illustrates the logic of both versions of the model. The circles are hypothetical reported probabilities of survival to the target ages presented in the PB. The model poses that reported probabilities are generated from true probabilities given by a Gompertz distribution, which is the solid black line in both panels. In both models true (Gompertz) probabilities are perturbed by recall errors. The distributions of those errors are given by the dark grey normal

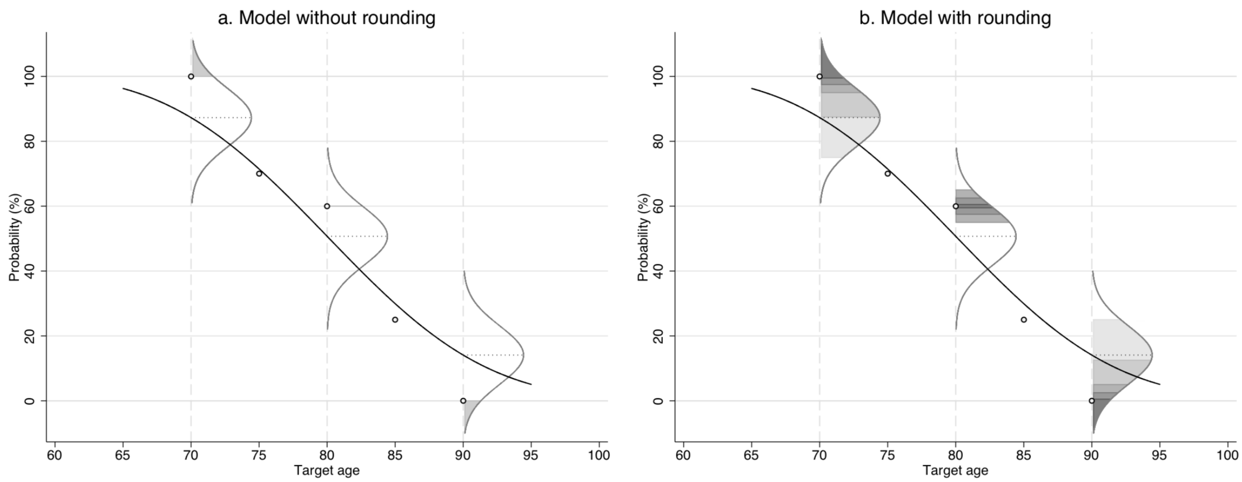



**Figure 2:** Illustration of the likelihood for hypothetical data

distributions for the target ages 70, 80 and 90. In the model without rounding, panel a., reported probabilities are either exactly equal to the perturbed true probabilities or censored between zero and the minimum of 100 and the lowest probability reported previously. This censoring is illustrated for the target ages of 70 and 90, which are censored at 100 and zero respectively. The other probabilities are not censored, so the likelihood for age 80 is given by the normal density itself rather than some area underneath. For the model with rounding, panel b., reported probabilities are not only perturbed by recall error but also rounded. As a result, the data only yield intervals within which perturbed probabilities fall even in case the reported probability has not been censored. Figure 2 displays those intervals for target ages 70, 80 and 90. Note that different levels of rounding imply different intervals that are shown in different shades of grey: the darkest areas correspond to rounding to multiples of 1 and the lightest intervals to multiples of 50. For instance, the darkest area for age 80 runs from 59.5% to 60.5%, which is the range in which the perturbed probability should fall if the reported probability of 60% is rounded to a multiple of 1. Rounding is not observed perfectly, so the likelihood calculates the probability corresponding to each shaded area and averages over the different levels of rounding.

# 5 Results

This section presents estimation results for the two models of subjective life expectancy explained above. Descriptive statistics for the covariates are reported in Appendix C. In the main text we only report estimates for the equations that govern expectations. Estimates of the recall error and rounding processes can be found in Table E1 of Appendix E. The sample from which the estimates presented in the main text are obtained limits the data to complete and consistent responses for *both* sets of probabilities. Moreover, we only use the probabilities corresponding to those target ages for which *both* a PB and a DHS probability are available (the "intersection" sample in Appendix C). Estimates based on all complete and consistent responses for *either* one of the questionnaires, regardless of whether the target age is included in both (the "union" sample in Appendix C), corroborate the findings from the main text and can be found in Appendix F. A graphical analysis of model fit with a focus on rounding is included in Appendix G.

## 5.1 Model without rounding of reported probabilities

Estimation results of the model without rounding are presented in the left panel of Table 3 (see section 4.1 for a description of this model). The first two columns on the left present hazard ratios that capture the relationships between covariates and the hazard of death and the third column contains differences between these hazard ratios across the two surveys. The estimated associations for most variables are both qualitatively and quantitatively similar for the PB and the DHS, with the exception of the cohort dummies. The baseline cohort 1942-1951 has a relatively low hazard of death according to the DHS: the hazard rates for the cohorts born between 1952 and 1981 are between 15 and 30 percent higher than the baseline. However, according to the PB only the cohort 1952-1961 has a significantly higher hazard than the baseline and the difference is only 12 percent. These large differences between cohorts in the DHS and smaller and mostly insignificant differences in the PB lead us to reject the null hypotheses of equal cohort effects for all cohorts. The cohort differences in the DHS mirror the relatively high average probabilities at old target ages observed in Table 1 and indicate that such differences are too large to be explained by the current age of respondents or other covariates such as health. However, the PB data for different target ages can be interpreted according to a common set of expectations without prominent cohort effects.

We do not find evidence to suggest that the two surveys generate different results for most other covariates. The dummy for the year 2012 is insignificant for both surveys. Women report a lower hazard of death compared to men, the hazard ratio is 93% according to the PB and 95% in the DHS. We find some disagreement between the PB and the DHS for the income dummy corresponding to a net household income of 1151-1800 euro per month. Based on the PB individuals in this group have a 18% higher hazard of death than the baseline of individuals in households that earn more than 2600 euro per month. However, in the DHS this difference is close to zero. Such disagreement is not there for the other income groups, for which we cannot reject the null of equal coefficients. The education dummies show similar patterns for the PB and the DHS: respondents in the middle education category have a 14-16% lower hazard of death than their less educated peers. Though the PB shows a statistically significant difference of 9% for the high education category, this difference is only 2% and not significant for the DHS. However, the hazard ratios are not significantly different between the surveys. As for self-reported health, respondents who rate their current health worse report substantially higher hazards of death regardless of the set of probabilities used. The average hazard of respondents who rate their health as “not good” or “poor” is 86-94% higher than that of respondents who rate their health as “excellent”. None of the coefficients for the health variables differs significantly between the two surveys. The Chi-squared tests for joint equality of coefficients across the PB and DHS reported in Table 3 reflect the differences between surveys in cohort effects and one income dummy: we reject the null of joint equality and much more strongly so if we take the cohort dummies into account.

The bottom of Table 3 reports other estimates. The baseline hazard is significant and positive for both surveys, which means that the hazard of death increases with age. Moreover, the estimated coefficients are very close, around 8.1 for both datasets, and the difference is not statistically significant. The estimated variances of the individual effects indicate that expectations are persistent at the level of the individual for both datasets: around 90% of the variance in expectations that cannot be explained by covariates is due to permanent unobserved heterogeneity. Furthermore, the test-retest correlation of individual effects is 0.87, which is much higher than that of individual probabilities.

Table E1 in Appendix E presents estimates of the coefficients that capture heteroskedasticity of the recall error, capturing variation in the extent to which reported probabilities fit the Gompertz distribution. In addition to some differences between cohorts, the only factor that affects recall error similarly in both sets of probabilities is education. The middle and high education categories report probabilities that deviate significantly less from Gompertz probabilities compared to respondents who have not finished vocational training. This may well reflect the education gradient in understanding of probabilities documented by Hudomiet et al. (2018).

Table F1 in Appendix F contains estimates of the exact same model, estimated on the larger sample of complete and consistent responses to either set of survival questions, using all available probabilities (including those target ages that are not in one of the questionnaires). The same general picture emerges, but the differences between surveys in terms of estimated cohort effects are markedly smaller (less than 10pp compared to 15-35pp for the estimates from common target ages). Using more probabilities for each survey wave reduces cohort differences for the DHS and hence brings the surveys more in line. Furthermore, we reject equality of coefficients for one additional income dummy (for an income between 1801 and 2600 euro per month).

**Table 3:** Gompertz models of subjective survival

| | Model 1 – No rounding | | | Model 2 – Rounding | | |
|---|---|---|---|---|---|---|
| | PB[a] | DHS[a] | Diff. PB - DHS | PB[a] | DHS[a] | Diff. PB - DHS |
| **a. Hazard ratios** | | | | | | |
| Wave 2012 | 1.009 | 0.993 | 0.0165 | 0.997 | 1.003 | -0.00603 |
| | (0.0236) | (0.0177) | (0.0272) | (0.0196) | (0.0151) | (0.0223) |
| Female | 0.927** | 0.948* | -0.0207 | 0.830*** | 0.904*** | -0.0741*** |
| | (0.0293) | (0.0290) | (0.0295) | (0.0225) | (0.0220) | (0.0244) |
| *Cohorts (baseline: 1942-1951)* | | | | | | |
| Coh. 1932-41 | 1.128 | 0.975 | 0.153** | 1.240*** | 0.888** | 0.352*** |
| | (0.0833) | (0.0646) | (0.0673) | (0.0672) | (0.0433) | (0.0589) |
| Coh. 1952-61 | 1.118** | 1.276*** | -0.158*** | 1.055 | 1.160*** | -0.105*** |
| | (0.0574) | (0.0607) | (0.0520) | (0.0354) | (0.0364) | (0.0389) |
| Coh. 1962-71 | 0.930* | 1.147*** | -0.217*** | 1.020 | 1.278*** | -0.258*** |
| | (0.0373) | (0.0559) | (0.0489) | (0.0360) | (0.0483) | (0.0452) |
| Coh. 1972-81 | 0.956 | 1.298*** | -0.342*** | 1.120** | 1.316*** | -0.195*** |
| | (0.0567) | (0.0831) | (0.0686) | (0.0501) | (0.0520) | (0.0558) |
| Coh. 1982-87 | 0.813 | 0.981 | -0.168* | 0.895 | 0.954 | -0.0590 |
| | (0.115) | (0.125) | (0.0931) | (0.104) | (0.0670) | (0.0737) |
| *Net household income (baseline: more than €2600)* | | | | | | |
| Net HH. Inc. $\leq$ €1150 | 0.980 | 0.928 | 0.0524 | 1.220*** | 1.094 | 0.126 |
| | (0.0748) | (0.0752) | (0.0730) | (0.0891) | (0.0737) | (0.0830) |
| Net HH. Inc. €1151-1800 | 1.181*** | 0.994 | 0.188*** | 1.274*** | 1.046 | 0.228*** |
| | (0.0522) | (0.0416) | (0.0521) | (0.0567) | (0.0372) | (0.0540) |
| Net HH. Inc. €1801-2600 | 0.933* | 0.925** | 0.00850 | 0.924*** | 0.938*** | -0.0138 |
| | (0.0332) | (0.0303) | (0.0336) | (0.0270) | (0.0229) | (0.0290) |
| *Education (baseline: lower secondary)* | | | | | | |
| Educ. higher sec./vocational | 0.858*** | 0.838*** | 0.0202 | 1.025 | 0.958 | 0.0668* |
| | (0.0344) | (0.0334) | (0.0353) | (0.0363) | (0.0299) | (0.0360) |
| Educ. (applied) university | 1.091*** | 1.024 | 0.0672 | 1.151*** | 1.057* | 0.0936** |
| | (0.0369) | (0.0419) | (0.0413) | (0.0323) | (0.0327) | (0.0373) |
| *Health (baseline: excellent)* | | | | | | |
| Health: good | 1.263*** | 1.346*** | -0.0825 | 1.437*** | 1.303*** | 0.134*** |
| | (0.0416) | (0.0554) | (0.0599) | (0.0407) | (0.0401) | (0.0512) |
| Health: fair | 1.725*** | 1.710*** | 0.0156 | 2.153*** | 1.717*** | 0.436*** |
| | (0.0755) | (0.0838) | (0.0923) | (0.0953) | (0.0691) | (0.0976) |
| Health: not good/poor | 1.859*** | 1.938*** | -0.0782 | 2.199*** | 2.001*** | 0.198 |
| | (0.139) | (0.143) | (0.157) | (0.117) | (0.0977) | (0.133) |
| Constant | 0.00650*** | 0.00526*** | 0.00124*** | 0.00531*** | 0.00436*** | 0.000950* |
| | (0.000335) | (2.069e-06) | (0.000335) | (0.000310) | (0.000430) | (0.000499) |
| Chi2 test joint equality (16df) | 86.90 ($p < 0.0001$) | | | 154.25 ($p < 0.0001$) | | |
| Chi2 test joint equality no cohorts (11df) | 36.04 ($p = 0.0002$) | | | 69.60 ($p < 0.0001$) | | |
| **b. Other estimates** | | | | | | |
| Baseline hazard ($t/100$) | 8.119*** | 8.084*** | 0.0342 | 8.104*** | 8.385*** | -0.282** |
| | (0.0765) | (0.0775) | (0.0992) | (0.0696) | (0.123) | (0.140) |
| Variance ind. effects | 0.771*** | 0.481*** | | 0.635*** | 0.431*** | |
| | (0.0400) | (0.0265) | | (0.0248) | (0.0185) | |
| Corr. ind. effects | 0.870*** | | | 0.787*** | | |
| | (0.0163) | | | (0.0155) | | |
| Variance seq. effects | 0.0818*** | 0.0610*** | | 0.112*** | 0.0300*** | |
| | (0.0153) | (0.0114) | | (0.00776) | (0.00489) | |
| Corr. seq. effects | 0.0324 | | | 0.239*** | | |
| | (0.0774) | | | (0.0743) | | |
| Fraction var. ind. effects | 0.904*** | 0.888*** | | 0.851*** | 0.935*** | |
| | (0.0175) | (0.0213) | | (0.0107) | (0.0107) | |
| No. individuals | | 1,470 | | | 1,470 | |
| No. probabilities | | 4,034 | | | 4,034 | |
| Log-likelihood | | -30,530.175 | | | -16,048.925 | |

