WHO CAN PREDICT THEIR OWN DEMISE? HETEROGENEITY IN THE ACCURACY AND VALUE OF LONGEVITY EXPECTATIONS

TERESA BAGO D' UVA, ESEN ERDOGAN-CIFTCI, OWEN O'DONNELL,
and EDDY VAN DOORSLAER[‡]

May 2017

Abstract

Inaccurate longevity expectations will lead to suboptimal life cycle planning with negative consequences for wellbeing in old age. We evaluate the accuracy of expectations by comparing probabilities of living to 75 reported in the Health and Retirement Study with the actual survival of the respondents to that age. On average, survival predictions are poor and downwardly biased. Men predict less accurately, but that is because they face greater uncertainty due to a higher mortality rate. Women underestimate their survival chances more than men. Predictions are least accurate and most noisy at the lowest levels of education and cognitive functioning. A simple model suggests that welfare would be higher if everyone made decisions on the basis of the base survival rate rather than relying on the individual-specific survival probabilities reported. Despite the predictions of the least educated being the least accurate, they are not unambiguously the least valuable for decisions.

Keywords: Longevity, Prediction, Subjective Probability, Education, Cognition

[*] We thank Rob Alessie, Susann Rohwedder, Paul Lau, Martin Salm and Sílvia Sousa, as well as seminar and conference participants at Cornell, Lancaster, Manchester, Venice, Netspar (Paris), SSPH+ (Grindelwald), ASHE (Los Angeles), and the University of Southern California for helpful comments. Early stages of this research were supported by Netspar under theme *Health, income, work and care across the life cycle II*. The research has also been supported by the National Institute of Aging (R01AG037398-01); from the project *Thalis Panteion Investigating Crucial Interdisciplinary Linkages in Ageing Societies*, which is part of the *Education and Lifelong Learning* programme of the Greek Ministry of Education, Religion, Culture and Sport (MIC 380266); and a Veni grant (451-09-015) of the Dutch Organisation for Scientific Research (NWO).

‡ Bago d'Uva: Erasmus School of Economics, Erasmus University Rotterdam and Tinbergen Institute, bagoduva@ese.eur.nl; Erdogan-Ciftci: Novartis Turkey; O'Donnell: Erasmus School of Economics, Erasmus University Rotterdam, Tinbergen Institute, University of Macedonia and University of Lausanne, odonnell@ese.eur.nl ; van Doorslaer: Erasmus School of Economics, Erasmus University Rotterdam and Tinbergen Institute, vandoorslaer@ese.eur.nl .

1. INTRODUCTION

Accurate expectations of longevity are critical to optimal life cycle planning. Pessimism about survival chances might lead to inadequate saving for retirement and an impoverished old age. It may also discourage investment in preventive care that slows health deterioration in old age.

Forming accurate longevity expectations involves acquisition of health knowledge, perception of mortality risks and processing of information. These are cognitive skills that may be improved through education. Variation in ability to forecast the lifetime horizon over which finances are planned may contribute to observed differences by education and cognition in saving, retirement and wealth management (Fang et al. 2008, Banks and Oldfield 2007, Banks et al. 2010, Christelis et al. 2010, Smith et al. 2010, Behrman et al. 2012, Agarwal and Mazumder 2013). It may also explain differences in health behavior by education (Kenkel 1991, Cutler and Lleras-Muney 2010).

This paper evaluates the accuracy of longevity expectations by examining the extent to which probabilities of living to 75 reported in the Health and Retirement Study (HRS) correspond to the rate of survival to that age of the respondents. We measure overall prediction accuracy and correct for the fact that prediction is more difficult when survival is more uncertain. We examine prediction calibration, discrimination and noise – properties of the joint distribution of predicted and actual survival that each influence accuracy. We compare accuracy and its components by sex, education and cognitive functioning.

On average, the predictions are poor. For men, they are worse than everyone reporting a 50-50 chance of surviving to 75. For women, they are no better than this low benchmark. The poorer prediction performance of men is entirely due to their survival rate being closer to 50 percent, which means the outcome is more uncertain and so more difficult to predict. Both sexes underestimate their survival chances, but women do so to a far greater degree.

The survival predictions of the least educated and cognitively able are the least accurate. This is the net outcome of opposing differences. On the one hand, the predictions of the low education and cognition groups discriminate better between those who live to 75 and those who die. On the other, the predictions of these groups are less well calibrated – their level corresponds less to the actual survival rate – and they display much more noise.[1] These latter effects dominate. The education gradient in accuracy is much stronger than the cognition gradient for women. For men, the two gradients are similar.

[1] There is a difference in calibration by education, but not cognition.

Limited ability to think probabilistically about longevity, and possibly other prospects, potentially explains the strong correlation between cognitive functioning and the quality of decision-making (Choi et al. 2014). It may also explain the difficulty that individuals with low education and numeracy skills have in valuing an annuity (Brown et al. 2017). Heimer et al (2015) show that pessimism about survival chances at younger ages and optimism at older ages respectively explain under-saving before retirement and sub-optimal depletion of assets after retirement. They find that the deviation from the optimal consumption path induced by biased mortality beliefs is greater at lower levels of education. This suggests that the less educated lose more by relying on inaccurate longevity expectations.

While we do not estimate the impact of biased expectations on behavior, we do assess the losses that would be incurred as a consequence of basing decisions on inaccurate predictions of longevity. Using a simple model that captures the structure of at least some decision problems that would optimally make use of longevity expectations, we measure and compare the value of the survival probabilities reported in the HRS relative to a scenario in which everyone decides on the basis of the base survival rate. Prediction value depends on the cost of insuring longevity risk relative to the loss incurred if it is not insured.

Over most of the range of the cost/loss ratio, welfare would be higher if everyone were to decide on the basis of the base rate rather than each using their reported survival probability. The losses arising from use of the subjective probabilities can be substantial. Despite the predictions of the least educated being the least accurate, they are not of least value in decision-making over the full range of the cost/loss ratio. However, at the extremes where overly optimistic or pessimistic predictions result in large losses of welfare, it is the least educated who lose most by relying on these inaccurate predictions.

Previous analyses of the HRS and other representative surveys have established that reported survival probabilities predict mortality in subsequent waves (Van Doorn and Kasl 1998, Smith et al. 2001, Hurd and McGarry 2002, Siegel et al. 2003, Hurd 2009). Elder (2013) argues that this correlation does not necessarily demonstrate that the reported probabilities provide useful information on longevity. He shows that life table survival probabilities do a better job of predicting mortality within the HRS, and that the perceived deviation of survival chance from the life table average accounts for little more than a third of the variation across individuals in the reported probabilities. This suggests that the reported probabilities contain a great deal of noise. We demonstrate that the signal-to-noise ratio is lowest for the least educated and cognitively able.

Delavande and Rohwedder (2011) are considerably more optimistic regarding the information content of reported survival probabilities. They show that the wealth gradient in these probabilities is remarkably similar to the gradient in actual survival in the HRS. However, the gradient in the reported probabilities by education is markedly shallower than that in actual survival (see Delavande and Rohwedder 2011, online Appendix). This is consistent with our more direct demonstration that individuals with lower educational attainment (and cognitive ability) are less accurate in predicting their longevity.

Our contribution to the literature on subjective predictions of longevity is fourfold. First, by using a longer follow-up period, we are able to compare predicted survival to 75 with actual survival to that age for a large number of HRS respondents. This allows us to measure prediction accuracy directly by comparing the outcome (survival to 75) with its prediction at the individual level. Second, we examine different dimensions of prediction accuracy – skill, calibration and discrimination – as well as the noise that contributes to inaccuracy. This helps us understand why accuracy varies and also to speculate about its consequences for decision-making. Third, we document heterogeneity in the accuracy of longevity expectations by education and cognition. In doing so, we move beyond the question of whether a representative agent possesses and can process information on longevity prospects that are instrumental to solving lifecycle planning problems. We address a question that has a distributional motivation: which individuals are more likely to hold inaccurate expectations and, consequently, make suboptimal decisions? Fourth, we assess not only the accuracy of longevity predictions but also their value in decision-making. As far as we are aware, the approach we take to assess prediction value has not previously been used to evaluate subjective probabilities reported in a survey.

In the next section, we explain how we select samples from the HRS and describe the measurement and distributions of the two key variables – reported probability of survival to 75 and actual survival to that age. In the third section, we present the measures used to assess prediction accuracy and value. Results are presented in the fourth section. We first examine the accuracy of predicted survival by sex and education. We then assess prediction value by the same two characteristics. Finally, we examine prediction accuracy by cognition.

2. Data and descriptives

2.1 Sample

We use data from the original cohort of the US Health and Retirement Study (HRS), which is a nationally representative sample of individuals born between 1931 and 1941. These individuals and their spouses were interviewed for the first time in 1992 and have been called for interview every two years since. We mainly use data on chances of survival to 75 reported in the first wave, when the cohort members were aged 50-61. Spouses of all ages were interviewed but we restrict attention to those younger than 66 in wave 1 because only respondents below this age were asked to predict their chances of survival to 75 in later waves used for part of the analysis. We further restrict the sample to those aged 54 and above in 1992 since younger respondents could not have reached the age of 75 in 2014 when the last wave of the HRS currently available was conducted.[2] Finally, we drop all respondents interviewed by proxy since the question about survival chances was not asked in such cases.

Applying these selection criteria gives a sample of 7,124 individuals aged 54-65 in 1992, of whom 6,367 are from the original HRS 1931-41 birth year cohort. Those aged 62-65 are older spouses of this cohort. Because men are much less likely than women to have an older spouse, there are relatively few females older than 61 in the sample.[3]

2.2 Measurement of key variables

We take all variables from the RAND HRS data files (Bugliari et al. 2016). Education groups are formed from years of schooling and highest qualification obtained: i) high school dropout (0-11 years of schooling and no qualification other than General Educational Development (GED); ii) high school graduate (12 years of schooling and high school diploma); iii) some college (more than 12 years of schooling and high school diploma or GED); and, iv) college graduate (bachelor degree or higher). The bottom two education groups are each roughly twice as large as each of the top two (Appendix Table A2). Women are underrepresented among college graduates and overrepresented among high school graduates. Mean age is very similar across the education groups (about 58).

---

[2] We set the lower limit at 54 rather than 53 to be sure that all those selected would have been 75 in 2014. This avoids the classification of vital status in 2014 being dependent on the timing of the wave 12 interview and birth date.

[3] The HRS oversampled Hispanics, Blacks and residents of Florida. Weights to adjust for this are available only for the direct members of the 1931-41 birth cohort. Using them would require dropping spouses aged above 61. Given item non-response and attrition (see next sub-section), application of these weights would, however, not make the sample representative even of the original HRS cohort.

Wave 1 respondents were asked: *Using any number from 0 to 10, where 0 equals absolutely no chance and 10 equals absolutely certain, what do you think are the chances that you will live to be 75 or more?*[4] We rescale the response to the (0,1) range and refer it as *subjective survival probability*. Among those asked the question, the item non-response rate is 2.5 percent (Appendix Table A1). High school dropouts are about three times more likely than any other education group to answer 'don't know' or not at all (idem.). This may indicate that the least educated hold less well-defined longevity expectations and/or have greater difficulty in completing the abstract task of formulating and reporting a probability.[5]

To determine whether each respondent actually survived to 75, we must establish vital status in 2014 of all HRS cohort members who were aged 54-65 and were interviewed in 1992. The study first determines the vital status of all individuals who were ever HRS respondents through tracking and exit interviews. In each wave, non-respondents are considered alive if they were contacted directly by an interviewer during the wave, or said be alive by a spouse or partner, or not reported dead. If no informative contact was made because the respondent dropped out of the survey when alive, for example, then the respondent's vital status is classified as unknown. For those reported dead, exit interviews include a question about the date of death. Additionally, the HRS ascertains vital status and obtains dates of death by matching records of all non-respondents in each wave to the National Death Index (NDI).

We use all of these sources to establish whether each wave 1 respondent in the selected age range survived to the age of 75. If we have information on date of death, we use this together with date of birth to determine whether the individual died before reaching 75. If we do not have information on the date of death, we consider individual to be alive by 75 if one of two conditions hold: i) the individual was a respondent, contacted by the HRS or confirmed to be alive by a partner in any wave after the person turned; ii) the individual is presumed to be alive in wave 12 on the basis of exit interviews and no match was found for the person in the NDI.

Additionally, we categorize as alive most individuals whose vital status in 2014 is classified by HRS as unknown (from exit interviews) but for whom there is no match with NDI records. This potentially introduces error. For example, someone who has emigrated, has been lost to HRS follow-up and who died before reaching 75 will not appear in the NDI and will be incorrectly classified as alive. But such cases are likely to be rare. The payoff to accepting these errors is that

[4] The respondent was allowed to report an integer number only.

[5] In addition to higher non-response, the less educated have been found to be more prone to violate monotonicity and binary complementarity in answering subjective probability questions (Dominitz and Manski 1997, Dominitz and Manski 2006, Delavande and Rohwedder 2008, Van Santen et al. 2012).

we correctly classify as alive a substantial number of cases who the HRS has not been able to trace but who are not recorded as dead in the NDI. We do this only for those aged 56 or older in wave 1 because NDI information is only available up until December 2011. Individuals aged 54-55 could have turned 75 between 2012 and 2014, but may also have died before reaching the threshold. These individuals are coded as missing on the indicator of survival to 75.

Following these procedures, we are unable to establish survival to 75 for 2.8 percent of respondents aged 54-65 in wave 1 who were asked the survival probability question (Appendix Table A1). The proportion missing on survival to 75 does not differs significantly by either sex or education.

