Measuring Probability Numeracy

Péter Hudomiet, Michael D. Hurd, Susann Rohwedder

RAND Labor & Population

WR-1270

September 2018

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

For more information on this publication, visit www.rand.org/pubs/working_papers/WR1270.html

Published by the RAND Corporation, Santa Monica, Calif.

Measuring Probability Numeracy

Péter Hudomiet (RAND)
Michael Hurd (RAND, NBER, NETSPAR)
Susann Rohwedder (RAND, NETSPAR)

Preface

This research was undertaken within RAND Labor and Population. RAND Labor and Population has built an international reputation for conducting objective, high-quality, empirical research to support and improve policies and organizations around the world. Its work focuses on labor markets, social welfare policy, demographic behavior, immigration, international development, and issues related to aging and retirement with a common aim of understanding how policy and social and economic forces affect individual decision-making and the well-being of children, adults, and families. For more information on RAND Labor and Population please contact: Krishna Kumar, Director, RAND Labor and Population, at kumar@rand.org.

Research support from the National Institute on Aging under grant P01 AG008291 is gratefully acknowledged.

Abstract

Probability numeracy is the ability of individuals to think in probabilistic terms and use probabilities effectively in everyday life. Measuring probability numeracy helps to understand how well individuals can plan for uncertainty, such as the likelihood of losing a job over the next year, and to identify groups who make suboptimal life decisions because of their limited understanding of probabilities. In this paper, we explore different batteries of questions to measure levels of numeracy. We do so by examining questions asked in the American Life Panel Financial Crisis Survey. We describe methods to create various versions of the numeracy score based on Item Response Theory. We evaluate the performance of alternative scores in predicting the quality of answers to subjective probability questions in the survey. We identify a four-item battery that captures relevant variation in probability numeracy; it takes about 90 seconds of survey time. We show that adding demographic covariates (e.g. gender, age, education, race) significantly improves properties of the score, but some additions such as education pose challenges for comparisons across cohorts or countries because of the differences in the meaning of these measures.

Table of Contents

Figures

Tables

1. Introduction

Life is full of uncertainties, and we all need to develop strategies to deal with them. These may include contingency plans, rainy day funds, insurances, social networks, or still some other system of support. In choosing which systems of support to construct and why, we need to better understand how likely a given event is to affect us.

Experts frequently use the language of probabilities to describe uncertainty. Doctors tell patients about the chances of full recoveries conditional on treatment. Financial advisors tell clients about the chances of various financial outcomes conditional on investment strategies. Newspapers talk about the chances of candidates winning an election.

Thus, understanding probabilities can make individuals more successful in planning their own futures, it signals when they should seek advice from experts, and it helps them effectively use such advice.

The concept of probabilities, however, is not easy. Understanding of this concept can vary considerably across individuals. In this work, we explore evidence of the understanding of probability among individuals, as well as the implications of this understanding.

A number of surveys, such as the Health and Retirement Study (HRS) and the American Life Panel (ALP), collect subjective probability data. They ask survey participants about the probabilities of selected future events in a percent chance format. These may include, for example, questions on the probability the respondent perceives of surviving to age 75 or working past age 65. For reviews of research on such questions, see Manski (2004), and Hurd (2009).

The main motivation for asking such questions is to use the resulting data to understand decision making under uncertainty. Yet respondents who have little understanding of probability may provide data on subjective probabilities that are not valid or reliable. Analyses of subjective probability would benefit from considering how well respondents understand the concept of probability as well.

In this work, we introduce a new survey measure of *probability numeracy*. We define this as the ability of individuals to think in probabilistic terms and to use probabilities effectively in everyday life. This probability numeracy score can be used to identify groups whose understanding is sufficient that it would be fruitful to study decision making under uncertainty using their stated subjective probabilities. Alternatively, it would identify other groups where such research is not likely to be fruitful—or, more generally, identify groups that might not be able to use probabilities in decision making for everyday life.

The main intent of this paper is methodological. We first discuss the statistical procedure we used to create probability numeracy scores. We then discuss and compare various alternatives with which we experimented to derive a numeracy score.

We used a generalized item response theory (IRT) model, estimated by Markov Chain Monte Carlo (MCMC), to derive the numeracy scores. This model assumes that probability numeracy is a latent variable that increases the chance that individuals answer questions about probabilities correctly. We call these questions about probabilities "numeracy items": they should be distinguished from questions about subjective probabilities, which are personal assessments of the probabilities of future events. By observing the responses of individuals to the numeracy items ("the test"), we predict the value of the latent variable.

Our model generalizes the standard modern item response theory model in several ways. First, our procedure models and removes question format effects. Second, we allow observable covariates (e.g. gender and education) to enter the numeracy score directly. Third, we allow repeated questions over time. Fourth, we model missing answers. Finally, we allow potential learning effects over time.

We derive a detailed numeracy score that uses all numeracy items from all three ALP waves in which our numeracy questions appeared. We then discuss how we selected a shorter battery that researchers can include in other surveys to measure numeracy. Our goals in constructing shorter batteries were to select items that 1) are most informative about the latent variable (formally we analyze the *item information functions*) and 2) provide good coverage in the sense that the resulting score successfully creates variation at all levels of numeracy (e.g. both at the top and at the bottom of the numeracy distribution).

We also discuss various alternatives we considered. First, we provide scores that do and do not include demographic predictors directly in numeracy. Including covariates in the model usually increases the quality of the score. Intuitively, this is a form of statistical discrimination: As we observe gender/race/education differences in numeracy, a good score would use that information as well to recover the expected value of the latent variable. However, using demographic information in the score makes difficult international comparison or even comparison of trends within a single country, because such a score may fail to recognize demographic differences between countries at a given time or within countries over time.