[a] Estimates reported as hazard ratios.
Standard errors in parentheses; ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

## 5.2 Model with rounding of reported probabilities

Estimates for the model that accounts for rounding, described in section 4.2, are reported in the right panel of Table 3. As was the case without rounding, the model with rounding shows that the significant relationships between the hazard of death and covariates that emerge for the PB and the DHS go in the same direction in almost all cases. The only exception is the oldest cohort, which has a 24% higher hazard than the baseline according to the PB but an 11% lower hazard based on the DHS. Moreover, the size of many correlations remains comparable between the surveys. However, incorporating rounding does not reduce the differences between the estimates from the two datasets and actually leads to more frequent rejections of equality. For the cohort dummies we find that rounding does not fill the gap in the DHS between older cohorts who expect to live long and younger ones who expect to live shorter. In addition to the dummy for household income between 1151 and 1800 euro per month, we also reject equality for the variables capturing gender and education and for two out of three indicators for health. The estimates regarding unobserved heterogeneity are similar to those for the model without rounding, but the difference between the baseline hazards is larger.

The right panel in Table E1 contains the remaining estimates. The third and fourth column in Table E1 show the estimates for the heteroskedasticity of recall errors in the PB and DHS respectively. The variance of the errors is significantly lower among higher education groups, as was the case in the model without rounding. Compared to the left panel there are two additional columns, which shows the estimated coefficients of the rounding equations for the PB and DHS. Standard errors are not reported for the rounding equation in the DHS due to flatness of the simulated log-likelihood function. Moreover, the estimated thresholds for the DHS rounding rule are very large. Thus, the estimates indicate that almost all probabilities in DHS are rounded to multiples of 10. The coefficients of the rounding equation for the PB, shown in the penultimate column, also come with large standard errors. However, we do

**Table 4:** Model-implied average rounding probabilities

| Multiples of... | Pension Barometer (%) | DNB Household Survey (%) |
|---|---|---|
| ...100 | 0 | 2 |
| ...50 | 5 | 4 |
| ...25 | 11 | – |
| ...10 | 47 | 94 |
| ...5 | 32 | – |
| ...1 | 4 | – |

estimate the thresholds for different levels of rounding in the PB precisely. Sample average simulated rounding probabilities are reported in Table 4, which shows that half of the reported probabilities in the PB are rounded to multiples of 10 and a third is rounded to multiples of 5. As suggested by the numerical issues associated with estimating the rounding equation for the DHS, 94 percent of probabilities reported in the DHS are rounded to multiples of 10.

While the model described in Section 4.2 accounts for rounding, it does not allow for excess bunching of probabilities at 50 due to focal answers. Appendix G shows that the model closely matches the observed fraction of 50s based on rounding alone: the model-implied fraction of 50s is 14% for the PB (16% in data) and 16% for the DHS (17% in data). This corroborates the notion that the observed 50s can be explained without invoking focal answers.

## 5.3 One-by-one vs. aggregate reliability

Having estimated models that relate survival to demographic variables, we evaluate whether aggregation increases the reliability of survival expectations relative to the comparison of individual probabilities conducted in section 3. As before, we look at reliability in terms of both levels and variation.

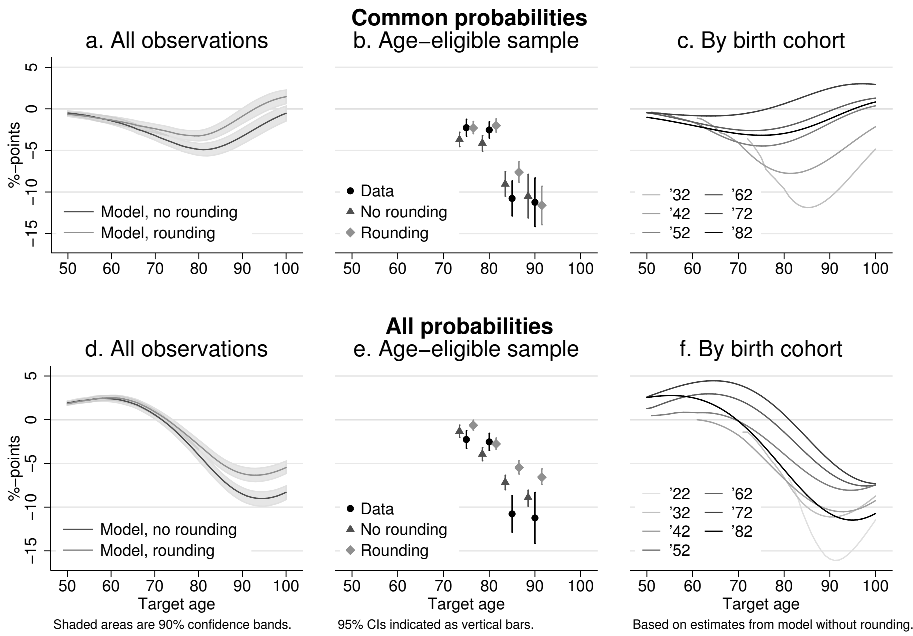



**Figure 3:** Average difference between probability of survival to different ages

### Levels

Figure 3 shows average differences between simulated PB and DHS survival probabilities for target ages between 50 and 100. All probabilities condition on the current age of the respondent and are the simulated analogues of the probabilities reported by survey respondents. While the top panels are based on estimates from the sample of those probabilities that are common to both questionnaires, the intersection sample analyzed in Section 3, the bottom panels are obtained from estimates based on the union of all valid probabilities reported in either one of the surveys.

Panel a. considers probabilities for all respondents who are at least two years younger than a target age (the eligibility criterion used by the PB). The two lines correspond to the models with and without rounding error. Both models perform similarly around age 50, where the survival curves are close to 1 for both surveys. Differences between surveys then grow, up to -3 to -5pp around age 80. This average difference is in line with that between reported probabilities shown in Table 2 and in panel b. of Figure 3, indicating that aggregation does not improve reliability at these ages. However, while the differences between surveys in the raw data are much larger for target ages 85 and 90 compared with ages 75 and 80, the size of the differences between simulated probabilities declines after age 80 to between -1 and 2pp around age 100. Hence, the average levels of simulated probabilities for the older target ages are closer than the raw data. Panel d. indicates that the same holds for the model estimated on all available probabilities, though differences at advanced age are larger at -6pp to -9pp.

Accounting for rounding improves reliability for older ages starting from 75 in both samples. The fact that average differences between simulated probabilities are smaller for the model that accounts for rounding seems to contradict the estimates in Table 3, which show no indication that modeling rounding improves similarities between hazard ratios estimated on the two surveys. The explanation is that the larger differences between the estimates for the hazard ratios and baseline hazards for the model with rounding cancel out partly when combined into survival probabilities. The estimated baseline hazard is higher for the DHS than for the PB, but the hazard ratios through which covariates influence mortality are lower in the DHS (especially for "poor" health, see Table 3). In the model without rounding the difference between baseline hazards is much smaller, so differences between the hazard ratios translate directly into differences between probabilities. In the model with rounding, differences in the two parameters partly cancel out.

Panels b. and e. of Figure 3 facilitate the comparison between simulations and data. They limit the sample to those individuals who were age-eligible for both surveys, reproducing the sample of Tables 1 and 2. Once we limit the sample in this way, panel b. indicates that model-implied differences are similar to those in the data for all target ages. Hence, the good performance of the models for older ages in panel a. is not due to errors canceling out within probability-sequences reported by given individuals. Instead it can be explained by differences being calculated from a different sample that includes younger respondents for the oldest target ages. Panel e., on the other hand, shows that cohort differences do not explain the smaller differences between average probabilities at advanced age when all available probabilities are used in estimation. Even when limiting the sample to those cohorts on which the direct comparison is based, the model that accounts for rounding improves reliability relative to the individual probabilities.

Panels c. and f. illustrate the importance of cohort effects in the reliability of expectations by plotting the average difference between the PB and the DHS separately for all birth cohorts in the estimation samples. While test-retest reliability for levels is high for most younger cohorts, differences are larger for the oldest cohorts.[7]

**Variation**

Figure 4 shows simulation results concerning variability and correlation. Panels a. and c. measure variation by Standard Deviations (SDs). The solid lines labelled "across unobserved heterogeneity" show the average over observations of the SD across draws of unobserved heterogeneity (both individual and sequence effects). The dashed lines labelled "across observations" are the SDs across observations of the average probability, where the average is taken across draws of unobserved heterogeneity. The main finding to emerge from panels a. and c. of Figure 4 is that variation across unobserved heterogeneity for fixed covariates is stronger than covariation with covariates, since the SD across heterogeneity is larger than that across observations for all target ages. Though covariates matter, there is important heterogeneity in expectations left after they are accounted for. Furthermore, as reflected in the estimates in Table 3, there is more such unobserved heterogeneity in the PB than in the DHS. Comparing Figures 3 and 4 shows that even variation across observations is large relative to the average difference between questionnaires for all target ages (the differences for the two datasets peak at -3pp and -6pp around the age where SDs are 10pp).