In total, 5.1 percent of the 7,124 wave 1 respondents in the appropriate age range are lost from the sample due to missing information on their predicted probability of survival to 75 and/or their actual survival to that age. We lose a larger fraction of high school dropouts than any other education group. We are left with a sample of 6,758 individuals aged 54-65 in 1992 who report their probability of surviving to 75 and for whom we can established whether they reached that age.

2.3 SUBJECTIVE AND ACTUAL PROBABILITY OF SURVIVAL BY SEX AND EDUCATION

Given differences in mortality rates and education by gender, we conduct most analyses separately for males and females. The distribution of subjective survival probabilities has a similar shape for men and women (Figure 1). Spikes at focal responses of 0, 0.5 and 1 are common in subjective probability data. Around a fifth of women and slightly more men report a probability of 0.5 (Table 2). Such bunching has been attributed to an extreme form of rounding (Gan et al. 2005, Manski and Molinari 2010, Kleinjans and Van Soest 2014), epistemic uncertainty (0.5='I haven't a clue') (Fischhoff and Bruine De Bruin 1999, Bruine de Bruin and Carman 2012) and extreme ambiguity (0.5='I am very unsure') (Hill et al. 2004, Hudomiet and Willis 2013). More than a fifth of both sexes report a value of 1, which they are instructed to use only if they are absolutely certain of living to 75. About 7% report 0. Taken literally, this means that they are absolutely certain of not reaching 75.

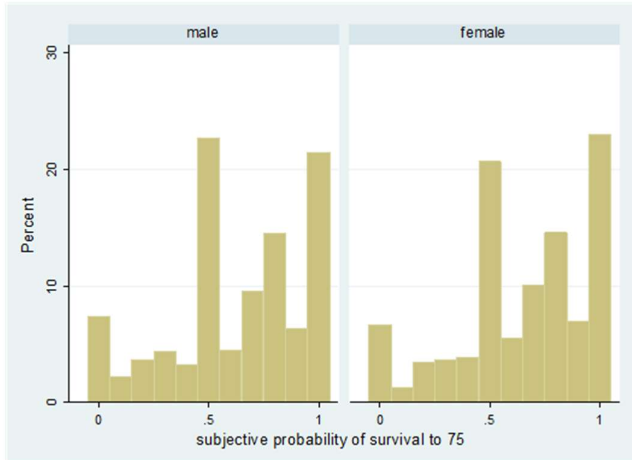


Figure 1: Distribution of subjective probabilities of survival to 75 by sex

On average, men report that their probability of living to 75 is 63 percent, which is a significant 4.8 percentage points less than the percentage of the sample that actually reached this age (Table 2). The median male response of 70 percent is about 2 percent above the fraction that survives to 75. So, while roughly half of men marginally overestimate their chances of living to 75, the mean is an underestimate because of the effect of the zero responses. Women report a higher probability of surviving to 75 but not to an extent that is anywhere near consistent with their greater longevity. The mean subjective probability of women is a substantial and significant 12.7 percentage point underestimate of their actual chance of survival. Not even the median response of women is above the objective probability. On average, women are clearly pessimistic about their chances of survival to 75.

Table 2: Subjective and actual probability of survival to 75 by sex

	Male	Female	All
Subjective probability of survival to 75			
Mean	0.6305	0.6503	0.6405
Std. Dev.	0.3023	0.2960	0.2992
Median	0.7	0.7	0.7
% reporting 0	7.4	6.8	7.1
% reporting 0.5	22.9	20.7	21.8
% reporting 1	21.4	22.9	22.2
Actual probability of survival to 75	0.6786	0.7768	0.7283
N	3,339	3,419	6,758

Notes: Top panel shows distribution of subjective probabilities of survival to 75 reported by wave 1 respondents. Second bottom row shows the proportion of these respondents who did survive to 75. For males, females and the full sample, the mean subjective probability is significantly different from the proportion surviving to 75 at the 5% level or less. The mean subjective probability and the proportion surviving to 75 each differ significantly between males and females.

Figure 2 provides greater insight into the degree to which the mean subjective survival probability corresponds to the actual probability of survival. For males up to the age of 60, the two statistics are reasonably close indicating little bias. Above that age, older men do not adjust their predictions upward sufficiently to capture the objectively rising probability of reaching a closer endpoint. This 'excess flatness' (Elder 2013) is responsible for men underestimating their survival chances, on average. In contrast, women underestimate at all ages to a similar degree.

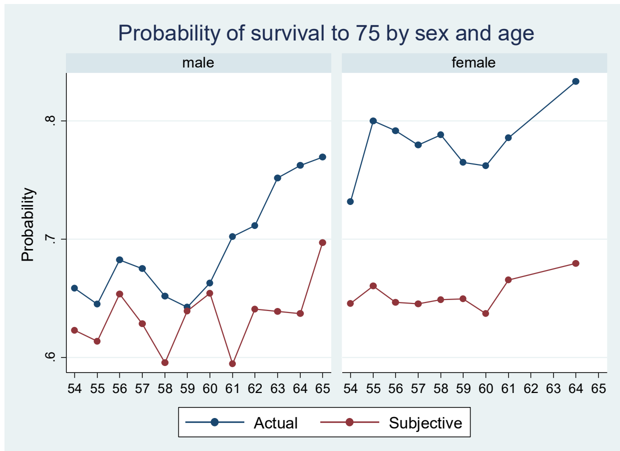


Figure 2: Mean subjective and actual probability of survival to 75 by sex and age

Notes: Because there are few women above the age of 61, they are all put into one group, the mean for which is shown at the point indicated by age 64 in the graph.

Table 3 shows summary statistics of the distribution of subjective survival probabilities by education level for each sex, along with the proportion of each group that actually lived to reach the age of 75. Both male and female high school dropouts are much more likely than the higher education groups to report a zero chance of surviving to 75. The fraction making this pessimistic forecast falls monotonically with rising education for men, and is close to doing so for women also. Male college graduates are least likely to report a probability of 1, while the most highly educated women are least likely to give a focal response of 0.5. Kleinjans and Van Soest (2014) find, using the HRS, that modelling focal responses, as well as non-response, makes little difference to the estimated education gradient in longevity expectations. We examine the contribution of focal responses to generating differences in the accuracy of longevity predictions by education.[6]

Consistent with the established education gradient in mortality risks, the mean subjective survival probability rises monotonically with education for both sexes, while the standard deviation falls. Interpretation of the latter relationship depends upon whether the greater variance in the probabilities reported by the less educated is systematic or not. If it is, then it would imply that there is more information about mortality risks in the predictions made by the lower education

[6] This analysis will be added in the next version of the paper.

groups. If it is not, then it would correspond to greater noise, possibly reflecting difficulties in forming or reporting beliefs about survival chances.

The average forecast is less than the actual probability of survival to 75 for men at all levels of education but it is college graduates who underestimate their survival chances by most – 10 percentage points. Highly educated men appear not to fully appreciate the extent to which they are privileged with respect to longevity. Little more than three fifths of the remarkable difference of 17.5 percentage points in the probability of living to 75 between the most and least educated men is reflected in the predictions reported by these groups. This suggests that while Delevande and Rohwedder (2011) are correct in arguing that inequality in subjectively predicted longevity can be used to proxy inequality in actual longevity, the estimate may well be a substantial underestimate. The median responses of men belonging to the lowest and highest education groups underestimate actual survival chances. Using this statistic, the bottom education group underestimates to a slightly greater extent. The median response of the other two groups is an overestimate.

Women at all levels of education underestimate their survival chances by at least 10 percentage points, on average. Female high school graduates underestimates by almost 15 points. For women, subjective survival probabilities provide a much better approximation to the education gradient in longevity. Almost all of the 16 percentage point difference in the probability of surviving to 75 between the top and bottom education groups is captured by the respective difference in the reported survival chances. The median forecast reported by women is also an underestimate at all levels of education, and the magnitude of this bias is smallest for the top two groups.

Table 3: Subjective and actual probability of survival to 75 by sex and education

	Males				Females			
	High school dropout	High school graduate	Some college	College graduate	High school dropout	High school graduate	Some college	College graduate
Subjective probability of survival to 75								
Mean	0.5776	0.6349	0.6571	0.6853	0.5861	0.6484	0.7089	0.7441
Std. dev.	0.3337	0.3018	0.2894	0.2420	0.3338	0.2877	0.2490	0.2246
Median	0.5	0.7	0.7	0.7	0.6	0.7	0.8	0.8
% reporting 0	11.48	7.41	6.06	1.90	13.10	5.33	1.52	1.86
% reporting 0.5	23.59	23.60	21.93	21.46	20.40	22.67	20.54	16.24
% reporting 1	22.59	22.86	23.73	15.33	22.60	23.15	23.06	22.74
Actual probability of survival to 75	0.6102	0.6761	0.6858	0.7854	0.6948	0.7955	0.8367	0.8561
N	1098	945	611	685	1137	1257	594	431

Note: As for Table 2.

2.4. Subjective and actual probability of survival by cognitive functioning

To examine the accuracy of longevity expectations by cognition, we use a sample selected from wave three (1996) when the HRS started to collect measures of intact mental status. We restrict attention to members of the original HRS cohort and their spouses who were aged from 58 to 65 in 1996.[7] This gives us 5,261 respondents who were asked to report their chances of survival to 75 and could potentially have lived to 75 by 2014.

The HRS contains measures of cognitive functioning in several domains based on validated tests (Ofstedal et al. 2005, Fisher et al. 2012). We focus on measures of *episodic memory* and *intact mental status* that are described in Appendix C. Following Ofstedal et al. (2005), we aggregate the measures in the two domains into a total score increasing in cognitive functioning. We categorize respondents into sex-specific quartile groups of the age-sex standardized score.[8]

In wave 3, respondents were asked to report their percent chance of surviving to 75 on a scale of 0 to 100, with the end points again specified to indicate absolutely no chance of survival and certainty of survival. We again rescale to the (0,1) range. The item non-response rate in wave 3 is more than double what it is wave 1: 5.5 percent of the full sample and 11.3 among high school dropouts (Appendix Table A1). There is no obvious explanation for this. Following the procedures described in section 2.2, we are unable to establish whether 1.8 percent of respondents aged 58-65 in wave 3 (who were asked the survival probability question) survived to 75 (Appendix Table A1). In total, we lose 7.4 percent of the selected wave 3 sample, and 12.9 percent of high school dropouts, because of missing information on their predicted or actual survival to 75. We are left with a sample of 4,870 individuals aged 58-65 in 1996.

Table 4 summarizes the distribution of subjective survival probabilities by sex and quartile of cognitive functioning. For both sexes, respondents in the bottom quartile are the most likely to report that they have no chance of surviving to 75. Men in this quartile are also the most likely to give the focal responses of 0.5 and 1. As with education in wave 1, the mean of the predictions rises and their standard deviation falls in moving from the bottom to the top quartile of cognitive functioning for both genders.[9] With the exception of men in the bottom quartile, all groups underestimate their chances of surviving to 75, on average. There is no consistent trend in the bias by levels of cognition.

[7] The upper age bound is actually defined by birth year after 1930 since this was the criterion for being asked about chances of survival to 75 in wave 3.

[8] The standardised score is the residual from a linear regression of the score on a full set of sex-specific age-year dummies.

[9] The same gradients by education that are observed in the wave 1 data are also present across the wave 3 observations (see Appendix Table A4).

Table 4: Subjective and actual probability of survival to 75 by sex and cognition quartile

	Males				Females			
	Bottom quartile	2nd quartile	3rd quartile	Top quartile	Bottom quartile	2nd quartile	3rd quartile	Top quartile
Subjective probability of survival to 75								
Mean	0.5955	0.6297	0.6486	0.7038	0.5745	0.6568	0.6658	0.7013
Std. dev.	0.3335	0.3064	0.2919	0.2595	0.3689	0.3108	0.2908	.2838
Median	0.50	0.63	0.70	0.75	0.50	0.75	0.75	0.75
% reporting 0	10.34	8.11	4.45	2.82	15.96	7.01	4.25	4.79
% reporting 0.5	31.03	26.98	25.79	23.16	20.64	24.23	28.55	23.35
% reporting 1	26.50	22.26	20.41	23.16	26.65	25.84	24.74	24.55
Actual probability of survival to 75	0.6098	0.7302	0.7718	0.7665	0.7262	0.7956	0.8536	0.8653
N	551	530	539	531	683	685	683	668

Note: Wave 3 respondents.

3. Measures of prediction accuracy and value

We evaluate the subjective survival probabilities by addressing two related but distinct questions. How accurate are the predictions? How valuable are they?[10] To answer the first question we need to establish the extent to which the probabilistic predictions of survival to 75 correspond to and vary with actual survival to that age. This involves examination of the joint distribution of these two variables. Addressing the second question requires assessment of the utility of the predictions in improving decisions about longevity-contingent behaviours, such as retirement, wealth annuitization and long term care insurance. This cannot be done on the basis of the joint distribution alone. The relative cost of different types of prediction errors must be specified (Lahiri and Yang 2013). We first describe the measures of prediction accuracy before outlining an approach to evaluating prediction value.

3.1 Accuracy of predictions

We use the sample mean square error (MSE) to measure of overall accuracy. In the context of forecasting a binary outcome, this is known as the Brier score (Brier 1951),

$$BS = \frac{1}{n}\sum\left(p_i - y_i\right)^2 \ \in\left[0,1\right] \qquad (1)$$

where p_i is the subjective probability of survival to 75 reported by individual *i,* $y_i = 1\left(age\,at\,death \geq 75\right)$ and *n* is the sample size, or the sub-sample size when we calculate for a sex, education or cognition group. For a perfect forecast that is 100% accurate, *BS* = 0. A useful benchmark is *BS* = 0.25, which is the score that would be achieved if everyone perceived their chances of surviving to 75 as equivalent to a coin toss and reported a probability of 0.5.

[10] See Lahiri and Yang (2013) for an excellent review of measures of the accuracy and value of predictions of binary outcomes.