Second, we considered reweighting the score to better reflect numeracy across the entire U.S. population. Our preferred model assumes that numeracy is standardized to have zero mean and variance of one in the ALP sample. The ALP sample, however, is somewhat selective: participants are better educated and richer compared to the general U.S. population, similarly to other internet surveys (Börsch-Supan et al., 2004; Craig et al., 2013). We developed a procedure to "inverse weight" the score, so that probability numeracy would be standardized in the Current Population Survey. This model is based on somewhat strong assumptions, and as we show, it does not change the score much.

We provide practical guidance in the Appendix to researchers who may want to construct their own battery of survey questions on numeracy.

This paper is most closely related to the literature on subjective probabilities, summarized by Manski (2004) and Hurd (2009). Related numeracy batteries are discussed by Schwartz et al. (1997), Lipkus et al. (2001), and Reyna et al., (2009). Our battery is broader in scope, including joint and inverse probabilities and probabilities of subset events. There is a related literature in cognitive psychology about misperceptions, biases and heuristics in probabilistic thinking and reasoning; see, for example, Kahneman and Tversky (1972) and Kahneman et al., (1982) for early contributions, and Gilovich et al., (2002) and Chernoff and Sriraman (2013) for more recent reviews. Finally, our paper is also related to methodology literatures on modern item response theory (summarized by Linden, 2016; or Fox, 2010) and Markov Chain Monte Carlo (summarized by Gelman et al., 2013).

This paper does not focus extensively on applications. Our companion paper, Hudomiet et al. (2018), shows that probability numeracy strongly predicts the quality of subjective probabilities in surveys. For example, individuals with high numeracy predicted future unemployment more accurately. Similarly, their demand for insurance products was more related to their previously stated subjective probabilities of adverse events.

Our next section reviews the data sources we use for this work. The third section provides our methods for estimating a probability numeracy score. The fourth section reviews the results of our estimates and compares alternative scores. We conclude with a summary of this work and its implications.

2. Data

2.1. The ALP Financial Crisis Surveys

To construct our measure of probability numeracy, we placed a battery of questions in the ALP Financial Crisis Surveys, a subset of the American Life Panel (ALP) Survey. While the ALP was launched in 2006, the ALP Financial Crisis Surveys were launched in November 2008 as a high-frequency panel survey to study how the "Great Recession" affected U.S. households. The Financial Crisis Surveys covered a broad range of topics including well-being, labor force status, consumption, housing, and health, for which the ability to estimate subjective probabilities would be helpful to individuals.

The first wave of the ALP Financial Crisis Surveys had about 2,500 households. Beginning in May 2009, a monthly interview schedule was established to collect high-frequency data on the most important variables (such as spending and labor force status), while the rest of the variables were collected quarterly. Because of panel attrition, about 400 households were added in November 2011, and a larger refresher sample of about 1,500 households was added in

October and November of 2012. After April 2013, all information was collected quarterly. Data collection for the Financial Crisis Surveys subset ended in January 2016.

Altogether 4,795 individuals participated in at least one of the 61 waves of the Financial Crisis Surveys. The Surveys asked our probability numeracy questions in waves 58, 60, and 61. Each survey wave collected basic demographic information. A small fraction of the sample reported inconsistent demographics over time; for an even smaller fraction, demographics were missing in some waves. We investigated these cases in detail (see Appendix B), ultimately dropping 35 (of 4,795) cases with inconsistencies in more than one basic covariate (gender, birth year, race, ethnicity, birth country, birth state, and education). We replaced inconsistent answers with the person's mode for respondents with inconsistencies in only one demographic variable dimension, most frequently for birth year or education. There were 19 cases with completely missing basic demographics which were also dropped, leaving a sample of 4,741 people. We further restricted the sample to 2,950 individuals who participated in waves 58, 60 or 61, when the probability numeracy battery was asked. Of these, 2,878 (98%) individuals answered the probability numeracy questions at least once. This is our final sample.

The ALP provides separate survey weights for each wave. The weights are based on a raking procedure, and adjust the sample to the CPS using information about gender, age, race, number of household members and household income.[1] Our preferred weight is the person-specific mean of the survey weights from waves 58, 60 and 61, the waves when our numeracy questions were asked.

2.2. Probability numeracy questions

Altogether, we asked 13 probability numeracy questions of ALP Financial Crisis Surveys respondents in waves 58, 60, and 61. Table 1 lists all 13 of these questions. They include items on estimating the probability of a ball being drawn from a bowl where half are of one color and half of another, of interpreting weather forecast probabilities, and of interpreting the likelihood of results from flipping a fair coin.

Table 2 lists a 4-item subset of the full 13-item battery. These have two questions on interpreting ball-drawing probabilities and two questions on interpreting weather probabilities. As we will discuss, this four-item battery performs almost as well as the full 13-item battery in measuring probability numeracy. In discussing results of each of these batteries, we will also explore the arguments for the choice of each (see Section 4.3 below).

Table 3 shows the average proportion correctly answering each of the 13 questions by survey wave. Majorities correctly answered the questions on ball-drawing probabilities. In both

[1] Details can be found at https://alpdata.rand.org/index.php?page=weights.

wave 58 and wave 61, the proportion of respondents able to correctly answer the probability of no rain given a 70 percent probability of rain exceeded 85 percent.

Some questions were repeated across waves 58, 60, and 61, while others were asked only once. As a result, wave 58 had a 10-item battery, and waves 60 and 61 used a 6- and a 7-item battery respectively. Some respondents did not participate in one or more of our waves. Among those who participated in at least one wave, some skipped all probability numeracy questions, typically because they left the online survey before reaching our battery of questions, while some skipped only a few items.

We randomized question formats in two ways to study the performance of the battery under different conditions. First, in waves 58 and 60, we offered a "Don't know" option response for the numeracy questions to a random half of the sample. Those who did not have this option could still skip the questions, but most did not. As a result, offering this option led to a lower proportion of correct answers. For simplicity, Table 3 considers "Don't know" answers and skipping a question as incorrect answers. We will consider "Don't know" options more carefully when discussing creation of our probability numeracy score.