[7] Panels c. and f. of Figure 3 use estimates from the model without rounding. Similar patterns were found in the model with rounding.

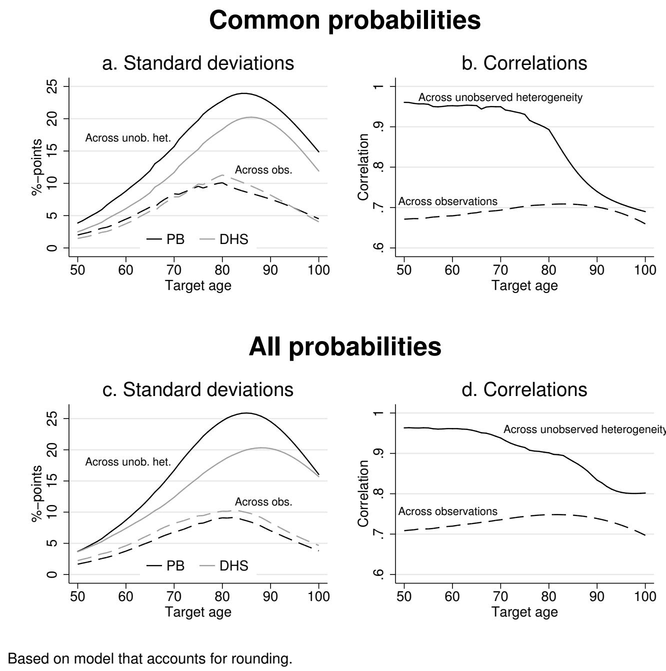



**Figure 4:** Variation in simulated expectations (left) and correlations between simulated probabilities (right)

Panels b. and d. of Figure 4 show test-retest correlations of simulated probabilities. Again we distinguish between the average (over observations) correlation across draws of unobserved heterogeneity and the correlation across observations of the average (over simulations) simulated probability. The correlation across observations is roughly constant around 0.7, which is stronger than the correlations between individual probabilities for all target ages when we pool men and women. Correlations across heterogeneity are above 0.9 at younger target ages and decline quickly after age 80 to approach the correlation across observations after age 90 for the estimates based on the probabilities that are common to both surveys (panel b.). In the extended sample based on all valid probabilities the correlation across individual effects remains higher and flattens out at 0.8. The high reliability of unobserved heterogeneity is reassuring since this is the dimension with larger variation according to panels a. and c.

The models show that cohort differences in the reliability of *levels* exist. Benefits of aggregation do arise, but only if we take into account all probabilities reported by an individual (raising the average number of probabilities from 1.9 to 4.4 per person/year). Level differences between surveys are small compared to the *variation* in expectations. This is true for variation with covariates and especially for unobserved heterogeneity, the latter being much stronger than the former. Aggregation does improve test re-test correlations. Unobserved heterogeneity is important and extremely reliable, with correlations in simulated probabilities in excess of 0.8 prior to target age 85.

# 6 Subjective longevity in a life cycle model

This section evaluates whether subjective life expectancy is sufficiently reliable to be used as input for the estimation of empirical life cycle models of saving and labor supply. We use the model from De Bresser et al. (2017) to map simulated probabilities into wealth and labor supply profiles. This allows us to quantify the consequences of variation in subjective probabilities between test and retest surveys, their *un*reliability, in terms of economic outcomes. We proceed in three steps. First, we use the estimates reported in Table 3 to simulate the probability of dying at a given age conditional on health. Three such sets of probabilities are computed: one for each survey and one based on actuarial tables adjusted for current health. We then use the model to link expectations to economic behavior. Preference are calibrated in two steps. Firstly, we fix the parameters of the utility function based on previous literature. Secondly, we calibrate the parameters that govern utility from bequests such that the model fits wealth quartiles and labor supply observed in the data. We compare the simulated wealth and labor supply obtained using the three sets of subjective probabilities. The following subsections describe these steps in turn.

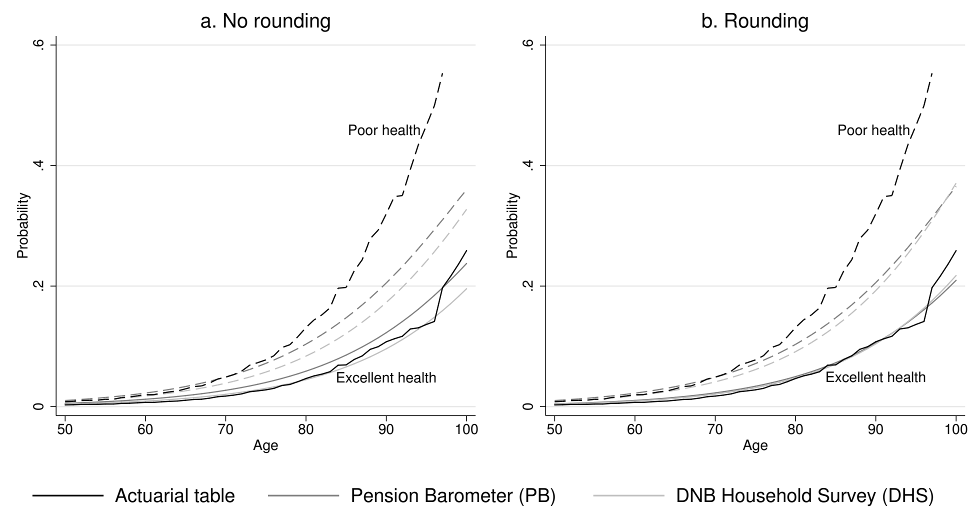



**Figure 5:** Simulated probabilities of dying at a given age conditional on survival up to that age

## 6.1 Simulating probabilities

The relevant input for a life cycle model is the probability of dying at a certain age conditional on having survived to that age. We simulate probabilities for a male born between 1952 and 1961; with a net household income between 1800 and 2600 euros/month; and with a medium level of education. Since the life cycle model uses a dichotomous measure of health, we alternatively fix health at "poor" and at "excellent". For covariates fixed at these values the relevant probabilities can be simulated from the estimates in Table 3 by integration over the distributions of individual and sequence effects. We use 10,000 draws of unobserved heterogeneity and simulate the relevant probabilities as the averages over those draws, using the same draws for all ages.

Figure 5 plots the simulated probabilities of dying at different ages conditional on surviving to those ages. The left panel uses estimates from the model without rounding, while the right panel simulates probabilities from the model that takes rounding into account. Both panels show that the probability of death increases with age and that the increase is markedly stronger for people in poor health. Death at age 50 is extremely unlikely regardless of current health, while above age 90 men in poor health face probabilities above 30% compared with 20% for those in excellent health.

Note that the differences between the PB and the DHS are larger for the model without rounding than for the model with rounding. This corroborates the pattern observed in panels a. and d. of Figure 3 and reflects the fact that differences between surveys in both parameters, hazard ratios and baseline hazards, cancel out. In the model without rounding the difference between baseline hazards is much smaller, so differences between the hazard ratios translate directly into differences between probabilities (as in panel a. of Figure 5).

In addition to the two sets of probabilities based on subjective survival, Figure 5 also shows analogous probabilities computed from life tables. We include these probabilities because they are the default way in which longevity risk is included in life cycle models. In particular, we adjust probabilities from the Human Mortality Database for current health using the method proposed in French (2005). This adjustment uses the Dutch sample from SHARE rather than the CentERpanel, since SHARE provides better follow-up after a panel member dies.[8] The probabilities that condition on good current health are close to those based on subjective data, for the model without rounding especially the DHS. However, the adjusted actuarial probabilities indicate a much higher likelihood of death when in poor health compared to either survey. While this discrepancy would seem to suggest that the actuarial figures yield much shorter lives, this is limited by the fact that relatively few individuals become unhealthy at young ages. Simulations based on these probabilities show that the average lifespan based on all three sets of probabilities are fairly close: the mean age of death is 79.0 according to the actuarial tables compared with 77.0 and 79.6 for the PB and DHS respectively.

[8]See http://www.share-project.org for more information on SHARE.

## 6.2 A life cycle model of saving and labor supply

We use the life cycle model proposed in De Bresser et al. (2017) to translate the probabilities reported in Figure 5 into saving and labor supply decisions. The model is specified to approximate the institutions in place in the Netherlands in 2011/2012. We present the main features here, a more detailed description is available on request.

The model is unitary: it assumes a single decision maker per household. It spans the age-range 50 to 100 with a resolution of one year. Every year agents decide how much to save. Up to age 70 agents also choose their labor supply (0, 1500, 2000 or 2500 hours of work per year). There are three exit routes from the labor market: disability insurance (DI), unemployment insurance (UI) and occupational pensions. Agents may decide to claim disability insurance as long as they are in bad health. Unemployment benefits can be claimed for a maximum of three years depending on one's work history (entitlements are accumulated at the rate of one month per year of work). The levels of unemployment and disability benefits are fixed at 70% of previous earnings. Moreover, both stop at age 65 when they are replaced by a flat-rate public pension. All 65-year olds receive the public pension regardless of their labor supply. If a worker's job includes an occupational pension plan the worker is obliged to participate. Occupational pension claiming can start at any age between 60 and 70 and benefits are a function of the number of years worked and final earnings. Occupational pension benefits are adjusted actuarially for the age at which they are first claimed: they are lowered by 7% per year for claiming prior to age 65 and raised analogously afterwards. Agents cannot work while they receive DI, UI or occupational pensions. In addition to paid employment and the four types of transfers mentioned above, additional income is provided by an exogenous income stream generated by the partner. The state variables included in the model are wealth, wage, a binary health variable and indicators for eligibility for DI, UI and occupational pensions.

There are three sources of uncertainty in the model: health, mortality and one's wage. Subjective survey information is only available for mortality, so uncertainty in health and wages is modeled from the transitions observed in the data. The probability of being in "excellent" health next period is a function of current health and age. We estimate this health process using the 2006-2016 waves of the DHS. The mortality processes consist of the probabilities reported in Figure 5. While of working age, agents face uncertainty in wages which we estimate from the Dutch sample of the European Community Household Panel (ECHP) following the methodology of Gourinchas and Parker (2002).

Agents with current age $t$ derive utility from consumption $c_t$ and leisure $l_t$ according to the following utility function:

$$u\left(c_t, l_t, t\right) = n_t \frac{\left(\left(\frac{c_t}{n_t}\right)^{\kappa} l_t^{1-\kappa}\right)^{1-\sigma} - 1}{1-\sigma} \tag{8}$$

$$\begin{aligned} l_t = 4000 - h_t - \xi \mathbb{I}\left\{h_t > 0\right\} - \delta \mathbb{I}\left\{\text{bad health}\right\} \\ - \phi \mathbb{I}\left\{di_t > 0\right\} - \zeta \mathbb{I}\left\{ui_t > 0\right\} \end{aligned}$$

$n_t$ is an equivalence scale that reflects family size and that decreases with age; $h_t$ is the number of hours worked; and all Greek letters denote parameters that are held constant in the simulations. The maximum amount of leisure is fixed at 4000 hrs/yr and bad health and claiming either UI or DI benefits carries stigma costs that are also measured in hrs/yr.