The Brier score depends on mean survival $(\bar{y})$. The closer the survival rate is to 0.5, the greater is the variance and the more difficult is the prediction task. Given differences in mortality by sex, education and cognition, this affects comparisons of prediction performance across categories of each of these characteristics. For some purposes, this is irrelevant. If the aim is to target groups that have the greatest difficulty in predicting longevity, then one wants to identify those making the greatest errors irrespective of whether their poor performance is attributable to the complexity of the task faced or their limited skill as forecasters. But assessment of prediction skill can also be interesting. Not least because it is an obvious potential explanation for variation in accuracy.

We measure prediction proficiency by the Brier skill score (BSS), which is one minus the ratio of the Brier score to the variance of survival $(\bar{y}(1-\bar{y}))$, and can also be thought of as indicating the accuracy of the subjective survival probabilities relative to the accuracy that would be achieved if everyone were to predict the (sub-sample) base survival rate,

$$BSS = 1 - \frac{BS}{\bar{y}(1-\bar{y})} = 1 - \frac{BS}{\frac{1}{n}\sum(\bar{y}-y_i)^2} = 1 - \frac{BS}{BS_b} \in [-\infty, 1] \ . \tag{2}$$

The upper limit of this skill score is reached when everyone predicts their survival perfectly. If everyone reports the base rate, then the score is 0. If the MSE of the subjective probabilities is greater than that generated by everyone reporting the base rate $(BS > BS_b)$, then BSS<0. This is usually interpreted as indicative of unskilful forecasts. This is too strict in this application. Retrospectively, we can calculate the survival rate of the cohort. But in 1992, the cohort members could not have known what proportion of them would survive to 75. The best estimate would have relied on mortality rates of older cohorts. Comparison of the subjective probabilities with a 'prediction' model exploiting information that only became available ex post sets the bar a little too high. The skill score can nonetheless be used to compare prediction performance across groups facing different degrees of uncertainty about survival to 75.

Standard errors for the Brier score and skill score are estimated using the formulae given in Bradely et al (2008).[11]

To determine the direction of any systematic deviation of predicted from actual survival, we examine sample bias in the subjective survival probabilities, $\left(\frac{1}{n}\sum p_i - y_i\right)$. This tell us whether

[11] Since the skill score is a ratio of sample estimators, it is a biased estimator of the population parameter. Since the bias needs to be approximated, the sampling variance of BSS can only be approximated. Bradely et al (2008) derive a first order approximation.

survival chances are overestimated or underestimated, on average. But since it allows errors in each direction to cancel out, it does not provide a measure of their average magnitude.

We assess and compare subjective survival probabilities in the two main dimensions of prediction accuracy: *calibration* and *discrimination*. The first is concerned with the correspondence between predicted and actual survival: do X% of individuals who report a chance of X in 100 of surviving to 75 actually survive to this age? The second property concerns the covariance between predicted and actual survival: do the predictions discriminate those who survive from those who die before reaching 75?

We use Hosmer-Lemeshow (1980) to test whether the subjective survival probabilities are calibrated.[12] We measure calibration using Murphy's (1973) reliability index, which is the weighted average of the squared deviation of the predicted from actual survival rate within deciles of the subjective survival probability distribution: $R = \frac{1}{n}\sum_{g=1}^{G} n_g \left(\overline{p}_g - \overline{y}_g\right)^2 \in \left[0,1\right]$, where $\overline{p}_g$ is the mean prediction in group *g*, $\overline{y}_g$ is the proportion of the group that survives to 75 and n_g is the size of the group. A lower value indicates better calibration, with the lower limit of 0 being achieved with perfectly calibrated subjective probabilities that correspond, on average, to the actual survival rate within each group.

To assess discrimination, we examine Receiver Operating Characteristic (ROC) curves, which plot the true positive rate against false positive rate as the probability threshold for predicting survival to 75 is varied. We measure discrimination by the area under the ROC curve – the concordance c statistic. This can be interpreted as the probability that a randomly selected respondent who survives to 75 reports a higher subjective survival probability than a random respondent who dies. If the subjective probabilities do not discriminate, then the true positive rate equals the false positive rate at all threshold probabilities and the c statistic is 0.5. We also measure discrimination by the difference in the means of the subjective survival probabilities of those who survive and those who die (the discrimination slope): $\frac{1}{n_1}\sum_i 1\left(y_i = 1\right) p_i - \frac{1}{n_0}\sum_i 1\left(y_i = 0\right) p_i$, where n_1 and n_0 indicate the number of respondents who survive and die respectively.

12 This test is often criticized for having low power. As seen from the results in Table 7, this is not a problem in this application. We deviate from the standard implementation of the test that involves splitting observations by quantiles of the predictions. Given the subjective survival probabilities are discrete, this would result in an uneven number of groups across sub-samples. This could be problematic since the test is known to be dependent on the number of groups. We therefore split observations into a fixed number (11 in wave 1 and 12 in wave 3) of intervals of subjective probability values within each sub-sample.

Decomposition of the Brier score reveals different dimensions of prediction accuracy and the contribution of each to overall performance. We use the decomposition proposed by Yates (1982) to show the extent to which variation in accuracy by education and cognition is driven by noise. For the population, the mean square error can be written as,

$$E\left(P-Y\right)^2 = Var\left(Y\right)+\left[E\left(P\right)-E\left(Y\right)\right]^2-2Cov\left(P,Y\right)+Var_{P.\min}\left(P\right)+\Delta Var\left(P\right) \ . \qquad (3)$$

The first term on the right hand side – the variance of survival – captures uncertainty and so the difficulty of the prediction task. The second term is the square of the bias and is an imperfect measure of calibration. The third term is twice the covariance of predicted and actual survival, which is related to discrimination and reduces the error. The fourth term is the part of the variance of the subjective survival probabilities that is related actual survival. It is equal to the product of the covariance and the discrimination slope, and so is the minimum variance of the predictions that could be reached given their observed covariance with survival.[13] The final term is the excess of the actual variance of the subjective survival probabilities over this lower bound variance. This captures noise – variance in the predictions that carries no information on survival.

3.2 VALUE OF PREDICTIONS

Since the Brier score is simply the average of the squared prediction errors, it treats overestimates and underestimates of survival chances equally. This may be inconsistent with the cost a decision maker attaches to false positive prediction relative to that arising from a false negative. Consequently, it is possible that the predictions of one group may be assessed as more accurate in the sense of having a lower Brier score and yet decisions based on those predictions may generate less surplus than would those taken on the basis of the less accurate predictions.[14]

We follow an approach to measuring the value of predictions that is suited to any situation in which the value of a binary decision depends on an uncertain binary outcome and the decision is taken on the basis of a probabilistic prediction of the outcome and a loss function (Thompson 1952). This structure captures the essence of many decision problems (Wilks 2001), and some that are related to longevity. For example, whether it pays off in strict financial terms to annuitize a pension depends upon whether the individual lives beyond a certain age. Of course, that age is unlikely to be 75 for all individuals. We claim only that the model provides a stylized representation of some binary decisions that would optimally be taken on the basis of survival chances and may therefore provide an approximation to the value of the survival predictions observed in our data.

---

13 $Var_{P.\min}\left(P\right)\equiv\left[E\left(P|Y=1\right)-E\left(P|Y=0\right)\right]Cov\left(P,Y\right)$

14 See Lahiri and Yang (2013) for a simple example.

Let $L(a,Y)$ represent the loss associated with the binary outcome Y when the agent has selected a from two possible actions (Lahiri and Yang 2013). Assume, following Thompson and Brier (1955) and Mylne (1999), that the loss function takes form: $L(1,1)=L(1,0)=C>0$, $L(0,1)=L>C$ and $L(0,0)=0$. Lahiri and Yang (2013) argue that this simple model captures the decision faced by someone contemplating the purchase of full insure ($a=1$) at a cost C against a loss L that is incurred if $Y=1$. If the event does not occur ($Y=0$) and no insurance has been purchased ($a=0$), then neither a loss is incurred or a cost borne.

With this setup, Wilks (2001) derives a measure of the value of a prediction model as a function only of the cost/loss ratio, C/L. A rational agent who perceives the probability of incurring the loss as P will insure if it costs no more than the expected loss without insurance, $PL \geq C \rightarrow P \geq C/L$. Note that C and L need not be defined in monetary terms and so the model does not impose risk neutrality.[15] If $C \geq L$, then insurance will never be taken. And if $C \leq 0$, it will always be taken. So the relevant range of the cost/loss ratio is $0<C/L<1$.

Consider the following example that demonstrates the relevance this simple model can have to some decisions that utilize survival predictions. A widow wishes to leave an inheritance of a given composition and value (e.g., the family home plus $). If she dies before a certain age ($Y=0$), then she will be able to realize this plan. But if she lives for longer ($Y=1$), which she puts the chance of happening at P, then her consumption needs in old age will eat into the inheritance. The family home will have to be sold (L). The woman can insure the risk to the inheritance by purchasing an annuity at a cost (C) that will reduce the amount of cash that can be left to her daughter but will ensure that the family home is passed on.[16] The insurance instrument need not be an annuity. The model would also apply to the decision of whether to purchase long-term care insurance to protect an inheritance.

If the agent has information only on the base rate π, she will insure if $\pi L \geq C$ and so incur a cost of C. Otherwise, she will face an expected loss of πL. If she optimally utilizes the information in the base rate, then her expected loss relative to the situation in which there is no loss to be faced or insured is $EL_b = \min(\pi L, C)$.

[15] To see the scope for risk aversion, let l and c be the loss and cost respectively in monetary units, and W be wealth. In the absence of insurance, expected utility is $EU=PU(W-l)+(1-P)U(W)$. If insurance is purchased, then utility is $U(W-c)$. An expected utility maximizer purchases insurance if $U(W-c) \geq PU(W-l)+(1-P)U(W) \rightarrow P[U(W)-U(W-l)] \geq U(W)-U(W-c) \rightarrow PL \geq C$, with L and C defined as the loss and cost respectively in utility terms.

[16] Normally, a bequest motive would reduce demand for an annuity. Here it raises demand because of the desire to leave an inheritance in a particular form – the family home.

If the agent could perfectly predict survival, then she would insure and incur C if she knows that she will survive and will not otherwise. The expected loss over a number of perfectly informed agents is $EL_{perf} = \pi C < \min(\pi L, C)$, with the inequality holding because $\pi \in [0,1]$ and $C<L$. The gain from using a perfect prediction rather than relying on the base rate is $EL_b - EL_{perf}$.

An agent optimally making use of an imperfect prediction P will insure only if this prediction exceeds the cost/loss ratio. So the predicted probability is transformed into a binary decision depending on its relation to the cost/loss ratio. Given a particular value of the cost/loss ratio, the proportion of agents who correctly predict survival, and so incur the cost of insurance but avoid the loss, is given by integrating the joint probability of predicted and actual survival evaluated at the positive outcome of the latter, *f(P,Y=1)* over all predictions in excess of the cost/loss ratio, $q_{11} = \int 1\left(P \geq C/L\right) f(P, Y=1) dP$ (Wilks 2001). The proportion of agents holding predictions that lead them not insure against longevity risk when in fact they do survive beyond the threshold age is $q_{01} = \int 1\left(P < C/L\right) f(P, Y=1) dP$. And the proportion of agents who decide to insure against longevity but who do not survive into old age is $q_{10} = \int 1\left(P \geq C/L\right) f(P, Y=0) dP$.

The expected loss across agents optimally making use of their imperfect predictions is given by the sum of the losses in each contingency weighted by their respective probabilities,

$$\begin{aligned} EL_P &= (q_{11} + q_{10})C + q_{01}L \\ &= C\sum_{j=0}^{1}\int 1\left(P \geq C/L\right) f(P, Y=j) dP + L\int 1\left(P < C/L\right) f(P, Y=1) dP \end{aligned}$$

Wilks (2001) suggested to measure the value of a forecast model, or in our case the subjective survival probabilities of a population, as the gain from using that model compared with the naïve base rate forecast relative to the gain from a perfect set of predictions,

$$VS = \frac{EL_b - EL_P}{EL_b - EL_{perf}} \tag{4}$$

This value score lies in the range of negative infinity to 1. If the predictions are informative in the sense of being better than the base rate forecast, then decisions based on these predictions will result in a lower loss than is obtainable using the base rate only and VS will be positive.

VS depends not only on the joint distribution of the predictions and outcomes, i.e. accuracy, but also on the loss function, which is individual specific. However, Wilks (2001) shows that VS is a function of C/L alone, and not C and L separately. This makes it possible to plot VS as a function of C/L and to compare sets of predictions over a range of interesting values of C/L. VS=0 when $C/L=0$ and when $C/L=1$.[17] Between these two extremes, VS can take positive and negative values.

4. RESULTS

4.1 PREDICTION ACCURACY BY SEX AND EDUCATION

Overall prediction accuracy

Table 5 presents the Brier score, the Brier skill score and the bias of survival predictions for the full wave 1 sample as well as by sex and education. The Brier score over all observations is slightly and significantly above 0.25, indicating that the reported survival probabilities perform worse than the accuracy that would be achieved if every respondent had declared a 50:50 chance of surviving to 75. The null hypothesis that the prediction is equal to the true probability of survival for each individual is decisively rejected using observations from the full sample and from all sub-samples.[18] The skill score is significantly negative, which implies that the predictions are less accurate than everyone reporting the average probability of survival to 75. On average, the predictions display no skill. That is, individuals are not starting from the base rate and using information relevant to their own longevity to deviate from it. This is not so surprising since the respondents could not have known the base rate when they reported their survival probabilities. The result does not rule out the possibility that survival chances were estimated by combining a biased estimate of the base rate with information on personal mortality risks. The bias indicates that respondents underestimate their chances of survival to 75 by almost 9 percentage points, on average.