Second, we randomized the placement of the numeracy battery to study how survey fatigue affects responses. In wave 58, a random half of the sample received the probability numeracy battery relatively early in the questionnaire (at about the 10^{th} minute), and the other half answered it at the end of the instrument (at about the 30^{th} minute). We expected the latter group to provide worse answers due to elevated survey fatigue. The placement of the battery was not randomized in the other two waves: The wave 60 battery was always at the end, while the wave 61 battery was always early in the questionnaire.

The proportion of correct answers varied widely by question, indicating the battery was successful in covering different parts of the numeracy distribution. We used the proportion of respondents answering questions correctly to classify each of the 13 questions as easy, medium, or hard. The three with roughly 90 percent correct answers we classified as easy, the six with around 75 percent correct answers we classified as medium, and the four with less than 50 percent correct answers we classified as hard.

Perhaps the most striking result was the low proportion who appear to understand complex probability laws such as autocorrelation or joint probabilities. Only 14 percent of respondents could compute the joint probability of two independent events with 50 percent marginal probabilities (Q9 in Table 1). Most respondents answered 50 rather than 25 percent to this question, that is, they averaged the marginal probabilities rather than multiplying them.

Most respondents did understand basic laws about the probability scale and how frequencies and probabilities are related to one another. Still, about 10 percent of the sample incorrectly answered even the most basic questions, such as the probability of no rain given a 70 percent chance of rain (Q6 in Table 1). There is some evidence that many incorrect answers on the basic questions were due to inattention rather than lack of knowledge. Table A2 in the appendix shows correlations across the items within and between survey waves. Most questions

use a continuous answer scale, so it is not possible to guess the correct answers. Thus, those who do not understand a particular probability concept should provide incorrect answers to questions on it in all waves, and the cross-wave correlations across the items should be close to 1. We find, however, low correlations across waves for items, which we take as evidence that inattention affects the proportion answering correctly in any given wave. We find, for example, more than half (123 of 215) who incorrectly answered Q6 in wave 58 answered it correctly in wave 61.

2.3. Subjective probability response anomalies

Beyond understanding levels of probability numeracy, we seek to understand how such numeracy affects the quality of answers to subjective probability questions. We expect that a better score is more predictive of the qualities of subjective probabilities.

We define two quality indicators. Our companion paper (Hudomiet, et al., 2018) analyzes many more.

First, we define an "inconsistency score". Some subjective probability questions are related, such as whether returns on stocks will be positive next year, and whether they will be higher than 20 percent. The laws of probability require the answer to the second question to be lower than that to the first question, because the second event is a subset of the first event. Thus, if the answer to the second question is higher than the answer to the first, we conclude that the answers are inconsistent with the laws of probability. Similar related question-pairs address changes in housing and gasoline prices, retirement probabilities at different ages, and survival chances to various target ages. We chose 13 such pairs of answers that were included in many ALP waves to derive an inconsistency score for each respondent. The inconsistency score, defined for each individual in the ALP, was the simple mean of inconsistent answers to these 13 pairs from all ALP waves (maximum number of waves was 61).

Second, we created a measure about the noisiness of answers to subjective probabilities. We computed the within-person standard deviation of answers to the subjective probability of living to 75 years or more. In doing so, we sought to measure instability in these subjective probability answers. Subjective probabilities in surveys may fluctuate because of survey noise, or because of real changes in expectations. We expected relatively little change in survival chances within a short period of time among relatively young adults. We controlled for the within-person standard deviation of subjective health to capture the effect of real health shocks on survival expectations.

3. Methodology

3.1. Estimating a probability numeracy score

An appealing naïve probability numeracy score is a simple average of correct answers to the numeracy items. This naïve score has some drawbacks:

1. It ignores the difficulty of the questions, even though some questions were much easier than others.
2. The list of questions changed over survey waves, and some waves were generally easier than others. An efficient score should adjust for such cross-wave differences, because some individuals missed some of the waves.
3. Individuals were randomly assigned whether "don't know" options are offered and whether the battery appeared early in the questionnaire. The naïve score does not adjust for these issues.

We used a generalized item response theory model to estimate a probability numeracy score that took the preceding issues into account. The logic of our score was as follows. We assumed probability numeracy is a latent variable that makes individuals more likely to succeed in answering numeracy questions correctly. (This is a standard assumption in item response theory.) We further assumed that question format can also affect performance on the items. (This is a non-standard assumption.) Then from observing the performance of individuals on the probability numeracy questions we estimate this latent variable, the probability numeracy score, by Bayesian numerical methods.

Formally, we assumed that there is a latent probability numeracy score with the distribution,

$$p_i^* = \beta' x_i + u_i, \tag{1}$$

$$u_i \sim N\left(\mu, \sigma^2\right), \tag{2}$$

$$E\left[\, p_i^* \,\right] = 0 \tag{3}$$

$$V\left[\, p_i^* \,\right] = 1 \tag{4}$$

where *i* indexes individuals, x_i is a vector of observable characteristics (there may be no covariates at all), and u_i is a normally distributed unobservable part.

Then we assumed that the probability of answering a particular numeracy question correctly depends on the latent probability numeracy and the question format:

$$\Pr\left(q_{ijt} = 1\right) = \Phi\left(a_j + b_j p_i^* + \gamma_j' z_{it}\right), \tag{5}$$

where subscript *j* refers to the particular numeracy question, *t* refers to the wave, a_j is a measure of question difficulty, b_j is the slope measuring how predictive numeracy is for the question (relative to guessing), and z_{it} is a vector of variables that can affect performance on the questions beyond the respondent's numeracy: whether the "don't know" option was offered or not, whether the numeracy battery appeared early or late in the survey, and whether respondents saw the questions in earlier waves or not (potentially bringing learning effects in).