In addition to consumption and leisure, agents value leaving behind a bequest according to the following bequest utility function:

$$b\left(w_t\right) = \exp\left(\theta_0 + \theta_1 n_t\right) \frac{\left(w_t + K\right)^{\kappa(1-\sigma)}}{1-\sigma} \tag{9}$$

where $w_t$ is wealth at age $t$ and $K$ and the Greek symbols are parameters. The strength of this bequest motive varies with the size of the household as captured by the equivalence scale. This variation accounts for the idea that agents care more about leaving wealth to their partner than to other individuals outside the household, such as adult children. Men

**Table 5:** Values for preference parameters used in simulations

| Utility function | | Leisure costs (hrs/yr) | | Bequest utility | |
|---|---|---|---|---|---|
| $\sigma$ – concavity | 4.5 | $\xi$ – fixed cost of work | 850 | $\theta_0$ – constant | -8.9 |
| $\kappa$ – consumption share | 0.6 | $\delta$ – cost of poor health | 300 | $\theta_1$ – HH size coefficient | 7.1 |
| $\beta$ – discount factor | 0.97 | $\phi$ – stigma costs DI | 2000 | $K$ – bequest concavity ($) | 984,171 |
| | | $\zeta$ – stigma costs UI | 3500 | | |

who expect to outlive their partner care less about bequests than do men who expect to die before their spouse.

The values of the parameters of both felicity functions are chosen to be in line with previous work where possible. In particular, we fix all utility and leisure cost parameters based on estimates from the 1993-2001 waves of the DHS (see De Bresser, 2019). The parameters that govern utility from leaving a bequest are calibrated such that wealth quartiles and average hours worked are close to the DHS data for the waves of 2006-2016 (see Appendix H for details). This calibration is done for mortality probabilities derived from the PB using the measurement model without rounding. All parameter values used in the simulations are shown in Table 5. Sensitivity analysis indicates that the findings described in the next subsection regarding the consequences of the reliability of subjective survival are qualitatively robust to variation in preference parameters.

## 6.3 Subjective survival and economic behavior

We use the life cycle model described above to simulate wealth and labor supply for 5000 workers. Initial conditions for wages, wealth, labor supply and social insurance entitlements are taken from the DHS. The age profiles used to summarize simulations are those typically employed to estimate preference parameters. We compute wealth quartiles at two-year age bins between ages 50 and 70 and five-year bins for ages 70-84. Labor supply is summarized by the average yearly hours worked by two-year bins for ages 50 to 70.

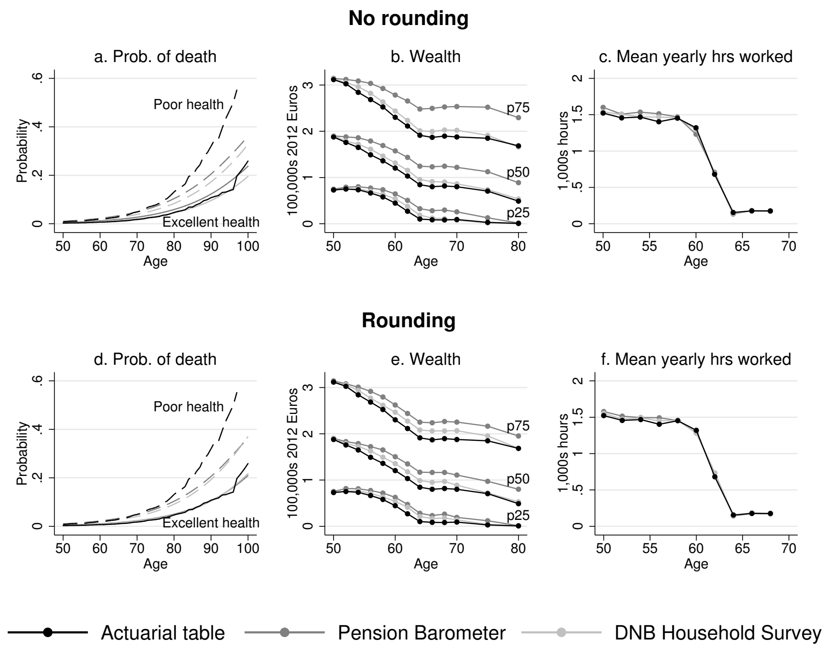


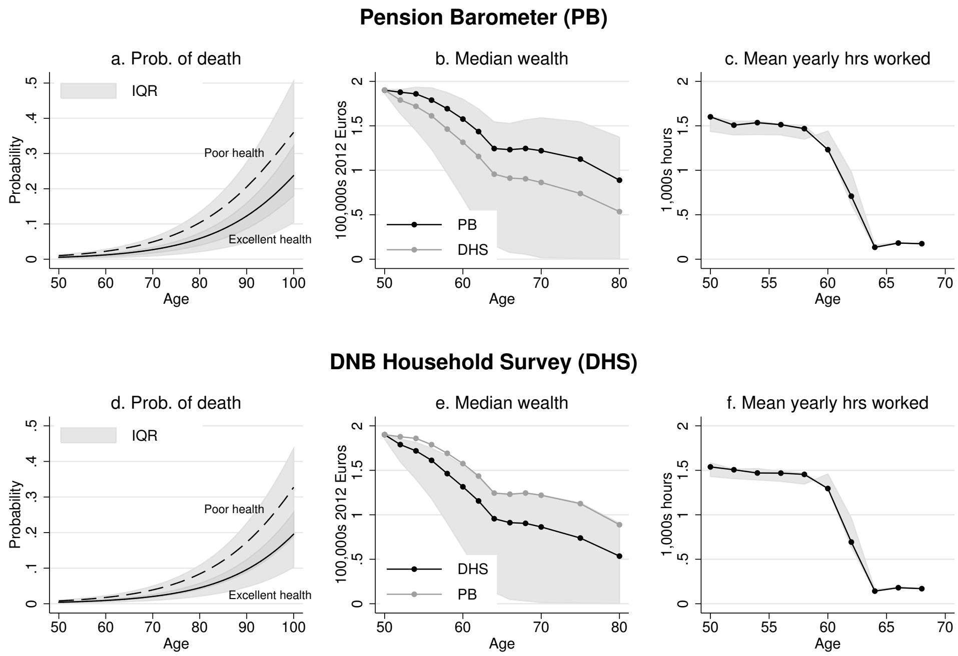

**Figure 6:** The impact of the reliability of subjective expectations on behavior in a life cycle model – average expectations

Figure 6 presents the simulations. Results for the model that does not account for rounding are shown in the top row while results for the model that does take rounding into account are in the bottom row. The leftmost panels a. and d. contain the mortality processes discussed in section 6.1 and shown in Figure 5. The middle column, panels b. and e., displays wealth quartiles. In line with Gan et al. (2015), we find that the level of wealth is sensitive to survival expectations. For the model without rounding, panel b., the difference between the medians simulated based on the PB and the DHS increases to 30,000-40,000 Euro after age 65 (around 30% of the PB profile). Simulated workers accumulate less wealth if we use the DHS mortality probabilities, despite the fact that they expect to live longer according to the DHS. This is due to the fact that the bequest motive is weaker at older ages, since household size declines with age. Men who expect to live longer have a weaker bequest motive and accumulate less wealth than those who expect to die younger. Life tables yield wealth profiles that are close to those based on DHS probabilities, because the probabilities conditional on good health are similar.

The model that does account for rounding leads to smaller differences in mortality expectations between the PB and DHS (see panel d.). It therefore also leads to smaller differences between simulated wealth profiles: the maximum difference between the medians in panel e. is around 20,000-25,000 Euros or 20% of the PB profile. Hence, our simulations indicate that while the level of wealth is sensitive to the set of mortality probabilities used, differences in simulated wealth are reduced substantially once we model rounding. This is primarily due to the PB: if we do not model rounding median wealth simulated using PB probabilities increases by up to 15%, while for the DHS rounding does not change median wealth by more than 5%.

Panels c. and f. show that labor supply is not sensitive to survival expectations. Regardless of whether we take rounding into account or not, the average hours worked per year is almost identical for the PB, DHS and actuarial tables. As a result the reliability of subjective expectations appears adequate if the main focus is on labor supply.

While differences in the levels of wealth and hours worked are interesting, the importance of unobserved heterogeneity suggests that the effect of variation in expectations on optimal behavior is at least as important. Figure 7 illustrates how such variation translates into differences in labor supply and saving. Panels a. and d. plot the interquartile range (IQR) of survival probabilities across the draws of unobserved heterogeneity at each target age. Differences between respondents that are not related to covariates affect subjective survival more strongly than does reported health: while poor health raises the probability of dying at a given age by 10pp around age 95, the IQRs for individuals in poor health and in excellent health are about twice as wide at that age and overlap at all ages. Given that unobserved heterogeneity is more important even than health, it should outweigh the differences between average expectations in the two surveys. Indeed, this is what the remaining panels




All simulations are based on estimates from model without rounding.

**Figure 7:** Heterogeneity in expectations and economic behavior

of Figure 7 indicate. The shaded areas in those panels correspond to simulations where we set expectations equal to the first and third quartiles of their distribution across unobserved heterogeneity. The average differences between surveys shown in Figure 6 are much smaller than the range of variation produced for each survey by individual effects. This underlines the point that heterogeneity in survival expectations can generate large differences in behavior for a given set of preferences. Moreover, it shows that such heterogeneity is far more important than differences between repeated measurements of expectations, even if those measurements use different answer scales. This is a novel piece of evidence indicating that the signal in expectations data outweighs the noise in terms of relevant economic implications. Given the strong correlation between individual effects across surveys, we conclude that subjective expectations are sufficiently reliable to enrich life cycle models.

# 7 Conclusion

A growing body of research recognizes the potential of data that directly elicits expectations of survey respondents, so-called subjective expectations, to enrich inter-temporal models. However, many economists remain sceptical of the reliability of such data. This paper investigates the validity of reported expectations by evaluating the test-retest reliability of the type of expectations that has received most attention from researchers: survival expectations. It shows that measured expectations are reliable even when elicited on different response scales. This supports the interpretation of answers on 11-point scales ranging from 0 to 10 as rounded probabilities.

Using two surveys that were administered to the same respondents within the same month, we compare the answers to items that ask for the likelihood of survival to various target ages. The questionnaires are the Pensioenbarometer (PB) and the DNB Household Survey (DHS), both of which were fielded in the CentERpanel, a household panel that is representative for the Dutch population. The PB allows respondents to report any integer probability between 0 and 100, while the DHS limits responses to an 11-point scale between 0 and 10. We first analyze reliability at the level of the reported probability by checking whether reported probabilities are consistent with each other one-by-one. We check whether the reported probabilities from both datasets are consistent with at least one underlying true probability under different degrees of rounding. We then analyze reported probabilities jointly and test whether the two surveys yield similar survival curves. This allows us to evaluate to what extent noise in the probabilities cancels out when those probabilities are aggregated. We compute survival probabilities from the model estimates to simulate labor supply and saving in a life cycle model and quantify the economic implications of test-retest variation in elicited expectations.