The Brier score for men is significantly larger than that for women. Men's predictions are worse than a 50:50 random guess. Women do no better than this low benchmark. However, men perform worse because they face a more difficult task. Their survival rate is closer to 50 percent and so they are predicting a more uncertain outcome. The skill score, which adjusts for this, is significantly below zero for both sexes, implying that both do worse than everyone reporting the sex-specific base rate, but it is significantly lower for women indicating less skill. In addition, women's predictions display much greater bias. They underestimate survival chances by around 12.7 percentage points.

---

[17] When $C=0 \rightarrow C/L=0 \rightarrow a=1$ for all P since there is nothing to lose (no cost of a false positive) and so $VS=0$. When $C=L \rightarrow C/L=1 \rightarrow a=0$ for all P since otherwise a positive cost would be incurred only to avoid the possibility of a same sized loss. So, again $VS=0$.

[18] The p-value for Spiegelhalter's (1986) one-sided test is well below 0.01 when conducted using the full sample and all sub-samples.

For both sexes, the Brier score of high school dropouts is significantly larger than that of any other education group. The predictions of the least educated are worse than a 50:50 guess. For this group, the forecast accuracy differs little between men and women. For women, the Brier score declines monotonically with rising education, indicating that prediction accuracy improves with education. The gradient is similar for men, although there is no difference between the two middle education groups.

For both sexes, the point estimate of the skill score is lowest for high school dropouts, which would suggest that their worse forecast performance as judged by the Brier score is not simply due to facing greater uncertainty because of higher mortality risk. However, this difference is not statistically significant for either sex.[19] After high school dropouts, the point estimate of the skill score is lowest (but not significantly) for college graduates. The reason this group makes more accurate predictions than the middle education groups is that it faces lower mortality risk and so can predict survival more easily. College graduates forecast survival more accurately but they are not more skilled forecasters. The skill score is significantly negative for all groups, indicating that they all do worse than equating personal risk with the sex-education specific survival rate. Of course, accurately estimating the latter is already a demanding task.

Among men, college graduates underestimate their survival chances the most, on average. However, comparison of the Brier scores makes clear that bias is not an appropriate metric to judge forecast accuracy. Male high school dropouts display less bias than college graduates because their larger positive and negative errors cancel out to a greater degree. The magnitude of the errors made by male college graduates is smaller, on average, but they are more systematically in the direction of underestimation. Among women, the average degree of underestimation is greatest for the two middle education groups.

[19] The insignificance may be due to a lack of power given the sizes of some sex-education groups. Taking men and women together, the skill score of high school dropouts is significantly lower than that of each of the two middle education groups at the 10% level.

Table 5: Predictive performance of subjective survival probabilities by sex and education

	Brier score	SE	Skill score	SE	Bias	SE	N
All	0.2588†	0.0037	-0.3079*	0.0290	-0.0878*	0.0061	6758
Male	0.2728†	0.0054	-0.2509*	0.0344	-0.0482*	0.0090	3339
Female	**0.2451**	0.0051	**-0.4139***	0.0509	**-0.1266***	0.0082	3419
Male							
dropout	0.3115†	0.0102	-0.3096*	0.0500	-0.0325	0.0168	1098
high school graduate	**0.2692**	0.0102	-0.2293*	0.0636	-0.0413*	0.0168	945
some college	**0.2719**	0.0125	-0.2619*	0.0848	-0.0286	0.0211	611
college graduate	**0.2165†**	0.0101	-0.2847*	0.1126	**-0.1001***	0.0174	685
Female							
dropout	0.3234†	0.0103	-0.5252*	0.0681	-0.1087*	0.0166	1137
high school graduate	**0.2263†**	0.0078	-0.3916*	0.0864	*-0.1471**	0.0128	1257
some college	**0.1850†**	0.0101	-0.3540*	0.1529	-0.1278*	0.0169	594
college graduate	**0.1761†**	0.0122	-0.4301*	0.2176	-0.1121*	0.0195	431

Notes: † indicates the estimate of the Brier score is significantly different from 0.25 at the 5% level. * indicates the estimate of the skill score/bias is significantly different from zero at 5%. Bold indicates that the estimate is significantly different from that of the first sub-group (males / high school dropouts) at the 1% level or less. Italics indicates the same at the 10% level.

Prediction accuracy conditional on covariates

To examine whether predictive accuracy continues to vary by sex and education after conditioning on other basic demographic and socioeconomic factors, we regress each individual's squared error $\left((p_i - y_i)^2\right)$ on sex and education, plus indicators of age (year dummies), race, marital status, employment status and wealth (see Appendix Table A6 for definitions and means). The least squares estimates in the right-hand column of Table 6 are from a pooled regression using both male and female observations. The significant coefficient on the female indicator is substantially greater in magnitude than the unconditional sex difference in the Brier scores reported in Table 5. Women appear to predict their survival chances even more accurately after controlling for other sociodemographic factors. But this because women face less mortality risk and so have an easier prediction task.

Differences by education also remain significant and substantial. For example, the Brier score of female college graduates conditional on covariates is 0.18 less than that of high school dropouts. Without conditioning, the difference is 0.15 (Table 5, column 1). For males, the conditional difference between the top and bottom groups is 0.079, which is only slightly less than the 0.095 unconditional difference. Clearly, the differences by education in prediction accuracy are not explained by correlation with other basic sociodemographics.

There appears to be no difference in prediction accuracy by race. Employed and wealthier individuals make more accurate predictions of their chances of surviving to 75.

Table 6: Regressions of squared error of predicted survival – wave 1

	Male	Female	Both
high school graduate	-0.0332**	-0.0777***	-0.0558***
	[0.0147]	[0.0135]	[0.0099]
some college	-0.0293*	-0.1129***	-0.0709***
	[0.0164]	[0.0149]	[0.0111]
college graduate	-0.0786***	-0.1174***	-0.0966***
	[0.0151]	[0.0166]	[0.0112]
white	0.0003	0.0055	0.0004
	[0.0318]	[0.0297]	[0.0218]
black	0.0373	0.0333	0.0323
	[0.0354]	[0.0321]	[0.0238]
cohabiting	-0.0257	-0.0158	-0.0200**
	[0.0175]	[0.0124]	[0.0100]
working	-0.0188	-0.0391***	-0.0321***
	[0.0128]	[0.0107]	[0.0082]
wealth	-0.0027*	-0.0037***	-0.0034***
female			-0.0477***
			[0.0079]
constant	0.3602***	0.3872***	0.4063***
	[0.0412]	[0.0352]	[0.0274]
R^2	0.0245	0.0517	0.0355
N	3339	3419	6758

Notes: Dependent variable is the squared deviation of an individual's reported probability of surviving to 75 from an indicator of whether they do survive. Estimation is by ordinary least squares. The regressions also include age dummies. The inverse hyperbolic sine transformation is applied to wealth. High school dropout is the reference category for education. 'All other' is the reference category for race. Descriptions and means of covariates in Appendix Table A6. Robust standard errors in brackets. *, **and *** indicate significance at the 10%, 5% and 1% levels.

Calibration and discrimination

Figure 3 groups observations by the reported probability of surviving to 75 and plots the proportion of each group that does survives to this age. The diameter of each circle is proportionate to the number of observations reporting each probability. If the predictions were perfectly calibrated, then all circles would line on the diagonal. Clearly, this is far from the case. The Hosmer-Lemeshow tests presented in the first column of Table 7 confirm that perfect calibration is very strongly rejected for both sexes.

Murphy's (1973) measure of reliability – the weighted average of the squared distance between each point on the calibration plot and the diagonal – presented in the second column is larger for women, indicating that their predictions are more poorly calibrated than those of men. Women underestimate their survival chances to a greater extent than men at reported probabilities below the respective base rate. More women get close to the 45^{o} line corresponding to calibrated predictions. But this appears to be a consequence of their higher survival rate, which makes prediction easier. Women reporting a probability of 0.7 and above are reasonably close to the diagonal, and yet survival chances rise very little with the reported probability in this region. Both men and women reporting probabilities above their respective mean survival rate are optimistic and men are more so.

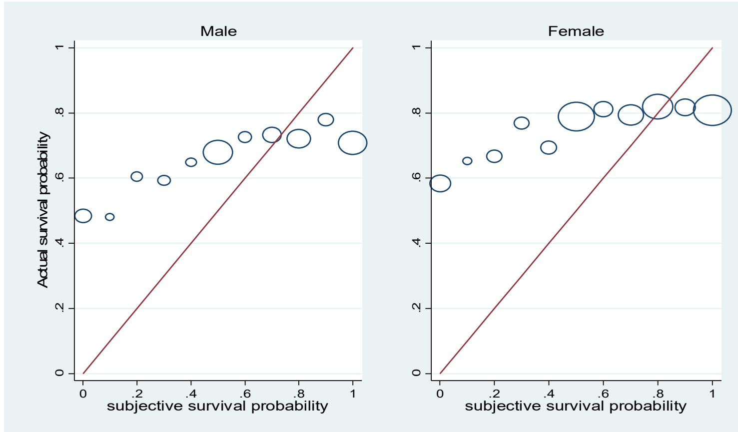


Figure 3: Actual against predicted probability of survival to 75 by sex

Note: The size of each circle is proportionate to the fraction reporting the respective probability.

Table 7: Calibration and discrimination of subjective survival probabilities

	Calibration		Discrimination				
	Hosmer-Lemeshow	Reliability	c statistic	SE	$\bar{p}_1 - \bar{p}_0$	SE	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
All	221602	0.0640	0.5766‡‡	0.0080	0.0918**	0.0087	6758
Male	118121	0.0584	0.5746‡‡	0.0107	0.0893**	0.0117	3339
Female	108458	0.0740	0.5762‡‡	0.0121	0.0919**	0.0131	3419
Male							
high school dropout	56504	0.0760	0.5667‡‡	0.0177	0.0814**	0.0208	1098
high school graduate	36151	0.0538	0.5876‡‡	0.0198	0.0971**	0.0214	945
some college	18853	0.0608	0.5483‡	0.0253	0.0651**	0.0267	611
college graduate	8345	0.0529	0.5473‡	0.0284	0.0609**	0.0257	685
Female							
high school dropout	78271	0.1119	0.5202	0.0186	0.0277	0.0219	1137
high school graduate	25040	0.0688	**0.6082‡‡**	0.0207	0.1250**	0.0222	1257
some college	4000	0.0531	**0.6035‡‡**	0.0329	0.1092**	0.0317	594
college graduate	5176	0.0513	0.5319	0.0412	0.0402	0.0355	431

Notes: Column (1) gives Hosmer-Lemeshow test statistic $\sim\chi^2(11)$. Critical value at 5% level of significance is 19.68. Test is done by grouping observations into 11 discrete values of reported probability (not quantiles). Column (2) gives Murphy's (1973) measure of reliability. ‡‡ and ‡ indicate that the c statistic is significantly greater than 0.5 at the 1% and 5% levels respectively. Bold indicates that the c statistic of a group is significantly different from that of the first group (males / high school drouputs) at the 5% level. Column (5) gives the discrimination slope, and ** and * indicate it is significantly different from zero at the 1% and 5% levels respectively.

The ROC curves for both males and females lie above the diagonal indicating that the reported probabilities do discriminate between individuals with higher and lower chances of survival (Figure 4). The area under the curve (c statistic) is significantly greater than 0.5 for both sexes (Table 7, column 3). However, the rather modest c statistics of around 0.58 for both sexes indicate that there is only a 58% chance that a randomly selected individual who survives to 75 reported a higher survival probability than did an individual who does not survive. There is little discernable divergence between the curves for men and women, indicating that there is no difference in discrimination power by gender. This is confirmed by a common discrimination slope: for both sexes, the mean survival probability reported by individuals who survive to 75 is about 9 percentage points higher than the mean probability reported by those who die before reaching this age (Table 6, column 5).[20]

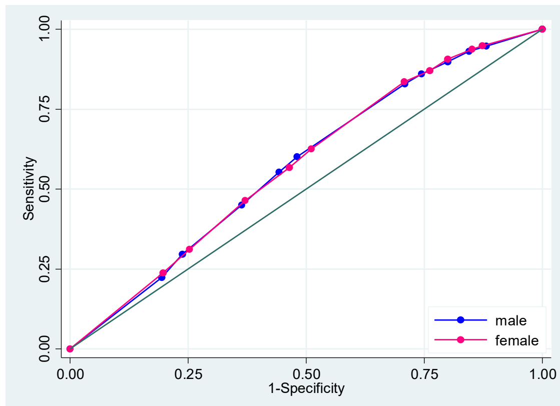


Figure 4: Receiver Operating Characteristic curves of subjective survival probabilities

For all education groups, the Hosmer-Lemeshow tests decisively reject perfect calibration of survival predictions (Table 7, column 1). Figure 5 appears to indicate that predictions of high school dropouts of both sexes are the least calibrated. This is confirmed by the larger reliability index for this group (Table 7). Four features of the calibration plots for the lowest education group stand out in comparison with the others. First, while individuals who report low probabilities grossly underestimate their survival chances in all groups, high school dropouts are more likely to make such pessimistic forecasts. Second, fewer dropouts report high probabilities, but those who do are more optimistic than higher educated individuals making such predictions. Third, as a consequence of being extremely pessimistic at the bottom of the range and highly optimistic at the top, the plot for high school dropouts is flatter than that for the other groups. Particularly for women who have not completed high school, there is little or no evidence of actual survival rising with the prediction. Fourth, unlike the other groups, there is not mass of observations lying on or close to the diagonal indicating correspondence between predicted and actual survival at that point.