We estimated this system by Markov Chain Monte Carlo (MCMC) procedures. Appendix C discusses the estimation procedure in more details. After estimating this model, we defined the probability numeracy score as the expected value of p_i^* for each individual in the sample, conditional on individual responses to the numeracy and demographic questions:

$$p_i^s \equiv E\left[p_i^* \mid q_{ijt}, x_i \right] \tag{6}$$

This probability numeracy score is similar to the principal component of the probability numeracy questions, with several differences. First, we model and control for the effect of question formats. Second, we allow observable covariates in the numeracy score in (1). Third, we allow for missing answers in some waves (in a missing-at-random fashion). Finally, we allow questions to be repeated across survey waves, yet allowing for learning effects through z_{it}.

3.2. Using the probability numeracy score in estimation

The probability numeracy score, p^s, is only an estimate of the true latent variable, p^*; however, the difference between them is not classical measurement error because p_i^s comes from the class of optimal prediction models. As a consequence, the result of using p_i^s in estimation depends on how it is used, on the predictor variables (the x in equation (1)), and on any control variables used in the estimation. A leading example is the use of p_i^s in regression. Suppose that for some outcome y $E(y \mid p^*, z) = \beta_0 + \beta_1 p^* + \beta_2 z$ and we would like to know the β: that is, we would like to know how true probability numeracy affects some outcome such as wealth. The regression calls for p^* but, not having data on p^*, we substitute p^s. The estimated regression will yield consistent estimates of the β as long as the x in equation (1) and the z are the same. When p^* is on the left-hand side as in $E(p^* \mid w) = \gamma_0 + \gamma_1 w$ and we would like to know the γ, using p^s instead of p^* will produce consistent estimates of the γ as long as all the w are included in the numeracy equation (1). See Hyslop and Imbens (2001) and Kimball et al. (2008). These requirements make the use of p^s difficult because the estimation of p^s

needs to be adjusted for each empirical model due to the likely variation in covariates z and w from model to model.

3.3. Using demographic predictors in the probability numeracy score

Adding covariates in the probability numeracy score (such as education) in (1), can improve the quality of the score. Yet psychometricians rarely include demographic covariates in their created scores. This is for two reasons.

First, it is never clear what covariates one should include in the score: When numeracy is on the right-hand side of regressions, it is optimal to include the same set of variables in the numeracy score as in the regression. Thus, the set of needed variables changes from regression to regression, while researchers need to decide a-priori what variables to include in the model.

Second, the correlation between the covariates and numeracy may differ by population. For example, gender differences in probability numeracy may be different in the United States, Europe, and Asia. A scoring algorithm derived from one population may therefore not work in other populations. This may be particularly problematic for covariates such as race or education that have different meanings by country. Overall, for international comparisons, or even for comparisons within the same country over time, it is conceptually cleaner to use numeracy scores without any covariates.

If we do not add covariates to the model, how much bias should we expect? This depends on many factors. Our experience was that the bias was detectable but smaller when numeracy was on the right-hand side of regressions, and was more of a problem when numeracy was on the left-hand side. (We will show some of these results in Section 4.4).

Altogether, our preferred method is to include a few basic demographic covariates in the numeracy score: gender, race, age, marital status and education. But we also derived numeracy scores that do not use any covariates. These latter scores are preferred for international comparisons or within country trends.

3.4. Standardizing the score for the U.S. population

The latent probability numeracy score discussed so far is assumed to have a zero mean and a standard deviation of one in the unweighted ALP sample. An alternative (perhaps more natural) assumption is for the score to be standardized for the U.S. population. Given that the ALP sample tends to be somewhat better educated and more affluent than the general population, the two assumptions are not the same.

This section shows how the previously estimated probability numeracy score can be renormalized to be standardized in the U.S. population. The idea is to “inverse-weight” the estimated score so that weighting would make it standard.

We assume that the distribution of numeracy conditional on a set of covariates, denoted by x_i^1, is the same in the ALP and the CPS:

$$f\left(p_i^{*,ALP} \mid x_i^1\right) = f\left(p_i^{*,CPS} \mid x_i^1\right) \tag{7}$$

This is a stronger assumption than is typically used by the research on survey weights. (7) implies that the entire conditional distribution, including its mean and variance, are the same in the two samples.

We further assume that x_i^1 is a subset of x_i, the variables used in the numeracy score in (1).

$$\beta' x_i = \beta^1{}' x_i^1 + \beta^2{}' x_i^2 \tag{8}$$

Thus, (8) implies that all variables used in the weights (and possibly more) should be added to the model in (1).

Then we apply the following adjustment to the score:

$$p_i^{s,adj} = \frac{p_i^s - e_2}{s_2} \tag{9}$$

$$e_2 = E\left[\beta^1{}' x_i^{1,CPS}\right] - E\left[\beta^1{}' x_i^{1,ALP}\right] \tag{10}$$

$$s_2 = \sqrt{V\left[\beta^1{}' x_i^{1,CPS}\right] - V\left[\beta^1{}' x_i^{1,ALP}\right] + 1} \tag{11}$$

We used the following set of variables to weight: gender, race (white vs. non-white), Hispanic origin, education (high school dropout, high school, some college, college or more), age (21-30, 31-40, …, 61-70, 71+). We used the 2016 March CPS, weighting the moments by the CPS survey weights.

This procedure adjusts the numeracy score to be standardized to the general U.S. population. If desired, the estimated model parameters can also be readjusted to correspond to this adjusted score, using the following formulas (results not shown in this paper):

$$\beta_0^{adj} = \frac{\beta_0 - e_2}{s_2} \tag{12}$$

$$\beta_{-0}^{adj} = \frac{\beta_{-0}}{s_2} \tag{13}$$

$$b_j^{adj} = b_j s_2 \tag{14}$$

$$a_j^{adj} = \alpha_j + b_j e_2 \tag{15}$$

3.5. Information functions

We also sought to select a subset of the 13 probability numeracy items that are sufficient to characterize the numeracy of individuals in surveys. To select the items "optimally", we need a metric to evaluate their usefulness. Following the literature, we use information theory to guide our item selection.