The response scales perform equally well in terms of non-response, internal consistency and 50/50s. Rounding is not related to education for either scale. We find the reliability of subjective survival expectations to be satisfactory. Test-retest correlations are in the 0.5-0.7 range, which is similar to the reliability of subjective well-being documented by Krueger and Schkade (2008). While around 20% of reported probabilities are equal in the PB and DHS, the fraction of consistent responses is much higher once we allow for rounding. Depending on the target age, 24-37% of reported probabilities are consistent if we assume that all PB probabilities are rounded to multiples of 1 and all DHS probabilities are rounded to multiples of 10. Common rounding as in Manski and Molinari (2010) raises the fraction of consistent probabilities to 32-46% and the most conservative degree of rounding for each reported probability increases it further to 61-77%. Better educated respondents are more likely to provide consistent responses on the two scales.

Joint models of all reported probabilities show that both datasets yield quantitatively and qualitatively similar associations between socio-demographic covariates and the hazard of death. The largest differences between the estimates occur for cohort dummies. Older individuals report a relatively high likelihood of surviving past the oldest target ages when forced to round in the DHS, but not on the more elaborate response scale of the PB. Other variables such as gender, income, education and self-assessed health enter the model in similar ways for both datasets. We find that unobserved heterogeneity at the level of the individual is important and that this heterogeneity is strongly positively correlated across questionnaires. As for the benefits relative to individual probabilities, aggregation improves reliability of *levels* only when we use all available probabilities for each person/year (which corresponds to 4 rather than 2 probabilities on average). It improves reliability of *variation*, i.e. correlation between individual effects, regardless of the set of probabilities considered, as evidenced by higher test-retest correlations across observations than were found in the raw data.

We simulate saving and labor supply in a calibrated life cycle model using survival curves constructed from the estimates of joint models of PB and DHS probabilities as well as life tables. The model with rounding yields more reliable survival curves than the model that does not take rounding into account. Saving is sensitive to survival expectations, to the extent that the variation in survival curves leads to a substantially different level of wealth if we construct curves from the estimates for the model without rounding (the difference between median wealth simulated based on the PB and DHS is around 30%, with simulations based on life tables close to the DHS). However, the model that does take rounding into account substantially reduces the difference between wealth profiles to around 20%. This improvement reflects the influence of rounding for the finer percentage scale. Moreover, differences between wealth profiles constructed at average probabilities are small relative to individuals effects, which is important since individual effects are reliable across surveys. Labor supply is less sensitive to survival expectations than is wealth.

Taking all results together we conclude that when probabilities are considered in isolation, the quality of subjective survival is comparable to that of other types of subjective data, such as subjective well-being. Aggregation of probabilities into individual-specific survival curves improves reliability, especially when rounding is taken into account. Within-individual variation is both quantitatively less important and less reliable than variation between individuals, so applied researchers are advised not to focus exclusively on the former. While this finding is plausible in the context of survival, within-variation may well be more reliable in different contexts in which expectations are revised frequently, such as when learning plays an important role. When aggregated into survival curves these data can be used to enrich inter-temporal models. Such models should account for heterogeneity in survival expectations, since that strongly affects saving before and during retirement. Answers on the 11-point response scale that is often used to elicit expectations can be interpreted as rounded probabilities.

# References


Bissonnette, L. and de Bresser, J. (2018). Eliciting subjective survival curves: Lessons from partial identification. *Journal of Business & Economic Statistics*, 36(3):505–515.

Bissonnette, L., Hurd, M. D., and Michaud, P.-C. (2017). Individual survival curves comparing subjective and observed mortality risks. *Health economics*, 26(12):285–303.

Blundell, R., French, E., and Tetlow, G. (2016). Retirement incentives and labor supply. In *Handbook of the economics of population aging*, volume 1, pages 457–566. Elsevier.

Bruine de Bruin, W. and Carman, K. G. (2012). Measuring risk perceptions: what does the excessive use of 50% mean? *Medical Decision Making*, 32(2):232–236.

Bruine de Bruin, W., Fischbeck, P. S., Stiber, N. A., and Fischhoff, B. (2002). What number is fifty-fifty?: Redistributing excessive 50% responses in elicited probabilities. *Risk Analysis*, 22(4):713–723.

Bruine de Bruin, W., Fischhoff, B., Millstein, S. G., and Halpern-Felsher, B. L. (2000). Verbal and numerical expressions of probability: "It's a fifty-fifty chance". *Organizational behavior and human decision processes*, 81(1):115–131.

De Bresser, J. (2019). Evaluating the accuracy of counterfactuals – the role of heterogeneous expectations in life cycle models. Mimeo Tilburg University.

De Bresser, J., Fonseca, R., and Michaud, P.-C. (2017). Retirement behavior in the U.S. and Europe. Netspar Academic Series, DP 11/2017-018.

De Bresser, J. and Van Soest, A. (2013). Survey response in probabilistic questions and its impact on inference. *Journal of Economic Behavior and Organization*, 96(December 2013):65–84.

De Nardi, M., French, E., and Jones, J. B. (2010). Why do the elderly save? the role of medical expenditures. *Journal of Political Economy*, 118(1):39–75.

De Nardi, M., French, E., and Jones, J. B. (2016). Savings after retirement: A survey. *Annual Review of Economics*, 8:177–204.

Delavande, A. (2008). Pill, patch, or shot? subjective expectations and birth control choice. *International Economic Review*, 49(3):999–1042.

French, E. (2005). The effects of health, wealth and wages on labour supply and retirement behaviour. *Review of Economic Studies*, 72(2):395–427.

French, E. and Jones, J. B. (2011). The effects of health insurance and self-insurance on retirement behavior. *Econometrica*, 79(3):693–732.

Gan, L., Gong, G., Hurd, M., and McFadden, D. (2015). Subjective mortality risk and bequests. *Journal of Econometrics*, 188(2):514–525.

Gourinchas, P.-O. and Parker, J. A. (2002). Consumption over the life cycle. *Econometrica*, 70(1):47–89.

Heimer, R., Myrseth, K. O. R., and Schoenle, R. (2018). Yolo: Mortality beliefs and household finance puzzles. Forthcoming in *Journal of Finance*.

Hudomiet, P., Hurd, M. D., and Rohwedder, S. (2018). Measuring probability numeracy. RAND Labor and Population Working Paper.

Hurd, M. (2009). Subjective probabilities in household surveys. *Annual Economic Review*, 1:543–562.

Hurd, M. and McGarry, K. (1995). Evaluation of the subjective probabilities of survival in the Health and Retirement Study. *The Journal of Human Resources*, 30(Special Issue on the Health and Retirement Study: Data Quality and Early Results):S268–S292.

Hurd, M. and McGarry, K. (2002). The predictive validity of subjective probabilities of survival. *The Economic Journal*, 112(482):966–985.

Kleinjans, C. and Van Soest, A. (2014). Non-response and focal point answers to subjective probability questions. *Journal of Applied Econometrics*, 29(4):567–585.

Krueger, A. B. and Schkade, D. A. (2008). The reliability of subjective well-being measures. *Journal of Public Economics*, 92(8-9):1833–1845.

Kutlu-Koc, V. and Kalwij, A. (2017). Individual survival expectations and actual mortality: Evidence from dutch survey and administrative data. *European Journal of Population*, 33(4):509–532.

Manski, C. F. (2004). Measuring expectations. *Econometrica*, 75(5):1329–1376.

Manski, C. F. and Molinari, F. (2010). Rounding probabilistic expectations in surveys. *Journal of Business and Economic Statistics*, 28(2):219–231.

Payne, J. W., Sagara, N., Shu, S. B., Appelt, K. C., and Johnson, E. J. (2012). Life expectancy as a constructed belief: Evidence of a live-to or die-by framing effect. *Journal of Risk and Uncertainty*, 46(1):27–50.

Perozek, M. (2008). Using subjective expectations to forecast longevity: do survey respondents know something we don't know? *Demography*, 45(1):95–113.

Smith, V. K., D. H. Taylor, J., and Sloan, F. A. (2001). Longevity expectations and death: can people predict their own demise? *American Economic Review*, 91(4):1126–1134.

Stinebrickner, R. and Stinebrickner, T. R. (2013). A major in science? Initial beliefs and final outcomes for college major and dropout. *Review of Economic Studies*, 81(1):426–472.

Teppa, F., Thorp, S., and Bateman, H. (2015). Family, friends and framing: A cross-country study of subjective survival expectations. De Nederlandsche Bank Working Paper.

Van der Klaauw, W. (2012). On the use of expectations data in estimating structural dynamic choice models. *Journal of Labor Economics*, 30(3):521–554.

Van der Klaauw, W. and Wolpin, K. I. (2008). Social security and the retirement and savings behavior of low-income households. *Journal of Econometrics*, 145(1-2):21–42.

Wu, S., Stevens, R., and Thorp, S. (2015). Cohort and target age effects on subjective survival probabilities: Implications for models of the retirement phase. *Journal of Economic Dynamics and Control*, 55(June 2015):39–56.

Zafar, B. (2011). How do college students form expectations? *Journal of Labor Economics*, 29(2):301–348.

# A Incidence of 50s

**Table A1:** Incidence of 50s

| | Pension Barometer (PB) | | DNB Household Survey (DHS) | |
|---|---|---|---|---|
| Target age | N | Fraction | N | Fraction |
| 65 | – | – | 1,504 | 0.08 |
| 70 | 2,194 | 0.14 | – | – |
| 75 | 2,457 | 0.13 | 2,335 | 0.14 |
| 80 | 2,633 | 0.18 | 2,724 | 0.19 |
| 85 | 2,742 | 0.16 | 630 | 0.22 |
| 90 | 2,775 | 0.12 | 394 | 0.18 |
| 95 | – | – | 260 | 0.13 |
| 100 | – | – | 127 | 0.06 |
| Overall | 12,801 | 0.15 | 7,974 | 0.15 |
| All 50s[a] | 2,775 | 0.03 | 3,245 | 0.04 |

These numbers are based on the separate samples of complete and internally consistent responses to the relevant survey. They do not limit the sample to the intersection of valid answers to both surveys.

[a] Equal to 1 if individual only reported 50s in a given survey wave, 0 otherwise.

# B Data quality for different levels of education

**Table B1:** Response rates by level of education

| | Pension Barometer (PB) | DNB Household Survey (DHS) |
|---|---|---|
| | Complete response | Complete response |
| Educ.: lower secondary | 0.92 | 0.89 |
| Educ.: higher secondary/vocational | 0.97 | 0.93 |
| Educ.: (applied) university | 0.96 | 0.92 |
| Overall | 0.95 | 0.91 |

These numbers are based on the separate samples for the relevant survey. They do not limit the sample to the intersection of observations to which both surveys were offered.

**Table B2:** Rates of internally consistent response by level of education (conditional on complete response)

| | Pension Barometer (PB) | DNB Household Survey (DHS) |
|---|---|---|
| | Consistent response | Consistent response |
| Educ.: lower secondary | 0.98 | 0.99 |
| Educ.: higher secondary/vocational | 0.98 | 0.99 |
| Educ.: (applied) university | 0.99 | 0.99 |
| Overall | 0.98 | 0.99 |

These numbers are based on the separate samples of complete responses to the relevant survey. They do not limit the sample to the intersection of complete answers to both surveys.