Differences in prediction calibration between the other education groups are smaller. Among men, high school graduates without any college education perform reasonably well judged by the

[20] Figure A1 in the Appendix reveals that it is not only the mean of the subjective survival probability (SSP) distributions that differ by survival status. For both males and females, the median and first quartile are also lower for those who die. In aggregate and separately by gender, the SSP distribution of those who survive stochastically dominates that of those who die. Dominance is tested using Bennett's (2013) asymptotic test.

reliability measure (Table 7). This is because the calibration plot runs close to the diagonal over a substantial range of the distribution of the predictions. Male college graduates also do well because a substantial number report high probabilities that are very close to the actual chance of survival. The same is true for women with at least some college education.

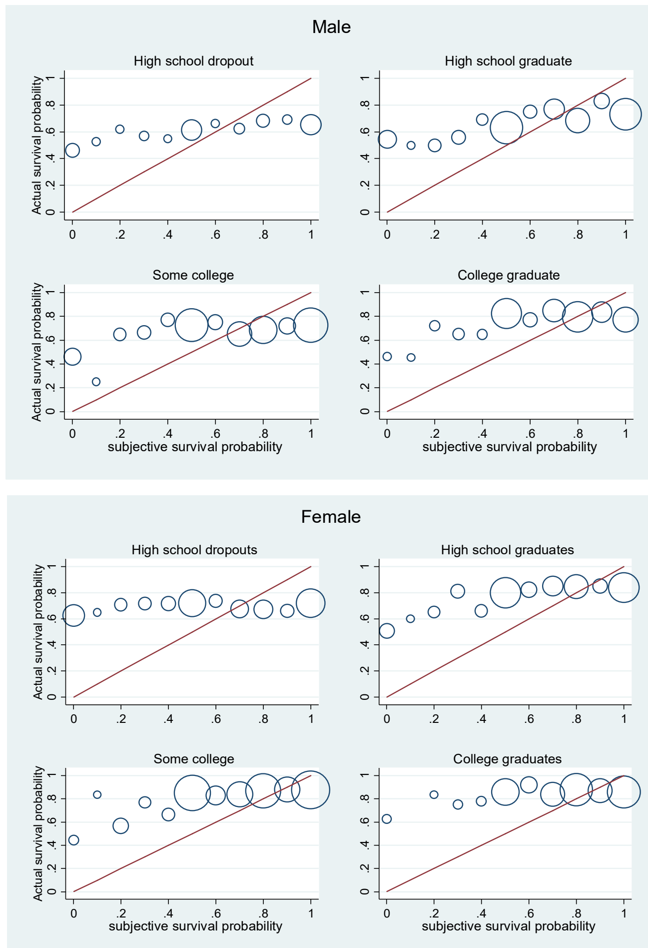


Figure 5: Actual against predicted probability of survival to 75 by education and sex

The ROC curves (Figure 6) reveal that education differences in the discriminatory power of predicted survival are quite different from those in calibration. The clearest differences are for women, with the predictions of the middle two education groups discriminating survivors from those dying before 75 considerably better than do the predictions of the bottom and top groups. The c statistics for the middle two groups imply that a survivor has around a 60% chance of having reported a higher probability of survival than that reported by someone who dies (Table 7). For the other two groups, we cannot reject that such a ranking occurs only with a random 50% chance. The mean probability of survival reported by survivors is at least 11 points greater than that reported by those who subsequently die in the two middle groups (Table 7, column 5). In the other two groups, there is no significant difference in the means of survivors and decedents. The c statistics differ significantly between each of the middle two groups and high school dropouts.

Among men, the pattern is less clear. The c statistic is significantly greater than 0.5 for all education groups, but there is no significant difference between groups. The same conclusion that the subjective survival probabilities have a similar degree of discriminatory power across education groups is reached using the discrimination slope.

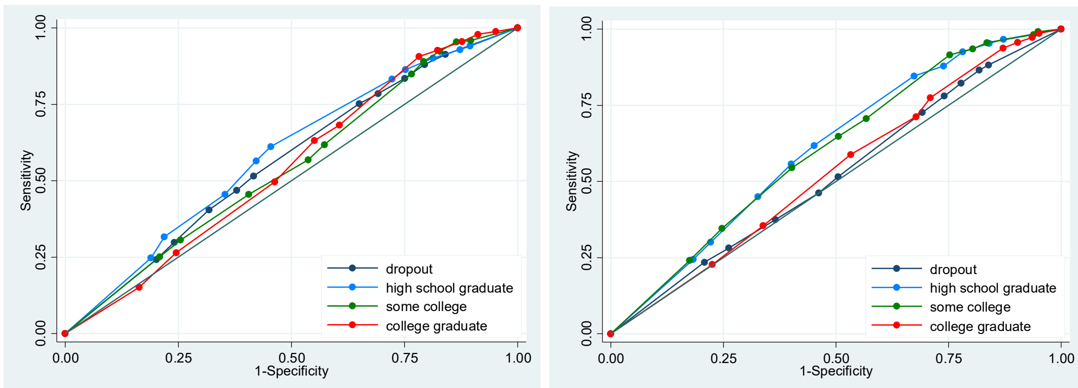


Figure 6: Receiver Operating Characteristic curves of subjective survival probabilities by education and sex

Decomposition of prediction accuracy

Table 8 shows the components of the Brier score obtained from the Yates (1982) decomposition (eq. (3)). The second column gives the variance of the binary indicator of survival to 75 and shows the greater uncertainty faced by men and lower education groups as a consequence of their higher mortality rates. The third column is the Brier score minus this variance. Consistent the skill score, women and high school dropouts display the least forecast skill by this net of variance measure. Different from the skill score, college graduates perform best among males.

The sixth column gives (twice) the covariance between predicted and actual survival, which is equal to the product of the discrimination slope and the variance of the survival indicator. This does not capture the degree to which predictions are calibrated. It is slightly larger for males than females simply because the former have a lower survival rate while survival predictions are approximately equally discriminatory for men and women. Among men, the covariance is greater for the bottom two education groups because of their lower survival and larger discrimination slopes. Among women, the covariance is larger for the middle two education groups due to the much greater discrimination power of the predictions reported by these groups.

If survival predictions are correlated with survival outcomes, then variation in the latter will drive variation in the predictions. Column (8) shows estimates of the final term in equation (3): the variation in the predictions that is not generated by information relevant to longevity. For both men and women, this noise increases monotonically as the level of education falls. This is a large part of the reason that the less educated predict their longevity less accurately. It could be that these groups respond more to irrelevant information. In which case, they would be more likely to base decisions on incorrect expectations. Another possibility is that their beliefs about mortality risks are less well defined and so they give wildly varying responses when invited to express those beliefs. In that case, they would be less likely to take account of their survival chances when making decisions that would optimally be based on expectations of longevity.

Table 7: Decomposition of Brier score for subjective survival probabilities

	Brier score (1)	Variance y (2)	(1)-(2) (3)	Bias^2 (4)	2×Covar (p,y) (5)	Minimum variance p (6)	Excess variance p (7)	N
All	0.2588	0.1979	0.0609	0.0077	0.0363	0.0017	0.0879	6758
Sex								
Male	0.2728	0.2181	0.0547	0.0023	0.0389	0.0017	0.0896	3339
Female	0.2451	0.1734	0.0717	0.0160	0.0319	0.0015	0.0861	3419
Male								
high school dropout	0.3115	0.2379	0.0736	0.0011	0.0387	0.0016	0.1097	1098
high school graduate	0.2692	0.2190	0.0502	0.0017	0.0425	0.0021	0.0890	945
some college	0.2719	0.2155	0.0564	0.0008	0.0280	0.0009	0.0827	611
college graduate	0.2165	0.1685	0.0480	0.0100	0.0205	0.0006	0.0579	685
Female								
high school dropout	0.3234	0.2120	0.1114	0.0118	0.0117	0.0002	0.1111	1137
high school graduate	0.2263	0.1627	0.0637	0.0216	0.0407	0.0025	0.0802	1257
some college	0.1850	0.1366	0.0484	0.0163	0.0299	0.0016	0.0603	594
college graduate	0.1761	0.1232	0.0530	0.0126	0.0099	0.0002	0.0501	431

Notes: Yates (1982) decomposition of Brier score (eqn. (3)).

4.2 Prediction value by sex and education

Figure 7 shows the value score of the subjective survival probabilities plotted against the cost/loss ratio. We do not show the curve at C/L<0.05 and C/L>0.95 because the large magnitudes in these regions obscure comparisons at more interesting values. Over nearly the entire range of the cost/loss ratio, the value of the subjective predictions of survival is negative for both males and females. Almost irrespective of the relative cost of insuring longevity risk through purchase of an annuity, for example, the respondents would gain more if they were to decide using the base rate probability of survival rather than the individual-specific probability they report. The subjective probabilities are not only less accurate than universal use of the base rate, which is evident from the negative valued skill scores reported in Table 5, but they also appear to be less useful for decisions of the sort described earlier.

This conclusion is subject to important caveats. Not all decisions that rational agents would take partly on the basis of their longevity risks fit into the structure modelled in section 4.2. And survival to 75 is unlikely to be the decisive outcome considered in such decisions. Further, we cannot be sure that the survival probabilities reported are those actually used, perhaps subconsciously, in decision making. The least we can infer is that if individuals are basing decisions with roughly a binary 2×2 structure on the survival probabilities they report, then they are worse off than they would be if they were to decide on the basis of the (gender-specific) population average risk rather than utilize what they believe to be information about personal risks. Of course, as emphasized earlier, the probability of survival to 75 was not known for the cohort back in 1992. An alternative measure of value would examine the gain from using subjective survival probabilities relative to basing decisions on the life table probabilities that were knowable in 1992.[21]

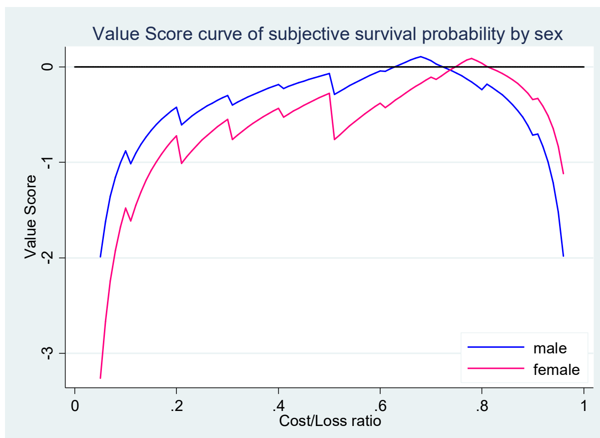


Figure 7: Value Score curve of subjective survival probabilities by sex

Note: Value Score given by equation (4). Curve is truncated at Cost/Loss ratio of 0.05 and 0.95 because values become very large in magnitude beyond these points.

At a cost/loss ratio of 0.09 the value score is about -1 for men. This implies that if men were contemplating taking an action that insured them against longevity risk at a cost equivalent to 9% of the loss, then deciding on the basis of their subjective survival probabilities (rather than using the base rate) would increase the expected loss by as much as it would be reduced if they could predict survival perfectly. For women, the same magnitude of deficit (VS=-1) is experienced when the cost/loss ratio is 0.16.

The inverted U shape of the curve is explained as follows. At low values of the cost/loss ratio it would be optimal for all to insure if they were relying on the base rate given it is 0.68 for men and 0.78 for women. This would result in a substantial number of false positive errors but this does not matter because the cost of insurance is low. If respondents were to rely on their subjective

[21] We plan to add this analysis in the next version of this paper.

probabilities, then even at this low cost those reporting a survival chance of 0 would not insure. Given that almost 50% of men and 60% of women who report that they have no chance of surviving to 75 do actually survive (see Figure 3), this results in a substantial number of false negative errors that are costly precisely because the cost/loss ratio is low. As the cost/loss ratio rises, the relative cost of the false negative errors arising from utilization of subjective survival probabilities falls while the cost of the false positive errors made using the base rate escalates. Consequently, the value score becomes less negative.

A similar logic explains the fall in magnitude of the value score as the cost/loss ratio decreases from a very high level. When the ratio is above the base rate, no one would insure if they were all relying on the base rate. This would produce many false negative errors. But the cost of avoiding these errors is high. Reliance on subjective survival probabilities allows individuals who perceive high risks to avoid false negative errors but this results in false positive errors that are expensive because the cost of insurance is almost as high as the loss insured. Note that among those reporting a survival probability of 1, less than 70% of men and 80% of women survive to 75. So, according to the decision model, 30% and 20% of males and females respectively in this group would needlessly insure at a very high cost. This brings down the gain from deciding on the basis of subjective survival probabilities rather than the base rate. As the cost/loss ratio decreases, the cost of false positive errors made using subjective survival probabilities falls and once the ratio falls below the base rate these errors are also made when that prediction model is employed.

There is an interval of the cost/loss ratio within which the value score is positive for men, and another in which it is positive for women. While these intervals appear small in relation to the full range of the ratio, they are potentially of some economic relevance. For men, the interval of positive scores runs from a cost/loss ratio of 0.63 to 0.72. For women, the respective interval is from 0.75 to 0.81. Each interval includes the mean survival rate of the sub-sample.[22] Insurance that pools across heterogeneous risks and is actuarially fair for the average risk implies a cost/loss ratio equal to the base rate. Therefore, provided the premium load is modest, insurance would be offered at a cost/loss ratio close to the base rate.[23] Risk neutral individuals would then decide

[22] In fact, for both sexes the Value Score reaches its maximum when the cost/loss ratio if equal to the base rate. This is not coincidental. It is when the decision threshold is close to the base rate that there is most scope to make gains from utilizing any person-specific information on risk incorporated in the subjective probability. When only the base rate is used, no one takes action when the cost/loss ratio is marginally above the base rate and everyone takes action when it is just equal to and marginally below the base rate. Many false negative errors are made in the first case and many false positive errors are made in the second. Provided the subjective probabilities do not display too much bias and contain relevant individual specific information, using them against a decision threshold close to the base rate will offer the greatest opportunity to cut down on the errors made through universal reliance on the base rate.