For each numeracy item (question) we define the item information functions as follows:

$$I_j\left(p^*\right)=\frac{\left(\frac{\partial}{\partial p^*}\Pr\left(q_j=1\mid p^*\right)\right)^2}{\Pr\left(q_j=1\mid p^*\right)\Pr\left(q_j=0\mid p^*\right)} \tag{16}$$

(16) shows the Fisher information in item *j* at numeracy value p^*. Higher information implies that the particular item is more informative (or better) for revealing individuals' numeracy at the particular value of p^*.

In our probit type model, the item response functions have the following form:

$$I_j\left(p^*\right)=\frac{\left(\beta_j\phi\left(\alpha_j+\beta_j p^*\right)\right)^2}{\Phi\left(\alpha_j+\beta_j p^*\right)\left(1-\Phi\left(\alpha_j+\beta_j p^*\right)\right)} \tag{17}$$

The overall value of an item can be characterized by the integrated information functions,

$$I_j=\int_{-\infty}^{\infty} I_j\left(p^*\right)f\left(p^*\right)dp^* \tag{18}$$

We use numerical integration to evaluate (18). The item response functions are additive under the assumptions of the IRT model. When all items of a test are summed, we get the test information function,

$$I\left(p^*\right)=\sum_{j=1}^{J} I_j\left(p^*\right) \tag{19}$$

We set the following goals for the selection of items:

1. We want items with high integrated information
2. We want a test information function that has high information at all parts of the numeracy distribution (i.e. both at high and at low values p^*)
3. But we particularly want to cover the low end of the numeracy distribution well, so that individuals with the lowest numeracy can be separated from the rest of the sample.

4. Results

In first presenting our results, we discuss the estimated item response theory model and test some of its assumptions. We then show basic properties of the estimated numeracy score. Following that, we show how we chose a four-item battery that can suffice for researchers wishing to derive a numeracy score for individuals but not wishing, for time or other reasons, to administer our full battery of numeracy questions. We conclude our discussion of our results by comparing the performance of the various scores we propose in regression equations.

4.1. The item response theory model

IRT models are based on two assumptions that we test here. These are

1. Local independence: conditional on probability numeracy, all items are independent from each other.
2. Unidimensionality: a single factor captures the correlation structure across the items.

The test items are expected to have a positive correlation due to numeracy. That is, persons with higher numeracy are more likely to pass all test items. Local independence and unidimensionality, however, add some restrictions on the correlation matrix.

Local independence assumes that any correlation between the items is due to (latent) numeracy. It can be violated if two items are more related to one another than to the other questions. There are at least two reasons to expect a potential deviation from local independence in our case. First, some questions were repeated over time. Performance in one wave may be correlated with performance in other waves even conditional on numeracy.

Second, some items asked related questions. For example, the answer to Q2 is 100% minus the answer to Q1. It is plausible that Q1 and Q2 would be positively correlated even conditionally on numeracy. Q4 and Q5 also form a similar, potentially problematic, pair.

Table A2 in the appendix shows pairwise correlations across all 13 items from all three waves. The correlations in wave 58 between items Q1 and Q2 (0.93) as well as between Q4 and Q5 (0.97) are indeed very large. The other correlation coefficients are moderate. Even the cross-wave correlations between the repeated items appear reasonably low.

We then simulated correlation coefficients that are consistent with the model's assumptions (including local independence): 1) We fitted a model on the 13 items and three waves; 2) We recovered individual's expected numeracy given their responses; 3) We simulated new responses for each person in all waves and items using the formulas in (5); 4) We estimated the pairwise correlations across the simulated answers.

Table A3 shows the correlations between the simulated answers and Table A4 shows the difference between the empirical and the simulated correlations. Larger differences mark potential violations of local independence. Q1 and Q2 as well as Q4 and Q5 clearly violate local independence, since the simulated correlations are far lower than the empirical ones (~0.5 vs. ~0.95). The cross-wave correlations across the items, however, are very similar to the simulated ones in most cases. The only exceptions are the hard questions Q9 and Q10, where the wave 1 and wave 3 responses have a higher correlation than what the simulated model implies. These questions may test textbook knowledge, which is fixed over time.

Overall, we decided to drop answers to Q1 and Q4 when we created our preferred numeracy score, because these questions were redundant conditional on Q2 and Q5. But, for the items we retained, we kept all answers from all waves, as the cross-wave correlations among them were sufficiently low to suggest independence. We even kept the wave 1 and wave 3 versions of Q9 and Q10 to be consistent with the other questions.

To test unidimensionality (whether one factor is enough), we used factor analysis. Table 4 presents our main results. Because the items are 0-1 variables, we used polychoric correlations between them for the factor analysis. We did not use the redundant Q1 and Q4 questions in the analysis. According to the Kaiser criterion, factors with eigenvalues above 1 should be kept. This suggests a one factor model, because the eigenvalue of the second factor is 0.76. A less formal test suggests using factors after which the eigenvalues drop sharply. This also suggests a one factor model: the first eigenvalue is 4.2, and the next four are all positive but below 0.8. Yet another sometimes-used criterion is to keep factors that explain most of the variation in the test. As factor 1 already explain 84% of the variation, one factor seems enough. Altogether the model passes the unidimensionality assumption.

Finally, we discuss the convergence of the preferred MCMC model. MCMC repeatedly samples simulated parameter values, which should eventually converge in distribution to the posterior distribution of the parameters. Early simulation draws, that are not converged yet are discarded as burn-ins, and the rest of the draws are used for inference.

Testing whether convergence has been achieved is not an easy task (Cowles and Carlin, 1996). We implemented many commonly used visual inspection techniques, and we found the performance of our simulations highly satisfactory.