**Table B3:** Incidence of 50s by education level

| | Pension Barometer (PB)[b] | | | DNB Household Survey (DHS)[b] | | |
|---|---|---|---|---|---|---|
| Target age | Lower secondary | Higher sec./ vocational | University | Lower secondary | Higher sec./ vocational | University |
| 65 | – | – | – | 0.11 | 0.10 | 0.06*** |
| 70 | 0.14 | 0.17 | 0.12 | – | – | – |
| 75 | 0.12 | 0.14 | 0.13 | 0.15 | 0.16 | 0.11** |
| 80 | 0.18 | 0.18 | 0.19 | 0.19 | 0.21 | 0.19 |
| 85 | 0.15 | 0.17 | 0.16 | 0.18 | 0.30** | 0.20 |
| 90 | 0.12 | 0.14 | 0.12 | 0.17 | 0.22 | 0.18 |
| 95 | – | – | – | 0.10 | 0.23* | 0.12 |
| 100 | – | – | – | 0.03 | 0.07 | 0.07 |
| Overall | 0.14 | 0.16 | 0.15 | 0.16 | 0.17 | 0.14** |
| All 50s[a] | 0.02 | 0.03 | 0.02 | 0.04 | 0.04 | 0.04 |

These numbers are based on the separate samples of complete and internally consistent responses to the relevant survey. They do not limit the sample to the intersection of valid answers to both surveys.

[a] Equal to 1 if individual only reported 50s in a given survey wave, 0 otherwise.

[b] Table entries are fractions of 50s in the sample. Stars refer to tests of the null that the fraction of 50s is the same as for those with *lower secondary* education (the baseline); $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$.

# C Descriptive statistics of covariates

**Table C1:** Descriptive statistics

| | Sample: intersection valid probs. in PB *and* DHS[a] | | Sample: union of valid probs. in PB *or* DHS[b] | | Sample: all potential observations[c] | |
|---|---|---|---|---|---|---|
| | Mean | Std. dev. | Mean | Std. dev. | Mean | Std. dev. |
| Coh. 1922-1931 | – | – | 0.04 | 0.19 | 0.04 | 0.19 |
| Coh. 1932-1941 | 0.13 | 0.34 | 0.13 | 0.33 | 0.13 | 0.34 |
| Coh. 1942-1951 | 0.28 | 0.45 | 0.27 | 0.45 | 0.27 | 0.44 |
| Coh. 1952-1961 | 0.24 | 0.43 | 0.23 | 0.42 | 0.23 | 0.42 |
| Coh. 1962-1971 | 0.21 | 0.41 | 0.18 | 0.39 | 0.18 | 0.39 |
| Coh. 1972-1981 | 0.11 | 0.32 | 0.13 | 0.34 | 0.13 | 0.34 |
| Coh. 1982-1987 | 0.02 | 0.14 | 0.02 | 0.14 | 0.02 | 0.15 |
| Wave 2012 | 0.48 | 0.50 | 0.51 | 0.50 | 0.51 | 0.50 |
| Female | 0.43 | 0.50 | 0.44 | 0.50 | 0.44 | 0.50 |
| Net HH. inc. $\leq$ €1150 | 0.06 | 0.24 | 0.08 | 0.27 | 0.08 | 0.27 |
| Net HH. inc. €1151-1800 | 0.16 | 0.36 | 0.15 | 0.36 | 0.15 | 0.35 |
| Net HH. inc. €1801-2600 | 0.28 | 0.45 | 0.24 | 0.43 | 0.24 | 0.43 |
| Net HH. inc. $\geq$ €2601 | 0.51 | 0.50 | 0.53 | 0.50 | 0.53 | 0.50 |
| Educ. lower secondary | 0.29 | 0.45 | 0.30 | 0.46 | 0.30 | 0.46 |
| Educ. higher secondary/vocational | 0.30 | 0.46 | 0.28 | 0.45 | 0.28 | 0.45 |
| Educ. (applied) university | 0.42 | 0.49 | 0.41 | 0.49 | 0.41 | 0.49 |
| Health: excellent | 0.14 | 0.34 | 0.14 | 0.34 | 0.14 | 0.34 |
| Health: good | 0.63 | 0.48 | 0.62 | 0.49 | 0.62 | 0.49 |
| Health: fair | 0.17 | 0.37 | 0.18 | 0.38 | 0.18 | 0.38 |
| Health: not good/poor | 0.07 | 0.26 | 0.07 | 0.25 | 0.07 | 0.25 |
| N (individuals) | 1,470 | | 2,323 | | 2,353 | |
| N (individual-years) | 2,073 | | 3,787 | | 3,840 | |

[a] This sample is used in the one-by-one comparison of PB and DHS probabilities and to estimate the joint model of survival probabilities reported in the main text and in Appendix E.
[b] This sample is used to estimate the joint model of survival probabilities based on all valid probabilities reported in Appendix F.
[c] All person-years that were offered the survival items in at least one survey. This sample drops 162 person-year observations based on missing background variables.

# D Distribution of rounding

**Table D1a:** Common rounding

| | PB (frac.) | DHS (frac.) |
|---|---|---|
| All 0 or 100 | 0.01 | 0.03 |
| All 0, 50 or 100 | 0.03 | 0.03 |
| All multiples of 10 | 0.25 | 0.94 |
| All multiples of 5 | 0.57 | |
| Some in $[1, 4]$ or $[96, 100]$ | 0.10 | |
| Other | 0.05 | |

$N = 1,549$ individuals (sample limited to the intersection of valid responses to both surveys)

**Table D1b:** General rounding

| Multiples of... | PB (frac.) | DHS (frac.) |
|---|---|---|
| ...100 | 0.08 | 0.07 |
| ...50 | 0.16 | 0.17 |
| ...25 | 0.09 | |
| ...10 | 0.52 | 0.76 |
| ...5 | 0.12 | |
| ...1 | 0.03 | |

$N = 4,062$ probabilities (sample limited to intersection of valid responses to both surveys)

# E Estimates for recall error and rounding

## *(for online publication)*

**Table E1:** Recall error and rounding estimates of Gompertz models of subjective survival

| | Model 1 – No rounding | | Model 2 – Rounding | | | |
|---|---|---|---|---|---|---|
| | Error PB | Error DHS | Error PB | Error DHS | Rounding PB | Rounding DHS[a] |
| Wave 2012 | -0.0281 | 0.000787 | -0.251*** | 0.0517 | 0.0325 | 1.237 |
| | (0.0468) | (0.0395) | (0.0566) | (0.0442) | (0.0613) | |
| Female | -0.0765** | 0.0226 | 0.0413 | 0.0916** | 0.0326 | -0.746 |
| | (0.0373) | (0.0344) | (0.0428) | (0.0371) | (0.0710) | |
| *Cohorts (baseline: 1942-1951)* | | | | | | |
| Coh. 1932-41 | 0.216*** | 0.198*** | 0.151* | 0.372*** | -0.0583 | -7.856 |
| | (0.0677) | (0.0619) | (0.0771) | (0.0610) | (0.128) | |
| Coh. 1952-61 | 0.0863 | -0.0411 | 0.0726 | -0.0269 | -0.000522 | -2.117 |
| | (0.0543) | (0.0482) | (0.0555) | (0.0496) | (0.0949) | |
| Coh. 1962-71 | 0.0211 | -0.142*** | 0.0434 | -0.0521 | -0.0438 | -0.720 |
| | (0.0533) | (0.0525) | (0.0604) | (0.0531) | (0.103) | |
| Coh. 1972-81 | 0.201*** | 0.102* | 0.123* | -0.0788 | -0.0763 | -0.00929 |
| | (0.0661) | (0.0613) | (0.0732) | (0.0641) | (0.120) | |
| Coh. 1982-87 | 0.155 | -0.194 | -0.296 | -0.605*** | 0.152 | 1.488 |
| | (0.130) | (0.129) | (0.189) | (0.196) | (0.234) | |
| *Net household income (baseline: more than €2600* | | | | | | |
| Net HH. Inc. $\leq$ €1150 | -0.0227 | 0.292*** | 0.146 | 0.208** | -0.156 | 2.450 |
| | (0.0791) | (0.0847) | (0.0979) | (0.0829) | (0.153) | |
| Net HH. Inc. €1151-1800 | 0.0289 | 0.00558 | 0.251*** | -0.0297 | -0.185* | 3.527 |
| | (0.0526) | (0.0515) | (0.0604) | (0.0557) | (0.104) | |
| Net HH. Inc. €1801-2600 | 0.0596 | -0.0724* | 0.237*** | -0.140*** | -0.0218 | 1.044 |
| | (0.0444) | (0.0411) | (0.0522) | (0.0464) | (0.0814) | |
| *Education (baseline: lower secondary)* | | | | | | |
| Educ. higher sec./vocational | -0.123** | -0.180*** | -0.146** | -0.272*** | 0.0348 | -1.607 |
| | (0.0541) | (0.0485) | (0.0574) | (0.0495) | (0.0966) | |
| Educ. (applied) university | -0.247*** | -0.152*** | -0.259*** | -0.221*** | -0.0714 | 0.448 |
| | (0.0503) | (0.0451) | (0.0525) | (0.0452) | (0.0930) | |
| *Health (baseline: excellent)* | | | | | | |
| Health: good | 0.0830 | -0.0962 | 0.0838 | -0.0350 | 0.00927 | 1.190 |
| | (0.0853) | (0.0643) | (0.0682) | (0.0571) | (0.115) | |
| Health: fair | 0.174* | 0.00757 | 0.209*** | 0.167** | -0.311** | -4.533 |
| | (0.0990) | (0.0748) | (0.0802) | (0.0678) | (0.138) | |
| Health: not good/poor | 0.0333 | 0.314*** | 0.201* | 0.101 | -0.148 | 5.886 |
| | (0.121) | (0.100) | (0.104) | (0.0901) | (0.176) | |
| Constant | 2.363*** | 2.586*** | 2.105*** | 2.337*** | | |
| | (0.118) | (0.0795) | (0.0894) | (0.0718) | | |
| $\mu_1$ | | | | | -2.491*** | 13.982 |
| | | | | | (0.169) | |
| $\mu_2$ | | | | | -0.584*** | 18.709 |
| | | | | | (0.167) | |
| $\mu_3$ | | | | | 1.154*** | |
| | | | | | (0.182) | |
| $\mu_4$ | | | | | 1.973*** | |
| | | | | | (0.199) | |
| $\mu_5$ | | | | | 3.088*** | |
| | | | | | (0.240) | |
| Variance ind. effects | | | | | 0.693*** | 56.865 |
| | | | | | (0.0995) | |
| Variance seq. effects | | | | | 0.0284* | 10.393 |
| | | | | | (0.0157) | |
| No. individuals | 1,470 | | 1,470 | | | |
| No. probabilities | 4,034 | | 4,034 | | | |
| Log-likelihood | -30,530.175 | | -16,048.925 | | | |

[a] Standard errors could not be calculated for the rounding equation in the DHS due to the likelihood being flat in those dimensions. These numerical problems indicate that DHS probabilities are almost exclusively rounded to the minimum degree possible, which is to multiples of 10. We fix all parameters of the DHS rounding process to the values given in the table and calculate SEs for the remaining parameters.
Standard errors in parentheses; $***p < 0.01$, $**p < 0.05$, $*p < 0.1$

**Table E2:** Correlation matrices of individual and question sequence effects

| | PB | DHS | Round PB | Round DHS |
|---|---|---|---|---|
| **a. Individual effects** | | | | |
| PB | 1 | | | |
| DHS | 0.787*** | 1 | | |
| Round PB | -0.0352 | 0.0969** | 1 | |
| Round DHS[a] | -0.201 | 0.156 | 0.869 | 1 |
| **b. Sequence effects** | | | | |
| | PB | DHS | Round PB | Round DHS |
| PB | 1 | | | |
| DHS | 0.239*** | 1 | | |
| Round PB | 0.386 | -0.775*** | 1 | |
| Round DHS[a] | 0.386 | -0.357 | 0.645 | 1 |

[a] Standard errors have not been calculated for correlations involving unobserved heterogeneity in rounding equation for DHS, because standard errors could not be computed for those elements of Cholesky matrix that pertain to rounding in the DHS.