[23] In this case, the cost and loss are being defined in monetary terms.

whether to take the offer by comparing their perceived risks with a ratio in proximity to the base rate. Therefore, over an interval that is most relevant for some decisions taken by some types of individuals, it is possible that the utilization of the subjective survival probabilities does generate greater value than universal reliance on the base rate. At its maximum, the value score is 0.11 for men and 0.09 for women. This implies that use of the subjective survival probabilities can reduce the expected loss by around 10% of the reduction that would be achieved with perfect knowledge of survival.

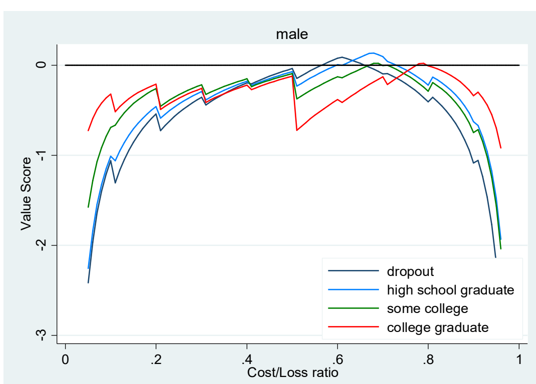
Risk averse individuals compare their perceived risks with the ratio of utility costs to utility losses, which is always less than the ratio of monetary costs to monetary losses. So the relevant interval for such agents is below the base rate, where the *VS* is predominantly negative.

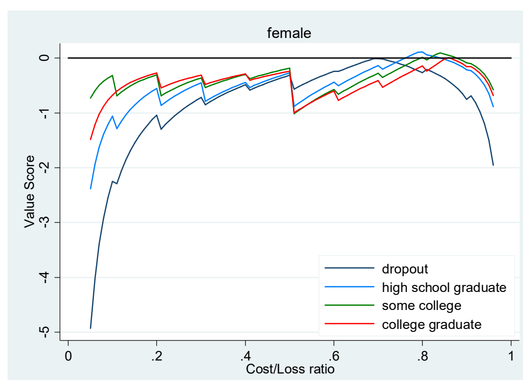
The value score curve for males lies above that for females at all cost/loss ratios up to 0.74, at which point the curves cross over. So, men generally lose less by relying on their subjective probabilities rather than the base rate. The ranking reverses at high cost/loss ratios because women have a higher survival rate. Only slightly more women than men report that they are certain to live to 75. But much fewer of these predictions made by women will be costly false positives.

For men, the value score is negative over most of the range at all levels of education (Figure 8, left panel). The greater accuracy of the survival predictions made by the more highly educated does not necessarily translate into predictions that are more valuable for decision purposes. At low cost/loss ratios, the score is less negative for the more educated men. This is because fewer of them report a subjective survival probability of zero, which results in a highly costly false negative decision to decline low cost insurance that turns out to be needed. This is a scenario in which the less educated underestimate their longevity and so are less likely to take precautions against the risk of living long enough to exhaust their wealth through consumption, medical care and nursing care needs in old age. The more highly educated are less likely to make this mistake, at least according to the model, because fewer of them report that they are certain to die before reaching 75.

As the cost/loss ratio increases, the curves quickly cross. At a ratio just above 0.5, the value score for college graduates plummets well below that of the other groups. At this point, it is high school dropouts who do least damage by relying on their predictions rather than the base rate. This is because the large proportion giving a focal response of 0.5 move from taking precautionary action to not at this threshold. These individuals would continue to insure if they were relying on the base rate. Since a larger fraction of college graduates who report this probability survive beyond 75, this group incurs more costs from false negative decisions. The cost/loss ratio interval within which the value score is positive is much narrower and the maximum positive value is much smaller for the top two education groups. It is only when the cost/loss ratio rises above 0.78 or so that college graduates regain an advantage. The reason is the same one that women have an advantage over men in this region: more college graduates giving themselves a very high chance of surviving into old age do actually survive. These individuals would lose out by relying on the base rate.

The value score curves for women by education display a similar pattern, with the high education groups suffering lower expected losses from utilization of their subjective probabilities at low and high cost/loss ratios, and greater expected losses in the middle of the range of ratios. A difference is that female high school dropouts incur much greater losses at low cost/loss ratios than their male counterparts. This is because a higher proportion of women with this level of education report a zero chance of survival and this is an even more pessimistic prediction than it is for males. Women giving this response would be much better off relying on the base rate. A further difference is that there is no cost/loss ratio interval over which the value score is positive for the most and least educated women. The middle two education groups fair better. This is due to the much better discriminatory power of the predictions made by women in these groups (see Table 7 and Figure 6). At a decision threshold close to the base rate, using personal predictions rather than the base rate is valuable if variation in those predictions is correlated with actual survival. This is what discrimination measures.





4.3 PREDICTION ACCURACY BY SEX AND COGNITION

This section compares prediction accuracy by levels of cognitive functioning using the survival probabilities reported in wave 3. In the Appendix, we use these data to estimate all the measures examined in the previous two sections by education and cognition. Here, we concentrate on the main indicators of prediction performance, which are presented in Table 8.

The Brier score calculated using all wave 3 observations is not significantly different from 0.25 indicating that the reported probabilities are as accurate as everyone making a 50:50 guess. This is a marginal improvement on the wave 1 reported probabilities, which performed worse than this low benchmark. This is due to a slight improvement in the prediction accuracy of male respondents. Consistent with the wave 1 reports, the skill score is significantly below zero for both sexes and is lower for females, indicating less skilled forecasts. Also consistent with wave, both sexes underestimate survival chances, and women do so more than men.

By cognition, the greatest difference in prediction accuracy indicated by the Brier score is between the least cognitively able and all other groups. This is particularly true among men. Among women, prediction accuracy rises in moving from the first to the second to the third quartile of cognition distribution, but does not differ between the top two quartiles. The skill score is significantly negative for both sexes only in the bottom quartile of cognitive functioning, although the point estimate is not lowest for this quartile. Bias is greatest in the two middle quartile groups for males and displays no clear relationship with cognition for females.

Table 8: Predictive performance of subjective survival probabilities by sex and cognition

	Brier score	SE	Skill score	SE	Bias	SE	N
All	0.2462	0.0045	-0.3885*	0.0412	-0.1226*	0.0069	4870
Male	0.2544	0.0067	-0.2587*	0.0490	-0.0748*	0.0108	2151
Female	*0.2397*	0.0060	**-0.5563***	0.0695	**-0.1605***	0.0089	2719
Male							
bottom quartile	0.3004†	0.0144	-0.2627*	0.0709	-0.0143*	0.0234	551
2nd quartile	**0.2395**	0.0128	-0.2156*	0.0968	**-0.1005***	0.0208	530
3rd quartile	**0.2438**	0.0129	-0.3842*	0.1224	**-0.1232***	0.0206	539
top quartile	**0.2325**	0.0127	-0.2988*	0.1265	-0.0627*	0.0208	531
Female							
bottom quartile	0.2965†	0.0138	-0.4915*	0.0958	-0.1517*	0.0200	683
2nd quartile	**0.2472**	0.0121	-0.5203*	0.1303	-0.1388*	0.0183	685
3rd quartile	**0.2077†**	0.0105	-0.6619*	0.1740	-0.1878*	0.0159	683
top quartile	**0.2064†**	0.0112	-0.7708*	0.2076	-0.1640*	0.0164	668

Notes: † indicates the estimate of the Brier score is significantly different from 0.25 at the 5% level. * indicates the estimate of the skill score/bias is significantly different from zero at 5%. Bold indicates that the estimate is significantly different from that of the first sub-group (males / high school dropouts) at the 1% level or less. Italics indicates the same at the 10% level.

Calibration measured by Murphy's reliability index does not differ greatly by cognition (Appendix Table A7). Discrimination, as measured by both the c statistic and the discrimination slope, is lowest in the top quartile of cognition for both sexes (ibid). In fact, for this group the hypothesis that the ROC curve lies on the diagonal and survival predictions do not discriminate at all between those who survive and those who die is not rejected. The Yates decomposition reveals that noise rises monotonically as cognition falls (Appendix Table A8), which is similar to the relationship of this characteristic of the predictions with education.

Overall, less cognitively able individuals do predict survival less accurately. This is not because of lower calibration, and it is certainly not because of lower discrimination power. Rather, it seems to be mainly attributable to making noisier predictions. The relationships of cognition with noise and discrimination are similar to those of education. It is the relationship with calibration that differs most between cognition and education. Prediction calibration improves with education, but not with cognition.

Table 9 presents regressions of the squared prediction error on quartiles of cognitive functioning, education and other covariates. Conditional of the other factors, prediction accuracy continues to be significantly lower in the bottom quartile of males by cognitive functioning and not to differ much between the other quartiles. For men, the difference between the Brier score of the bottom cognition group and each of the others is 51-73% of the unconditional difference. Conditional on cognition, prediction accuracy differs significantly only between the bottom and top education groups of males, with the magnitude of that difference being similar to that between the bottom cognition group and the others. For men, accuracy differs at least as much by cognitive functioning as it does by education.

Among women, accuracy differs much more strongly by education than by cognition. There are no significant differences between the cognition quartiles, conditional on education and the other covariates. Differences by education remain strong after cognition and are affected little by doing so. Pooling observations of both sexes, the education gradient in prediction accuracy is much more pronounced than the gradient by cognition. Variations in accuracy with the other covariates are similar to those observed in the wave 1 data, with an exception being that female wave 3 respondents are less accurate relative to males.

Table 9: Regressions of squared error of predicted survival – wave 3

	Male	Female	Both
Cognition			
2nd quartile	-0.0446**	-0.0046	-0.021
	[0.0197]	[0.0187]	[0.0136]
3rd quartile	-0.0344*	-0.0252	-0.0274**
	[0.0202]	[0.0182]	[0.0135]
top quartile	-0.0347*	-0.0143	-0.0221
	[0.0209]	[0.0190]	[0.0141]
Education			
high school graduate	-0.0197	-0.0715***	-0.0489***
	[0.0184]	[0.0167]	[0.0123]
some college	-0.0121	-0.1009***	-0.0617***
	[0.0205]	[0.0193]	[0.0141]
college graduate	-0.0467**	-0.1339***	-0.0895***
	[0.0204]	[0.0200]	[0.0144]
White	-0.0699	0.0386	-0.0149
	[0.0430]	[0.0365]	[0.0281]
black	-0.0638	0.0603	0.0006
	[0.0477]	[0.0397]	[0.0307]
cohabiting	-0.0149	-0.0113	-0.013
	[0.0194]	[0.0138]	[0.0111]
working	-0.0268*	-0.0443***	-0.0376***
	[0.0143]	[0.0122]	[0.0092]
wealth	-0.0016	-0.0037**	-0.0031***
	[0.0016]	[0.0015]	[0.0011]
female			-0.0298***
			[0.0092]
constant	0.3981***	0.3343***	0.3903***
	[0.0512]	[0.0427]	[0.0335]
R^2	0.0213	0.05	0.0313
N	2146	2713	4859

Notes: As for Table 6. *, **and *** indicate significance at the 10%, 5% and 1% levels.

5. CONCLUSION

Older Americans can predict their longevity only to a very limited extent. While survival to 75 is positively associated with reported chances of survival to this age, the correlation is weak. On average, individuals aged 54-65 underestimate their survival chances. On top of this systematic bias, there is also a great deal if fluctuating error. Predictions would be more accurate if everyone viewed their chances of surviving to 75 as the same as those for tossed coin showing heads. This suggests that the average individual does not hold longevity beliefs that are based on accumulated information on mortality risks.

Better educated and more cognitively able individuals predict their survival chances with greater accuracy. The predictions of the least educated are poorly calibrated and are very noisy. This heterogeneity in accuracy of longevity expectations could possibly result from differences in ability to exploit information made available by the experience of ill-health. We have previously assessed the plausibility of this explanation by testing for heterogeneity by education and cognition in the updating of longevity expectations in response to the onset of serious health conditions, such as cancer and lung disease. This reveals little support for the hypothesis, but it does show that the longevity expectations reported by high school dropouts and the least cognitively able are less stable and display greater unexplained variability (see Bago d'Uva et al. 2015).

These difficulties potentially result in sub-optimal decisions concerning household finances and health behavior. We have not tested whether there are behavioral consequences of mistaken longevity expectations. Heimer et al (2015) simulate the consequences for savings and there is evidence that behavior is correlated with longevity expectations (Hurd et al. 1998, Hurd et al. 2004, Salm 2010, Spaenjers and Spira 2015). Assuming decisions are taken in accordance with a simple decision rule, we calculate that decisions based on individual specific subjective survival probabilities generate substantially less value than decision made on the basis of the base rate alone. There are good reasons to target the least educated and cognitively able for the provision of advice regarding complex financial decisions. Our findings suggest that such advice should include help in forming more accurate expectations of longevity that determine the appropriate planning horizon.

REFERENCES

Agarwal, S. and B. Mazumder (2013) 'Cognitive Abilities and Household Financial Decision Making', *American Economic Journal: Applied Economics* 5(1): 193-207.

Anderson, N.D. and F.I.M. Craik (2000) 'Memory in the Aging Brain', in E. Tulving and F.I.M. Craik (eds) *The Oxford Handbook of Memory,* pp. 411-425. Oxford: Oxford University Press.

Bago d'Uva, T., E. Ciftci, O. O'Donnell and E. Van Doorslaer, E. (2015). 'Who can predict their own Demise? Accuracy of Longevity Expectations by Education and Cognition' (No. 15-052/V). Tinbergen Institute.