We ran the simulation a large number of times to assure convergence. We used 3,000,000 simulation draws and we discarded the first 10 percent as burn-in draws. The twelve panels of Figure A1 in the appendix show the time series plots of selected representative parameters. Noisy plots that appear to sample from a single distribution are said to “mix well” and are more likely converged. In contrast, random walk-like plots raise concerns about convergence. In our case all twelve parameters in the figure mix well.

The twelve panels of Figure A2 show the histograms of the simulated values of the same parameters. We prefer unimodal, smooth, bell-shaped distributions. All distributions pass this informal criterion.

The twelve panels of Figure A3 implement a visual version of the Geweke test. We plotted two kernel densities of each of the twelve parameters: one based on the first half, and one based on the second half of the simulation draws. Parameters that converged in distribution should provide similar kernel densities in the two samples. All twelve pairs of the kernel densities are right on top of each other.

Altogether, we conclude that the MCMC model converged in distribution to the posterior distribution of the parameters.

4.2. Properties of the full probability numeracy score

Panel A of Figure 1 shows the histogram of the naïve probability numeracy score (average correct answers). On average, respondents answered 66 percent of questions correctly. The distribution is strongly skewed to the left. Most respondents answered 70 to 90% of the questions correctly. Respondents typically answered the easy and medium hard questions well but missed the harder questions about the joint events and autocorrelation. A relatively small fraction of the sample missed most or even all questions: one in ten had fewer than 30% correct answers, and one in five had fewer than 50% correct answers. Given that it is hard to guess the correct answers (recall that most questions are continuous variables with infinitely many possible values), we conclude that the large majority of the ALP sample understands basic probability principles.

The mean score weighted for demographic characteristics is about 1.4 percentage point below the mean of the unweighted score. This implies that the ALP sample has about 0.06 standard deviation higher numeracy than the average U.S. population.

Panel B of Figure 1 shows scores based on a model that uses the basic demographic information in the numeracy score (gender, education, race, marital status) as well as 11 numeracy items (all items except the redundant items 1 and 4) from all three waves. This score is standardized on the ALP sample and not adjusted to the CPS. The unweighted mean of the score is zero. The standard deviation of the score is 0.9, which means that the score explains about 81% ($=0.9^2$) of the variation in latent numeracy, which is quite high. The weighted average of the score is -0.073, implying (again) that the ALP sample is a bit more numerate than the average U.S. population. This model based score is bell-shaped (a consequence of the modeling assumptions), and much less skewed than the naïve score in panel A. Nevertheless, the correlation between the model-based and the naïve scores is very high (0.94).

The scores in panel C of Figure 1 are based on the same model used to derive scores for panel B, but are standardized to the U.S. population using CPS data as discussed in Section 3.4. Using this procedure, we expected the unweighted mean score to increase and the weighted mean to be close to zero. We saw a little more change than we expected: the unweighted mean was 0.1 and the weighted mean was 0.03, or 3% of a standard deviation above the expected zero

value. Nevertheless, the discrepancy is not large, and may be a result of differences between the logic of our adjustment and ALP's weighting procedure.

The scores in panel D are based on a simpler four-item battery with no demographic information. We will later discuss how we chose the four items. Here we note that the score is based on an estimation model that uses all items from all three waves (but no demographics) but for the probability numeracy score uses only answers to items 2, 5, 6, and 9 from wave 58. The weighted and the unweighted means are close to zero. The score explains about 58% of the variation in numeracy ($=0.76^2$), which is reasonably high for a 4-item battery. Recall that the 11 items from all three waves and demographics together could explain only 81% of the variation in numeracy.

Column 1 of Table 5 presents regression versions of equation (1). It shows how some basic observable characteristics predict the latent numeracy. Other things equal, females have less probability numeracy than males, minorities less than whites, and generally less-educated people less than those who have gone to college. These coefficients are large. Female non-white high school dropouts, for example, are 1.5 standard deviations below male, white high school graduates in numeracy, and college graduates are almost 0.9 standard deviation above the numeracy of high school graduates. We also found a falling age gradient, with those over 70 doing poorly.

Table 6 shows how an individual's probability numeracy and the question format used predict performance on the 11 numeracy questions. These outputs are the regression versions of equation (5), and we report average partial effects. The output of the probit models are in Table A1 in the appendix.

We expected the coefficients on numeracy to be large and positive since numeracy should help individuals answer the test questions correctly. We did indeed find numeracy to be a significant positive predictor of performance on all questions except Q8 on autocorrelation (asking respondents the probability of rain on a subsequent day following a day of rain). There were five answer options and individuals needed to identify the correct 26-49% range. It turned out to be a very hard question, and performance appeared to be mostly random, independent of performance on any other questions.

Across the set of 11 questions, those with the strongest association with probability numeracy was Q5 on white balls. This is perhaps the most useful question for measuring probability numeracy, because it explains most of the variation in it. Performance on other questions was in between these cases.

Placing the numeracy battery earlier in the survey instrument led to better performance. On Q5, for example, the probability of a correct answer was as much as 5.5 percentage points higher for those who saw the question earlier. Our explanation for this disparity is that probability numeracy questions are hard and respondents can answer them better when they are less tired. Other supporting analyses included offering "don't know" options, which led to significantly worse performance in most cases (fewer correct answers; recall that DK answers

are coded as incorrect). We saw weak learning effects: those who answered questions for the second or third time did a little better on average, but the differences were usually not statistically significant.

4.3. An optimal 4-item battery

Because an 11-item battery would still take a substantial amount of time to administer, we explored whether a smaller subset of our items would perform almost as well as the full set.

Figure 2 illustrates how we considered and selected questions. Its six panels show the item information functions and how the probability of correct answers vary by latent numeracy.

The questions form three distinct groups. Items 7, 8, and 11, as shown in panels E and F, are the least-informative questions about numeracy, and thus, they are the least important to be included in the battery. All of these questions were very hard, and apparently performance on them did not vary much with performance on the other items. Item 8 was the single least-informative question, in accordance with our findings in Table 6.