***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

# F Estimates based on all valid probabilities

## *(for online publication)*

**Table F1:** Gompertz model of subjective survival – estimates based on all valid probabilities

| | Model 1 – No rounding | | | Model 2 – Rounding | | |
|---|---|---|---|---|---|---|
| | PB[a] | DHS[a] | Diff. PB - DHS | PB[a] | DHS[a] | Diff. PB - DHS |
| **a. Hazard ratios** | | | | | | |
| Wave 2012 | 1.012 | 1.008 | 0.00365 | 1.022 | 1.035*** | -0.0128 |
| | (0.0186) | (0.0161) | (0.0212) | (0.0143) | (0.0122) | (0.0176) |
| Female | 1.013 | 1.028 | -0.0152 | 0.852*** | 0.890*** | -0.0375* |
| | (0.0267) | (0.0253) | (0.0273) | (0.0166) | (0.0178) | (0.0198) |
| *Cohorts (baseline: 1942-1951)* | | | | | | |
| Coh. 1922-31 | 1.145* | 1.171* | -0.0265 | 0.985 | 0.971 | 0.0147 |
| | (0.0932) | (0.0999) | (0.104) | (0.0709) | (0.0768) | (0.0817) |
| Coh. 1932-41 | 1.053 | 1.136** | -0.0825 | 0.962 | 1.057 | -0.0948** |
| | (0.0612) | (0.0634) | (0.0518) | (0.0371) | (0.0420) | (0.0447) |
| Coh. 1952-61 | 1.036 | 1.072** | -0.0365 | 0.926*** | 1.103*** | -0.177*** |
| | (0.0326) | (0.0360) | (0.0378) | (0.0219) | (0.0275) | (0.0294) |
| Coh. 1962-71 | 0.928** | 0.997 | -0.0696** | 0.925** | 1.123*** | -0.198*** |
| | (0.0322) | (0.0363) | (0.0340) | (0.0305) | (0.0336) | (0.0318) |
| Coh. 1972-81 | 0.777*** | 0.869** | -0.0920** | 1.067** | 1.179*** | -0.112*** |
| | (0.0561) | (0.0531) | (0.0371) | (0.0344) | (0.0352) | (0.0385) |
| Coh. 1982-87 | 1.114 | 0.951 | 0.163 | 1.216*** | 1.060 | 0.156** |
| | (0.155) | (0.108) | (0.104) | (0.0810) | (0.0491) | (0.0730) |
| *Net household income (baseline: more than €2600)* | | | | | | |
| Net HH. Inc. $\leq$ €1150 | 1.108 | 1.031 | 0.0771 | 1.009 | 1.062 | -0.0532 |
| | (0.0766) | (0.0507) | (0.0737) | (0.0580) | (0.0414) | (0.0607) |
| Net HH. Inc. €1151-1800 | 1.046 | 0.940* | 0.107*** | 1.024 | 0.990 | 0.0340 |
| | (0.0410) | (0.0333) | (0.0384) | (0.0370) | (0.0283) | (0.0379) |
| Net HH. Inc. €1801-2600 | 1.039 | 0.966 | 0.0736** | 1.017 | 1.025 | -0.00787 |
| | (0.0273) | (0.0236) | (0.0295) | (0.0223) | (0.0216) | (0.0265) |
| *Education (baseline: lower secondary)* | | | | | | |
| Educ. higher sec./vocational | 0.852*** | 0.824*** | 0.0288 | 0.908*** | 0.904*** | 0.00386 |
| | (0.0356) | (0.0321) | (0.0283) | (0.0239) | (0.0220) | (0.0266) |
| Educ. (applied) university | 0.995 | 1.017 | -0.0219 | 0.883*** | 0.912*** | -0.0288 |
| | (0.0312) | (0.0318) | (0.0301) | (0.0160) | (0.0204) | (0.0226) |
| *Health (baseline: excellent)* | | | | | | |
| Health: good | 1.212*** | 1.210*** | 0.00245 | 1.361*** | 1.230*** | 0.131*** |
| | (0.0387) | (0.0397) | (0.0357) | (0.0236) | (0.0266) | (0.0322) |
| Health: fair | 1.719*** | 1.618*** | 0.101 | 1.975*** | 1.613*** | 0.362*** |
| | (0.0680) | (0.0657) | (0.0655) | (0.0606) | (0.0485) | (0.0673) |
| Health: not good/poor | 2.044*** | 1.858*** | 0.186 | 2.183*** | 1.817*** | 0.366*** |
| | (0.116) | (0.105) | (0.114) | (0.0885) | (0.0833) | (0.103) |
| Constant | 0.00310*** | 0.0222*** | -0.0191*** | 0.00307*** | 0.0188*** | -0.0157*** |
| | (0.000103) | (0.00157) | (0.00160) | (8.70e-05) | (0.000815) | (0.000820) |
| Chi2 test joint equality (17df) | 203.25 ($p < 0.0001$) | | | 577.95 ($p < 0.0001$) | | |
| Chi2 test joint equality no cohorts (11df) | 161.24 ($p < 0.0001$) | | | 411.33 ($p < 0.0001$) | | |
| **b. Other estimates** | | | | | | |
| Baseline hazard ($t/100$) | 9.091*** | 6.211*** | 2.880*** | 9.123*** | 6.480*** | 2.643*** |
| | (0.0680) | (0.0690) | (0.114) | (0.0344) | (0.0667) | (0.0742) |
| Variance ind. effects | 0.809*** | 0.505*** | | 0.850*** | 0.437*** | |
| | (0.0413) | (0.0312) | | (0.0256) | (0.0159) | |
| Corr. ind. effects | 0.834*** | | | 0.781*** | | |
| | (0.0393) | | | (0.0115) | | |
| Variance seq. effects | 0.106*** | 0.0350*** | | 0.104*** | 0.0234*** | |
| | (0.00647) | (0.0131) | | (0.00457) | (0.00346) | |
| Corr. seq. effects | 0.442*** | | | 0.604*** | | |
| | (0.123) | | | (0.0644) | | |
| Fraction var. ind. effects | 0.884*** | 0.935*** | | 0.891*** | 0.949*** | |
| | (0.00696) | (0.0214) | | (0.00521) | (0.00759) | |
| No. individuals | 2,323 | | | 2,323 | | |
| No. probabilities | 16,540 | | | 16,540 | | |
| Log-likelihood | -74,126.826 | | | -40,588.262 | | |

[a] Estimates reported as hazard ratios.
Standard errors in parentheses; ***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

**Table F2:** Gompertz model of subjective survival – estimates based on all valid probabilities

| | Model 1 – No rounding | | Model 2 – Rounding | | | |
|---|---|---|---|---|---|---|
| | Error PB | Error DHS | Error PB | Error DHS | Rounding PB | Rounding DHS[a] |
| Wave 2012 | -0.0221 | 0.0484* | -0.0514** | 9.60e-05 | -0.0330 | 0.303 |
| | (0.0184) | (0.0257) | (0.0232) | (0.0293) | (0.0334) | |
| Female | 0.0512*** | -0.00434 | 0.0316 | 0.0325 | 0.0503 | -0.683 |
| | (0.0169) | (0.0236) | (0.0207) | (0.0256) | (0.0404) | |
| *Cohorts (baseline: 1942-1951)* | | | | | | |
| Coh. 1922-31 | -0.105 | 0.348*** | -0.166* | 0.368*** | 0.108 | -2.859 |
| | (0.0740) | (0.0825) | (0.0853) | (0.0853) | (0.151) | |
| Coh. 1932-41 | -0.0216 | 0.251*** | -0.0265 | 0.311*** | 0.101 | -2.297 |
| | (0.0291) | (0.0519) | (0.0356) | (0.0484) | (0.0707) | |
| Coh. 1952-61 | 0.0594*** | -0.0159 | 0.0212 | -0.0638 | 0.0162 | 1.0670 |
| | (0.0225) | (0.0412) | (0.0268) | (0.0390) | (0.0550) | |
| Coh. 1962-71 | -0.0260 | -0.0391 | -0.0280 | -0.0604 | -0.0232 | 0.824 |
| | (0.0240) | (0.0414) | (0.0297) | (0.0402) | (0.0590) | |
| Coh. 1972-81 | 0.0931*** | 0.0569 | 0.0589 | -0.161*** | 0.0407 | 1.403 |
| | (0.0294) | (0.0441) | (0.0360) | (0.0444) | (0.0690) | |
| Coh. 1982-87 | 0.0260 | -0.139 | -0.266*** | -0.435*** | 0.0630 | 1.005 |
| | (0.0593) | (0.0903) | (0.0758) | (0.0926) | (0.126) | |
| *Net household income (baseline: more than €2600)* | | | | | | |
| Net HH. Inc. $\leq$ €1150 | 0.148*** | 0.204*** | 0.179*** | 0.140*** | 0.0804 | 1.046 |
| | (0.0358) | (0.0488) | (0.0460) | (0.0538) | (0.0877) | |
| Net HH. Inc. €1151-1800 | 0.0896*** | 0.105*** | 0.168*** | 0.0291 | -0.117** | 1.399 |
| | (0.0251) | (0.0361) | (0.0303) | (0.0391) | (0.0586) | |
| Net HH. Inc. €1801-2600 | 0.0350* | 0.0124 | 0.0854*** | -0.0569* | -0.0733 | 0.623 |
| | (0.0198) | (0.0288) | (0.0248) | (0.0319) | (0.0473) | |
| *Education (baseline: lower secondary)* | | | | | | |
| Educ. higher sec./vocational | -0.0410* | -0.0833** | -0.0990*** | -0.117*** | 0.0967* | 1.108 |
| | (0.0223) | (0.0352) | (0.0279) | (0.0340) | (0.0547) | |
| Educ. (applied) university | -0.192*** | -0.154*** | -0.224*** | -0.126*** | 0.0119 | -0.00616 |
| | (0.0208) | (0.0338) | (0.0263) | (0.0317) | (0.0532) | |
| *Health (baseline: excellent)* | | | | | | |
| Health: good | 0.00435 | 0.00544 | 0.0157 | 0.0328 | 0.0277 | -0.0478 |
| | (0.0261) | (0.0488) | (0.0313) | (0.0387) | (0.0582) | |
| Health: fair | 0.0297 | 0.148*** | 0.0930** | 0.272*** | -0.128* | -1.461 |
| | (0.0316) | (0.0564) | (0.0384) | (0.0469) | (0.0721) | |
| Health: not good/poor | 0.0230 | 0.250*** | 0.0413 | 0.335*** | -0.0670 | -1.112 |
| | (0.0413) | (0.0734) | (0.0522) | (0.0702) | (0.0974) | |
| Constant | 2.550*** | 2.479*** | 2.404*** | 2.311*** | | |
| | (0.0331) | (0.0767) | (0.0408) | (0.0523) | | |
| $\mu_1$ | | | | | -1.985*** | 6.744 |
| | | | | | (0.0854) | |
| $\mu_2$ | | | | | -0.374*** | 10.052 |
| | | | | | (0.0855) | |
| $\mu_3$ | | | | | 1.271*** | |
| | | | | | (0.0914) | |
| $\mu_4$ | | | | | 1.981*** | |
| | | | | | (0.101) | |
| $\mu_5$ | | | | | 3.124*** | |
| | | | | | (0.134) | |
| Variance ind. effects | | | | | 0.440*** | 14.194 |
| | | | | | (0.0476) | |
| Variance seq. effects | | | | | 0.0253*** | 0.467 |
| | | | | | (0.00813) | |
| No. individuals | 2,323 | | 2,323 | | | |
| No. probabilities | 16,540 | | 16,540 | | | |
| Log-likelihood | -74,126.826 | | -40,588.262 | | | |

[a] Standard errors could not be calculated for the rounding equation in the DHS due to the likelihood being flat in those dimensions. These numerical problems indicate that DHS probabilities are almost exclusively rounded to the minimum degree possible, which is to multiples of 10. We fix all parameters of the DHS rounding process to the values given in the table and calculate SEs for the remaining parameters.