Banks, J. and Z. Oldfield (2007) 'Understanding Pensions: Cognitive Function, Numeracy and Retirement Saving', *Fiscal Studies* 28: 143-170.

Banks, J., C. O'Dea and Z. Oldfield (2010) 'Cognitive Function, Numeracy and Retirement Saving Trajectories', *The Economic Journal* 120(548): F381-F410.

Bradley AA, Schwartz SS, Hashino T. 2008. 'Sampling uncertainty and confidence intervals for the Brier score and Brier skill score'. *Weather and Forecasting* 23: 992–1006.

Behrman, J., O.S. Mitchell, C. Soo and D. Bravo (2012) 'How Financial Literacy Affects Household Wealth Accumulation', *American Economic Review* 102(3): 300-304.

Bennett C. (2013). Inference for dominance relations. *International Economic Review* 54(4), 1309-1328.

Binswanger, J. and Salm, M. (2014). 'Does Everyone use Probabilities? The Role of Cognitive Skills'. Tilburg: Department of Economics, Tilburg University.

Brier GW (1950). 'Verification of Forecasts Expressed in Terms of Probability'. *Monthly Weather Review*. 78: 1–3.

Brown, J.R, A. Kapteyn, E.F.P. Luttmer and O.S. Mitchell (2016). Cognitive Constraints of Valuing Annuities, *Journal of the European Economic Association* forthcoming

Bruine de Bruin, W., B. Fischhoff, S.G. Millstein and B.L. Halpern-Felsher (2000) 'Verbal and Numerical Expressions of Probability: `It's a Fifty Chance'', *Organizational Behavior and Human Decision Processes* 81(1): 115-131.

Bruine de Bruin, W. and K.G. Carman (2012) 'Measuring Risk Perceptions: What does the Excessive use of 50% Mean?', *Medical Decision Making* 32(2): 232-236.

Bugliari, D., Campbell, N., Chan, C., Hayden, O., Hurd, M., Main, R., Mallett, J., McCullough, C., Meijer, E., Moldoff, M., Pantoja, P., Rohwedder, S. and P. St.Clair (2016). RAND HRS Data Documentation, Version P. Santa Monica, CA.: RAND Labor and Population Program, RAND Center for the Study of Aging.

Christelis, D., T. Jappelli and M. Padula (2010) 'Cognitive Abilities and Portfolio Choice', *European Economic Review* 54(1): 18-38.

Cutler, D.M. and A. Lleras-Muney (2010) 'Understanding Differences in Health Behaviors by Education', *Journal of Health Economics* 29(1): 1-28.

Delavande, A. and S. Rohwedder (2011) 'Differential Survival in Europe and the United States: Estimates Based on Subjective Probabilities of Survival', *Demography* 48(4): 1377-1400.

Delavande, A. and S. Rohwedder (2008) 'Eliciting Subjective Probabilities in Internet Surveys', *Public Opinion Quarterly* 72: 866-891.

Dominitz, J. and C. Manski (1997) 'Using Expectations Data to Study Subjective Income Expectations', *Journal of the American Statistical Association* 92: 855-867.

Dominitz, J. and C.F. Manski (2006) 'Measuring Pension-Benefit Expectations Probabilistically', *Labor* 20: 201-236.

Elder, T. (2013) 'The Predictive Validity of Subjective Mortality Expectations: Evidence from the Health and Retirement Study', *Demography* 50(2): 569-589.

Fang, H., M.P. Keane and D. Silverman (2008) 'Sources of Advantageous Selection: Evidence from the Medigap Insurance Market', *Journal of Political Economy* 226(2): 303-350.

Fischhoff, B. and W. Bruine De Bruin (1999) 'Fifty?Fifty=50%?', *Journal of Behavioral Decision Making* 12(2): 149-163.

Fisher, G.G., H. Hassan, W.L. Rodgers and D.R. Weir (2012) 'Health and Retirement Study Imputation of Cognitive Functioning Measures: 1992 – 2010 Early Release.'. Ann Arbor, MI: Survey Research Centre, University of Michigan.

Gan, L., M.D. Hurd and D. McFadden (2005) 'Individual Subjective Survival Curves', in D.A. Wise (ed.) *Aspects of Economics of Aging,* pp. 377-412. Cambridge MA: NBER.

Hamermesh, D.S. (1985) 'Expectations, Life Expectancy and Economic Behavior', *Quarterly Journal of Economics* 100(2): 398-408.
Heimer, R.Z., K.O.R. Myrseth and R.S. Schoenle (2015). YOLO: Mortality beliefs and household finance puzzles. Brandeis University, Department of Economics, Working Paper.
Hendren, N. (2013). 'Private information and insurance rejections', *Econometrica* 81(5): 1713-1762.
Hill, D., Perry, M. and Willis, R.J. (2004). Estimating Knightian Uncertainty from Survival Probability Questions on the HRS. Ann Arbor: University of Michigan.
Hosmer, D. W., Jr., and S. A. Lemeshow (1980). Goodness-of-fit tests for the multiple logistic regression model. *Communications in Statistics* A9: 1043–1069.
Hudomiet, P. and R.J. Willis (2013) 'Estimating Second Order Probability Beliefs from Subjective Survival Data', *Decision Analysis* 10(2): 152-170.
Hurd, M.D. (2009) 'Subjective Probabilities in Household Surveys', *Annual Review of Economics* 1: 543.
Hurd, M.D., D. McFadden and L. Gan (1998) 'Subjective Survival Curves and Life Cycle Behavior', in D. Wise (ed.) *Inquiries in the Economics of Aging,* Chicago: University of Chicago Press.
Hurd, M.D. and K. McGarry (2002) 'The Predictive Validity of Subjective Probabilities of Survival', *Economic Journal* 112(442): 966-985.
Hurd, M.D. and K. McGarry (1995) 'Evaluating Subjective Probabilities of Survival in the Health and Retirement Study', *Journal of Human Resources* XXX(supplement): S268-S292.
Hurd, M.D., J.P. Smith and J.M. Zissimopoulos (2004) 'The Effects of Subjective Survival on Retirement and Social Security Claiming', *Journal of Applied Econometrics* 19: 761-775.
Lahiri K, Yang L (2013). Forecasting Binary Outcomes, in G Elliott and A Timmermann (Eds.) *Handbook of Economic Forecasting*, Vol. 2, pp 1025-1106. Amsterdam: Elsevier
Kenkel, D., S. (1991) 'Health Behavior, Health Knowledge, and Schooling', *Journal of Political Economy* 99(2): 287-305.
Kleinjans, K.J. and A. Van Soest (2014) Rounding, focal point answers and nonresponse to subjective probability questions, *Journal of Applied Econometrics* 29(4): 567-585.
Ludwig, A. and A. Zimper (2013) 'A Parsimonuous Model of Subjective Life Expectancy', *Theory and Decision* 75: 519-541.
Murphy, A. H. (1973). A new vector partition of the probability score. *Journal of Applied Meteorology* 12: 595–600.
Ofstedal, M.B., G.G. Fisher and A.R. Herzog (2005) 'Documentation of Cognitive Functioning Measures in the Health and Retirement Study', No. HRS Documentation Report DR-006. Ann Arbor, Michigan: Survey Research Center, University of Michigan.
Peters, E. (2008) 'Numeracy and the Perception and Communication of Risk', *Annals of the New York Academy of Science* 1128: 1-7.
Prentice, R. and L. Gloeckler (1978) 'Regression Analysis of Grouped Survival Data with Applications to Breast Cancer Data', *Biometrics* 34: 57-67.
Reyna, V.F. and C.J. Brainerd (2007) 'The Importance of Mathematics in Health and Human Judgment: Numeracy, Risk Communication, and Medical Decision Making', *Learning and Individual Differences* 17(2): 147-159.
Salm, M. (2010) 'Subjective Mortality Expectations and Consumption and Saving Behaviors among the Elderly', *Canadian Journal of Economics* 43(3): 1040-1057.
Schwartz, L.M., S. Woloshin, W.C. Black and H.G. Welch (1997) 'The Role of Numeracy in Understanding the Benefit of Screening Mammography', *Annals of Internal Medicine* 123(11): 966-972.
Siegel, M., E.H. Bradley and S.V. Kasl (2003) 'Self-Rated Life Expectancy as a Predictor of Mortality: Evidence from the HRS and AHEAD Surveys', *Gerontology* 49: 265-71.
Smith, J.P., J.J. McArdle and R. Willis (2010) 'Financial Decision Making and Cognition in a Family Context', *The Economic Journal* 120(November): F363-F380.
Smith, V.K., D.H. Taylor, F.A. Sloan, F.R. Johnson and W.H. Desvousges (2001) 'Do Smokers Respond to Health Shocks', *Review of Economics and Statistics* 83(4): 675-687.
Smith, V.K., D. Taylor and F. Sloan (2001) 'Longevity Expectations and Death: Can People Predict their Own Demise?', *American Economic Review* 91(4): 1126-1134.
Spaenjers, C. and S.M. Spira (2015) 'Subjective Life Horizon and Portfolio Choice', *Journal of Economic Behavior & Organization* 116(0): 94-106.

Spiegelhalter, D. J. 1986. Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine* 5: 421–433.
Steffick, D.E. (2000) 'Documentation of Affective Functioning Measures in the Health and Retirement Study'. Ann Arbor, MI: HRS/AHEAD Documentation Report DR-05. Survey Research Center, University of Michigan.
Thompson, J. C. and Brier, G. W. (1955), 'The Economic Utility of Weather Forecasts', *Monthly Weather Review* 83: 249–254
Van Doorn, C. and S.V. Kasl (1998) 'Can Parental Longevity and Self-Rated Life Expectancy Predict Mortality among Older Persons? Results from an Australian Cohort', *Journal of Gerontology: Social Sciences* 53B(1): S28-S34.
Van Santen, P., R. Alessie and A. Kalwij (2012) 'Probabilistic Survey Questions and Incorrect Answers: Retirement Income Replacement Rates', *Journal of Economic Behavior & Organization* 82(1): 267-280.
Wallace, R.B. and A.R. Herzog (1995) 'Overview of the Health Measures in the Health and Retirement Study', *Journal of Human Resources* 30(5): 84-107.
Wilks, D. (2001). A skill score based on economic value for probability forecasts. *Meteorological Applications*, 8: 209-219.
Wu, S., R. Stevens and S. Thorp (2014) 'Die Young Or Live Long: Modelling Subjective Survival Probabilities'. Sydney: School of Risk and Actuarial Studies, University of New South Wales.
Xu, J., K.D. Kochanek, S.L. Murphy and B. Tejada-Vera (2010) 'Deaths: Final Data for 2007', *National Vital Statistics Report* 58(19): 1-133.
Yates, J.F. (1982). External correspondence: Decompositions of the mean probability score. *The* 30: 132-156.
Zikmund-Fisher, B., D.M. Smith, P.A. Ubel and A. Fagerlin (2007) ' Validation of the Subjective Numeracy Scale: Effects of Low Numeracy on Comprehension of Risk Communications and Utility Elicitations', *Medical Decision Making* 27: 663-671.

APPENDICES

Appendix A: Additional Tables

Table A1: Missing on subjective survival probability (SSP) and actual survival

	Wave 1				Wave 3			
	Missing information on:			N	Missing information on:			N
	SSP to 75 (%)	Survival to 75 (%)	SSP and/or survival to 75 (%)		SSP to 75 (%)	Survival to 75 (%)	SSP and/or survival to 75 (%)	
All	2.5	2.8	5.1	7,124	5.7	1.8	7.4	5,260
Sex								
Male	2.3	2.7	4.9	3,511	4.8	1.9	6.6	2,303
Female	2.6	2.8	5.4	3,613	6.5	1.7	8.1	2,957
Education								
High school dropout	4.7	2.6	7.1	2,405	11.3	1.9	12.9	1,654
High school graduate	1.4	3.1	4.4	2,304	4.0	1.5	5.5	1,752
Some college	1.4	2.7	4.1	1,256	2.5	1.8	4.1	972
College graduate	1.2	2.6	3.7	1,159	2.4	2.3	4.5	882

Notes: First column shows the percentage who do not report a subjective probability of survival to 75. Second column shows the percentage for whom survival to 75 cannot be established at wave 12. Third column shows the percentage missing on one or both variables.

Table A2: Sample sizes by sex and education

	Wave 1			Wave 3		
	Male	Female	Total	Male	Female	Total
high school dropout	1,098	1,137	2,235	619	816	1,435
	(49.13)	(50.87)		(43.14)	(56.86)	
high school graduate	945	1,257	2,202	626	1,025	1,651
	(42.92)	(57.08)		(37.92)	(62.08)	
some college	611	594	1,205	417	515	932
	(50.71)	(49.29)		(44.74)	(55.26)	
college graduate	685	431	1,116	484	357	841
	(61.38)	(38.62)		(57.55)	(42.45)	
Total	3,339	3,419	6,758	2,146	2,713	4,859
	(49.41)	(50.59)		(44.17)	(55.83)	

Notes: Table shows observations in the estimation sample of wave 1 respondents who are not missing on subjective survival probability to 75 and vital status at age 75. Figures in parentheses are row percentages.

Table A3: Descriptives for subjective and actual probability of survival to 75 by sex – Wave 3

	Male	Female	All
Subjective probability of survival (SSP) to 75			
Mean	0.6440	0.6493	0.6470
Std. Dev.	0.3017	0.3187	0.3113
Median			
% reporting 0	6.46	8.02	7.33
% reporting 0.5	26.78	24.20	25.34
% reporting 1	23.11	25.45	24.41
Actual probability of survival to 75			
N	2,151	2,719	4,870

Notes: As for Table 2, in this case for wave 3 respondents.