Items 2, 5, and 6, as shown in panels A and B, are the most informative items. They have the highest item information and the highest integrated information functions (the latter is not shown). Even though they are most informative at the low end of the numeracy distribution, they have considerable discriminatory power even at the median (zero). Items 2 and 5 have the additional advantage of using the same setup (drawing balls from bowls), making it less burdensome to administer both to respondents.

The remaining items (3, 9, 10, 12, 13) are moderately informative, as shown in panels C and D. Items 3 and 13 are quite informative at the very low end of the numeracy distribution. There are relatively few respondents in that range and the other items already cover the low end of the distribution well, so we do not find these items particularly useful. Item 9 is very informative at the top end of the distribution. This question tested computation of joint probabilities. Only 14% of the sample could answer this question correctly, and this question seems very useful to distinguish the best performers in particular. Items 10 and 12 are potentially useful: they are moderately informative about numeracy throughout its distribution. Of the two we prefer item 10, because item 12 may not work well with item 9 in a short battery.

Given the results above, we concluded

- Items 2, 5 and 9 form our preferred 3-item battery. Though item 6 is more informative than item 9, item 9 is needed to cover the top of the distribution.
- Our preferred 4-item battery adds item 6.
- Our preferred 5-item battery adds item 10.
- Our preferred 6-item battery adds item 3. This battery uses all items from wave 58 (the first wave we included numeracy items) except the redundant items 1 & 4 and the non-informative items 7 & 8.

Figure 3 shows the test information functions based on tests with different number of items. The 4-item battery, shown in the black dashed line, appears to be a good compromise: this is our preferred short battery. It is substantially more informative than the 3-item battery, but only marginally less informative than the 5-item battery. The 6-item battery makes the score more informative at the bottom, but where the test is already reasonably good. Nevertheless the 6-item battery is our preferred long battery. The 11-item battery is moderately more informative than the 4-item battery but at a relatively large cost (more than doubling survey time). All tests are most informative at around the 20th percentile of the numeracy distribution, which means that the tests are best at identifying who is below or above the 20th percentile.

4.4. Comparing alternative scores

Having derived alternative scores, we seek to compare the performance of several simulated probability numeracy scores in regression analysis. The main goal is to test the 4-item battery. To do so, we use

- A total score based on all items (except redundant 1 & 4) from all three waves and basic demographic covariates.
- Three versions of a 4-item battery:
  - Only using four numeracy items (excluding any demographic information)
  - Using the four numeracy items and including demographic information
  - Including demographic information and adjusting the score to be standardized in the CPS
- The four item batteries are only based on wave 58 responses.

Table 5 shows regressions with numeracy on the left-hand side and basic demographic covariates on the right hand side. Column 1 is based on the MCMC model's output. Because the MCMC model used the latent variable on the left-hand side (in a multiple imputation fashion using MCMC), the coefficients are unbiased and we use them as benchmarks. Scores shown are

- Total numeracy score in column 2
- Total numeracy score restricted to the four-item score available only in wave 58 in column 3
- Three differing versions of the four-item score in columns 4-6

Previous theory suggests the coefficients in columns 2-5 should be consistent, and the coefficients in column 6, which includes no demographic variables, should be biased toward zero. This is exactly what we found. The coefficients in columns 2-5 are very similar to those coefficients in column 1 (the benchmark), while the coefficients in column 6 are about 40% lower in absolute value than those in other columns. The bias therefore is substantial when demographic variables are not directly included in the numeracy score. According to theory all standard errors in columns 2-6 are biased toward zero, because these models do not take into

account the fact that the scores are only estimates of the latent score. The standard errors are indeed smaller in columns 2-6 compared to column 1, but they are similar to each other.

Table 7 shows regressions with numeracy on the right hand side. Columns [1] – [4] refer to inconsistent answers to questions about subjective probabilities. The average rate of inconsistent answers is 0.072. Black or Hispanic respondents averaged 0.036 and 0.026 more inconsistent answers; the rate was essentially flat by age except for the age band 21-30. The main interest is the relationship between probability numeracy and the rate. Based on the 11-item battery, a one standard deviation increase in the score reduces the rate of inconsistent answers by 0.035.

Columns [5] – [8] refer to the individual level standard deviation in the subjective probability of survival to age 75 as assessed over many waves in the ALP, which we think of as mainly a measure of noise. The most noticeable coefficient is on the standard deviation of health: those who experienced high levels of health variability reported high levels of survival variability. As for probability numeracy, an increase of one standard deviation in the 11-item probability numeracy battery is associated with a reduction of 3.0 in the standard deviation of subjective survival on a base of about 11.0.

We find that the 4-item batteries that include basic demographics perform quite well: the coefficients in columns 2-3 are similar to those in column 1, and those in columns 6-7 are similar to those in column 5. The 4-item battery that excludes demographic information does less well: the coefficients on numeracy are about 20% lower compared to the coefficient on the total score. The adjustment to the CPS seems to matter very little in regression analysis.

Despite the favorable results of Table 7, we should be cautious about conclusions: according to the theory we outlined in section 3.2, the estimates are not consistent because the control variables are not the same as the variables included in the numeracy scores. We do not have a benchmark, which would require jointly estimating the numeracy model and the regressions in the MCMC model. But we expect the 4-item batteries to be more biased than the 11-item total score. Our goal here is to compare the performance of the 4-item batteries to the total score.

5. Conclusion

Identifying numeracy capabilities can help in analysis of how individuals respond, and are able to respond, under conditions of uncertainty. To measure probability numeracy, the ALP Financial Crisis Survey included a 13-item battery of questions on the topic. In this paper, we showed how the items can be optimally combined into a probability numeracy score using a generalized Item Response Theory model estimated by Markov Chain Monte Carlo.