Standard errors in parentheses; $^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

**Table F3:** Correlation matrices of individual and question sequence effects based on all valid probabilities

| | PB | DHS | Round PB | Round DHS |
|---|---|---|---|---|
| **a. Individual effects** | | | | |
| PB | 1 | | | |
| DHS | 0.781*** | 1 | | |
| Round PB | -0.123*** | -0.127*** | 1 | |
| Round DHS[a] | -0.162 | 0.0234 | 0.956 | 1 |
| **b. Sequence effects** | | | | |
| | PB | DHS | Round PB | Round DHS |
| PB | 1 | | | |
| DHS | 0.604*** | 1 | | |
| Round PB | -0.985*** | -0.731*** | 1 | |
| Round DHS[a] | -0.912 | -0.474 | 0.882 | 1 |

[a] Standard errors have not been calculated for correlations involving unobserved heterogeneity in rounding equation for DHS, because standard errors could not be computed for those elements of Cholesky matrix that pertain to rounding in the DHS.
***$p < 0.01$, **$p < 0.05$, *$p < 0.1$

# G Model fit

## *(for online publication)*

The results in the main text show that the reliability of associations between subjective longevity and covariates is similar regardless of whether we account for rounding in our model of expectations. However, level differences between the questionnaires are smaller once we account for rounding. Accounting for rounding also improves model fit. Figure E1 shows six histograms of reported probabilities in the data and of simulated probabilities from the models with and without rounding, pooling together all target ages. Even though the PB allows respondents to report any probability between zero and one hundred, panel a. shows that resulting answers are bunched at multiples of 10. In fact, the lower part of the distribution, up to and including 50 percent, is similar to that of the DHS shown in panel d. The model without rounding cannot mimic such bunching, see panels b. and e., but the model that accounts for rounding does fit the data relatively closely (panels c. and f.). Hence, censoring by itself does not produce the heaping at multiples of 10 that we observe in the data.

While the histograms in Figure E1 illustrate the importance of rounding, we may prefer to look at estimated densities in order to evaluate model fit. It is difficult to compare the fit of the models with rounding and without rounding, since the former is discrete while that latter is mostly continuous. As a consequence, the model without rounding necessarily smooths the data more. Figure E2 displays estimated densities for the data and for simulated probabilities from both models. We find that the density of the model without rounding fits the data much better than might be expected from the histograms: it provides a reasonable smoothed approximation of the bumpy density fitted on the data. This illustrates that even without rounding the model is successful in distributing probability mass over the interval between 0 and 100, even if it does not place the mass at the limited set of probabilities that we observe in the data.

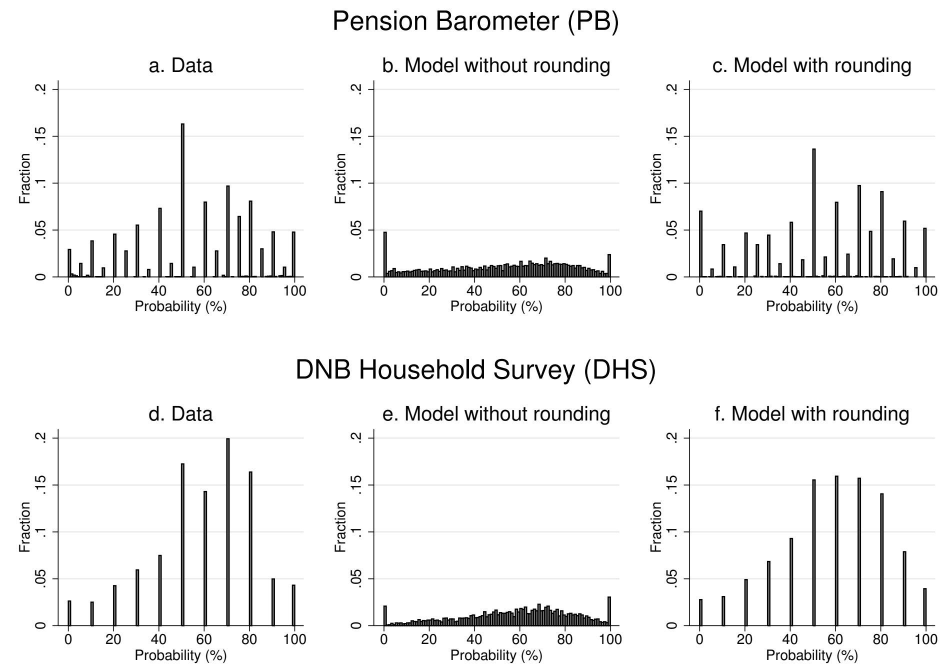



**Figure G1:** Histograms of data and simulated probabilities

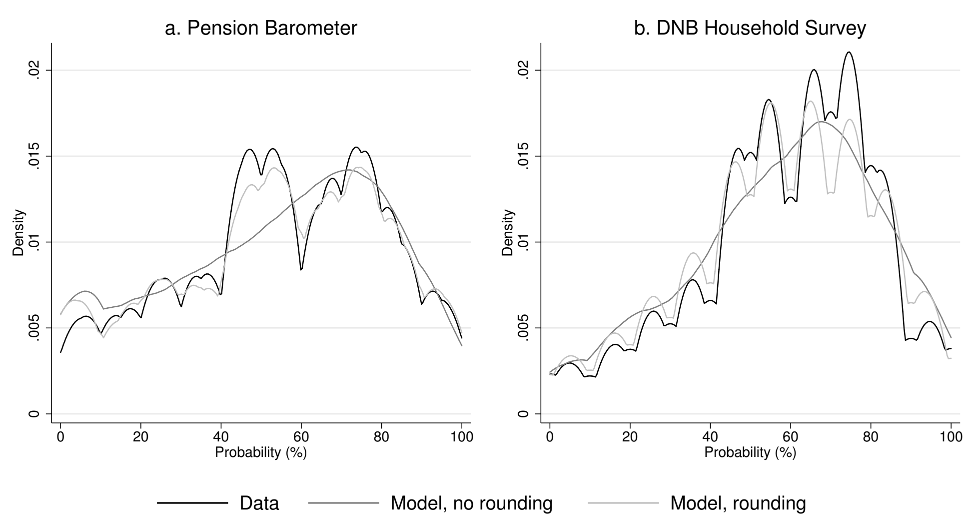



**Figure G2:** Kernel densities of data and simulated probabilities

# H Calibration of the life cycle model

## *(for online publication)*

While estimation of all parameters of the utility and bequest utility functions is beyond the scope of this paper, we do aim to simulate reasonable profiles for both wealth and labor supply. To this end we base the values of all utility parameters and leisure costs on the estimates reported in De Bresser (2019), who uses earlier waves of the DNB Household Survey to estimate a similar life cycle model. The parameters that drive the utility from leaving a bequest are then calibrated to match average yearly hours worked and wealth quartiles for the years 2006-2016, around the same time as our data on subjective probabilities. Moments are calculated by 2-year age bins up to age 70 and 5-year bins for ages 70-79. For labor supply we restrict the data to the cohorts born after 1949, because earlier cohorts had access to a generous early retirement scheme from which later cohorts were excluded. The wealth moments, on the other hand, do use respondents from all cohorts in order to extend the age range across which these moments can be computed. We remove cohort effects from both wealth and labor supply using fixed effects models as proposed in French (2005).

All simulations use the *same* preferences since the focus is on differences between sets of survival probabilities keeping all other aspects of the model constant. The fact that we only calibrate the model once means one should not compare model fit across sets of expectations. Different preferences could probably be calibrated for each set of mortality expectations such that all models fit the data roughly equally well (De Bresser, 2019). We calibrate bequest utility for survival probabilities derived from the Pensioenbarometer using the measurement model without rounding.

Figure H1 illustrates model fit. Panel a. shows that the model provides a reasonable fit of observed wealth quartiles, especially at the median. Panel b. focuses on labor supply. Average yearly hours worked are matched very closely for the ages 50-61. However, the model

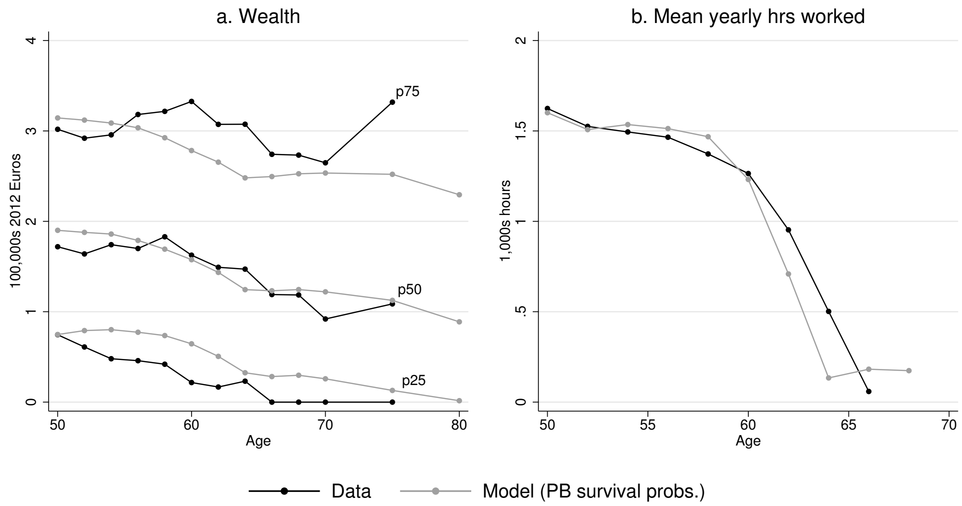



**Figure H1:** Calibration of the life cycle model: data and model simulations

produces a steeper drop in hours worked between the ages of 61 and 65 than is observed in the data. Overall the model fit is adequate given that we only calibrate three parameters out of ten.