Table A4: Descriptives for subjective and actual probability of survival to 75 by sex and education – Wave 3

	Males				Females			
	High school dropout	High school graduate	Some college	College graduate	High school dropout	High school graduate	Some college	College graduate
Subjective probability of survival to 75								
Mean	0.5832	0.6478	0.6604	0.7028	0.5510	0.6534	0.7214	0.7591
Std. dev.	0.3373	0.2936	0.2922	0.2548	0.3698	0.3011	0.2736	0.2234
Median								
% reporting 0	11.90	5.42	4.08	2.89	16.61	5.93	2.91	1.68
% reporting 0.5	27.65	31.26	26.38	20.21	22.34	28.79	21.17	19.61
% reporting 1	24.60	25.36	24.22	17.32	25.89	25.19	27.96	21.57
Actual probability of survival to 75	0.6559	0.7177	0.7266	0.7938	0.7265	0.8278	0.8505	0.8908
N	622	627	417	485	819	1028	515	357

Note: As for Table 3, in this case for wave 3 respondents.

Table A5: Definitions and means of covariates used in Table 6

Variable	Description	Mean Wave 1	Mean Wave 3
Wealth	Inverse hyperbolic sine of total wealth (excluding value of secondary residence)	10.813	11.160
	(standard deviation)	(4.913)	(4.826)
Working	1 if currently working for pay, 0 otherwise	0.624	0.500
White	1 if race classified as White/Caucasian, 0 otherwise	0.803	0.822
Black	1 if race classified as Black/ African American, 0 otherwise	0.165	0.148
Cohabiting	1 if married/partnered, 0 if widowed, divorced or never married	0.777	0.733

Table A6: Predictive performance of subjective survival probabilities by sex and education – wave 3

	Brier score	SE	Skill score	SE	Bias	SE	N
Male							
dropout	0.290†	0.013	-0.284*	0.075	-0.073*	0.021	622
high school graduate	*0.249*	0.012	-0.227*	0.088	-0.070*	0.020	627
some college	*0.256*	0.015	-0.287*	0.117	-0.066*	0.025	417
college graduate	**0.215†**	0.013	-0.316*	0.145	-0.091*	0.021	485
Female							
dropout	**0.323†**	0.013	-0.625*	0.091	-0.175*	0.019	819
high school graduate	**0.228†**	0.009	-0.602*	0.123	-0.174*	0.014	1,028
some college	**0.191†**	0.012	-0.501*	0.196	*-0.129**	0.018	515
college graduate	**0.151†**	0.013	-0.555	0.304	-0.132*	0.019	357

Notes: † indicates the estimate of the Brier score is significantly different from 0.25 at the 5% level. * indicates the estimate of the skill score/bias is significantly different from zero at 5%. Bold indicates that the estimate is significantly different from that of the first sub-group (males / high school dropouts) at the 1% level or less. Italics indicates the same at the 10% level.

Table A7: Calibration and discrimination of subjective survival probabilities – wave 3

	Calibration		Discrimination				
	Hosmer-Lemeshow (1)	Reliability (2)	c statistic (3)	SE (4)	$\overline{p}_1 - \overline{p}_0$ (5)	SE (6)	N (7)
All	149723.5	0.074	0.595‡‡	0.010	**0.121**	0.012	4,870
Male	57990.1	0.058	0.592‡‡	0.014	**0.109**	0.015	2,151
Female	94991.8	0.089	0.598‡‡	0.015	**0.135**	0.017	2,719
Male							
high school dropout	27130.7	0.072	0.597‡‡	0.024	**0.121**	0.029	622
high school graduate	18177.1	0.052	0.605‡‡	0.025	**0.111**	0.027	627
some college	6638.8	0.063	0.570‡	0.032	**0.082**	0.034	417
college graduate	7159.4	0.054	0.551‡	0.034	*0.065*	0.033	485
Female							
high school dropout	58194.1	0.129	0.580‡‡	0.023	**0.108**	0.030	819
high school graduate	26247.6	0.088	0.594‡‡	0.025	**0.123**	0.028	1,028
some college	8587.0	0.067	0.581‡	0.038	**0.109**	0.040	515
college graduate	3782.2	0.055	0.548	0.054	*0.067*	0.048	357
Male							
Cognition bottom quartile	25792.3	0.069	0.582‡‡	0.025	**0.102**	0.030	551
2nd quartile	10555.1	0.056	0.630‡‡	0.028	**0.156**	0.032	530
3rd quartile	13987.8	0.073	0.599‡‡	0.029	**0.092**	0.030	539
Top quartile	10782.2	0.058	0.524	0.031	*0.049*	0.030	531
Female							
Cognition bottom quartile	41268.4	0.102	0.612‡‡	0.025	**0.154**	0.033	683
2nd quartile	24714.9	0.084	0.577‡‡	0.028	**0.096**	0.032	685
3rd quartile	11420.8	0.091	0.612‡‡	0.034	**0.148**	0.038	683
Top quartile	19111.5	*0.092*	0.539	0.036	*0.075*	0.038	668

Notes: Column (1) gives Hosmer-Lemeshow test statistic $\sim\chi^2(11)$. Critical value at 5% level of significance is 19.68. Test is done by grouping observations into 11 discrete values of reported probability (not quantiles). Column (2) gives Murphy's (1973) measure of reliability. ‡‡ and ‡ indicate that the c statistic is significantly greater than 0.5 at the 1% and 5% levels respectively. Bold indicates that the c statistic of a group is significantly different from that of the first group (males / high school drouputs) at the 5% level. Column (5) gives the discrimination slope, and ** and * indicate it is significantly different from zero at the 1% and 5% levels respectively.

Table A8: Decomposition of Brier score for subjective survival probabilities – wave 3

		Brier score (1)	Variance y (2)	(1)-(2) (3)	Bias^2 (4)	2×Covar (p,y) (5)	Minimum variance p (6)	Excess variance p (7)	N
All		0.246	0.177	0.069	0.015	0.043	0.003	0.094	4870
	Male	0.254	0.202	0.052	0.006	0.044	0.002	0.089	2151
	Female	0.240	0.154	0.086	0.026	0.042	0.003	0.099	2719
Male									
	high school dropout	0.290	0.226	0.064	0.005	0.055	0.003	0.110	622
	high school graduate	0.249	0.203	0.046	0.005	0.045	0.002	0.084	627
	some college	0.256	0.199	0.057	0.004	0.033	0.001	0.084	417
	college graduate	0.215	0.164	0.052	0.008	0.021	0.001	0.064	485
Female									
	high school dropout	0.323	0.199	0.124	0.031	0.043	0.002	0.134	819
	high school graduate	0.228	0.143	0.086	0.030	0.035	0.002	0.088	1028
	some college	0.191	0.127	0.064	0.017	0.028	0.002	0.073	515
	college graduate	0.151	0.097	0.054	0.017	0.013	0.000	0.049	357
Male									
	Cognition bottom quartile	0.300	0.238	0.062	0.000	0.049	0.002	0.109	551
	2nd quartile	0.239	0.197	0.042	0.010	0.061	0.005	0.089	530
	3rd quartile	0.244	0.176	0.068	0.015	0.033	0.002	0.084	539
	Top quartile	0.232	0.179	0.053	0.004	0.018	0.000	0.067	531
Female									
	Cognition bottom quartile	0.297	0.199	0.098	0.023	0.061	0.005	0.131	683
	2nd quartile	0.247	0.163	0.085	0.019	0.031	0.001	0.095	685
	3rd quartile	0.208	0.125	0.083	0.035	0.037	0.003	0.082	683
	Top quartile	0.206	0.117	0.090	0.027	0.017	0.001	0.080	668

Notes: Yates (1982) decomposition of Brier score (eqn. (3)).

Appendix B: Additional Figures

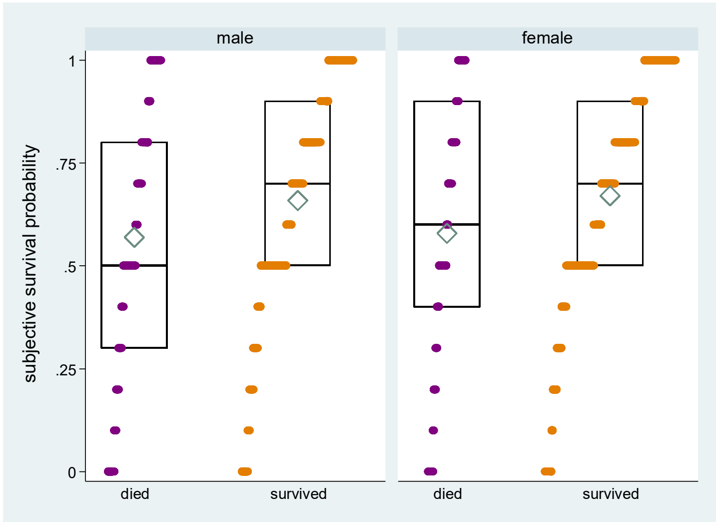


Figure A1: Strip plot of distribution of subjective survival probabilities by survival to 75

Notes: The coloured bars show the cumulative distribution of subjective survival probabilities. The lower and upper edge of each box indicates the first and third quartile respectively. The line with the box indicates the median. The diamond indicates the mean.

Appendix C: Cognition and numeracy scores

Following Ofstedal et al. (2005), we use a cognition score based on measures of *episodic memory* and *mental status*. The former enables recollection of experiences and specific events from the past (Tulving 1972). It may be required for reasoning (Smith et al. 2010). In the HRS it is assessed through two word recall tasks. The interviewer reads a list of ten common nouns (e.g., book, child, hotel) and the respondent is asked immediately to recall as many as possible in any order.[24] After five minutes or so, during which other questions are answered, the respondent is again requested to recall the words. The total recall score is the sum of the number of words recalled in these immediate and delayed recall tests (ranging from 0 to 20).

Mental status refers to the intactness of the neuro-cognitive system essential to communication and learning (Smith et al. 2010). It is measured through parts of the Telephone Interview for Cognitive Status (TICS) (Brandt et al. 1988). One dimension of mental status is *executive functioning*, which refers to the cognitive processes that facilitate the use of past experience in current action (National Center for Learning Disabilities,. 2012). These processes are used in planning, organization and management and are likely to be of considerable importance in relation to decisions taken concerning pensions, saving, health insurance, retirement and health behaviour. The cognitive processes relevant to the tasks we use from the HRS are working memory, attention and problem solving. Working memory refers to the ability to store and process information simultaneously (Baddeley and Hitch 1974). It, and the other two processes referred to, is assessed by asking respondents to subtract 7 from 100 and continue subtracting 7 from the answer for a total of five subtractions. The test score is the number of correct subtractions (0-5). This serial 7's test is part of the Mini-Mental State Examination (Folstein et al. 1975), in which it assesses the attention and calculation dimensions of mental status. A second test, which additionally assesses processing speed, asks the respondents to count backwards as quickly as possible, starting from 20 for 10 continuous numbers. Respondents are allowed two trials for this exercise. Scores are recorded as 0 if incorrect or "don't know/unable to do" on both tries, 1 if incorrect on the first try but correct on the second try, and 2 if correct on the first try. Three additional elements of mental status are assessed by orientation in time, ability to name objects and to recall the names of the President and Vice-President.[25] The total mental status score sums those of the serial 7's, backwards counting, date/object/president naming tests and ranges from 0 to 15.

The total cognition score is the sum of the total recall score and the total mental status score. It ranges from 0 to 35.

The numeracy score is derived from the following questions: 1) If the chance of getting a disease is 10 percent, how many people out of 1,000 would be expected to get the disease? 2) If 5 people all have the winning numbers in the lottery and the prize is two million dollars, how much will each of them get? 3) Let's say you have $200 in a savings account. The account earns 10 percent interest per year. How much would you have in the account at the end of two years?.

[24] To avoid respondents acquiring familiarity with the words, they are asked to recall from a different list in each successive wave.

[25] Orientation in time is tested by asking the date and the day of the week. Object naming asks questions such as: "What do you usually use to cut paper?". The score of each test is the sum the number of correct answers and ranges from 0 to 4 (date), and 0 to 2 (object and President, Vice-President). These questions are asked of all respondents in wave 3 but only of new entrants and those aged 65 and above from wave 4 onwards.

Respondents who answer 1) and 2) incorrectly are not asked 3). The numeracy score (0-3) is the sum of correct answers to the three questions. 'Don't know' responses and refusals are treated as incorrect.

References

Baddeley, A.D. and G.J.L. Hitch (1974) 'Working Memory', in G.A. Bower (ed.) *The Psychology of Learning and Motivation: Advances in Research and Theory,* (vol. 8 edn). pp. 47-89. New York: Academic Press.

Brandt, J., M. Spencer and M.F. Folstein (1988) 'The Telephone Interview for Cognitive Status', *Neuropsychiatry, Neuropsychology and Behavioral Neurology* 1: 111-117.

Folstein, M.F., S.E. Folstein and P.R. McHugh (1975) 'Mini-Mental State: A Practical Method for Grading the Cognitive State of Patients for the Clinician', *Journal of Psychiatric Research* 12: 189-198.

National Center for Learning Disabilities, (Last updated 2012) 'What is Executive Function?'. Accessed September 13 2012 <http://www.ncld.org/types-learning-disabilities/executive-function-disorders/what-is-executive-function>.

Ofstedal, M.B., G.G. Fisher and A.R. Herzog (2005) 'Documentation of Cognitive Functioning Measures in the Health and Retirement Study', No. HRS Documentation Report DR-006. Ann Arbor, Michigan: Survey Research Center, University of Michigan.

Smith, J.P., J.J. McArdle and R. Willis (2010) 'Financial Decision Making and Cognition in a Family Context', *The Economic Journal* 120(November): F363-F380.

Tulving, E. (1972) 'Episodic and Semantic Memory', in E. Tulving and W. Donaldson (eds) *Organization of Memory,* pp. 381-403. New York: Academic Press.