The model handled the varying difficulty and relevance of the items, and allowed us to assess the effects of different question formats and repeated items in the panel. We also identified advantages and disadvantages of different modeling approaches. In particular, we showed that adding demographic covariates directly to the model improves its measurement properties, although this makes the model difficult to use for populations in other settings because of how demographic characteristics may be defined or how their influence operates elsewhere. We derived a numeracy score that is standardized for the ALP sample, but we showed methods to standardize it for the general U.S. population.

We also derived short 4-item batteries that maximize the information content of the items. We compared the performance of various 4-item batteries as left- and as right-hand variables in regressions. We found that 4-item batteries that included basic demographic covariates performed very well both as left- and as right-hand variables. The 4-item battery that excluded demographic covariates performed noticeably worse, but qualitatively similarly to the more complex scores.

Altogether, the recommended 4-item battery seems sufficient to capture significant variation in probability numeracy. We suggest further investigation on understanding how less-numerate individuals make life-decisions involving risk and uncertainty. We also suggest further investigation on how probability numeracy can be used to improve survey measures of probabilistic expectations. To facilitate such research, we include in the Appendix to this paper a step-by-step guide to use our numeracy items.

Tables

Table 1. List of the original 13 probability numeracy questions used in the ALP

Intro	Now we would like to ask you some questions to find out how much you are at ease with questions about the chance of something happening or not happening. Consider a bowl with 10 balls in total. Some of the balls may be white and some red.
Q1	First, suppose this bowl has 10 white balls and no red balls. You will be asked to draw one ball without looking. On a scale from 0 percent to 100 percent, what is the percent chance that the ball you draw is white?
Q2	On a scale from 0 percent to 100 percent, what is the percent chance that the ball you draw is red?
Q3	Now suppose that the bowl has 7 white balls and 3 red balls. You will be asked to draw one ball without looking. What is more likely? That the ball you draw is red, or that the ball you draw is white?
Q4	On a scale from 0 percent to 100 percent, what is the percent chance that the ball you draw is red?
Q5	What is the percent chance that the ball you draw is white?
Q6	Imagine that the weather report tells you that the chance it will rain tomorrow is 70%. Assuming the weather report accurately reports the chance of rain, what is the chance it will NOT rain tomorrow?
Q7	Suppose the chance it will rain tomorrow is 70%. Someone tells you that the chance it will rain both today and tomorrow is 80%. Is this possible?
Q8	Imagine that if it rains one day, then it is more likely that it will also rain the next day. If the chance of a rainy day is 50%, then what could be the chance of two rainy days in a row? 1. 0-25%; 2. 26-49%; 3. 50-59%; 4. 60-69%; 5. 70-100%
Q9	Imagine that whether it rains in your town and whether it rains in Paris are unrelated. The chance that it will rain in your town tomorrow is 50%. The chance that it will rain in Paris is also 50%. What is the chance that it will rain both in your town and in Paris tomorrow?
Q10	Imagine your friend has a FAIR coin, that means that when flipping this coin the chance of it coming up heads is the same as the chance of it coming up tails. Imagine that your friend has flipped this fair coin 3 times, and each time it came up heads. What is the chance that the next result will be a tail?
Q11	Suppose that the chance of a sunny day is 80%. Also suppose that a sunny day is more likely to be followed by another sunny day. If it is sunny today, what is the percent chance that tomorrow will also be sunny?
Q12	Suppose that whether it rains in your town and whether it rains in Paris are unrelated. The chance that it will rain in your town tomorrow is 10%. The chance that it will rain in Paris tomorrow is also 10%. If it does rain in your town tomorrow, what is the percent chance that it will rain in Paris tomorrow?
Q13	You flip a fair coin twice. If the first flip is head, what is the percent chance that the second will be tail?

Table 2. The 4-item probability numeracy battery

Intro	Now we would like to ask you some questions about the chance of something happening or not happening. We would like you to give a number from 0 to 100, where "0" means that you think there is absolutely no chance, and "100" means that you think the event is absolutely sure to happen. Consider a bowl with 10 balls in total. Some of the balls may be white and some red.
F1	First, suppose this bowl has 10 white balls and no red balls. You will be asked to draw one ball without looking. On a scale from 0 percent to 100 percent, what is the percent chance that the ball you draw is red?
F2	Now suppose that the bowl has 7 white balls and 3 red balls. You will be asked to draw one ball without looking. What is the percent chance that the ball you draw is white?
F3	Imagine that the weather report tells you that the chance it will rain tomorrow is 70%. Assuming the weather report accurately reports the chance of rain, what is the chance it will NOT rain tomorrow?
F4	Imagine that whether it rains in your town and whether it rains in Paris* are unrelated. The chance that it will rain in your town tomorrow is 50%. The chance that it will rain in Paris* is also 50%. What is the chance that it will rain both in your town and in Paris* tomorrow?

* We recommend replacing “Paris” with “New York” in European surveys.

Table 3. Average correct answers to the numeracy questions by wave, ALP, weighted

				Avg. correct answers		
#	Keyword	Question	Difficulty	Wave 58	Wave 60	Wave 61
Q1	10-WHITE-BALLS-A	10 white balls, no red. Probability draw is white?	Medium	0.768	-	0.808
Q2	10-WHITE-BALLS-B	10 white balls, no red. Probability draw is red?	Medium	0.774	0.759	0.802
Q3	7-WHITE-BALLS-A	7 white, 3 red. Which is more likely?	Easy	0.879	0.868	-
Q4	7-WHITE-BALLS-B	7 white, 3 red. Probability of red?	Medium	0.702	-	0.743
Q5	7-WHITE-BALLS-C	7 white, 3 red. Probability of white?	Medium	0.701	0.681	0.744
Q6	INVERSE-PROB	Chance of rain is 70%. Probability of not rain?	Easy	0.871	-	0.893
Q7	SUBSET-EVENT	Chance of rain is 70%. Can chance of rain both today and tomorrow be 80%?	Hard	0.243	-	-
Q8	AUTOCORR-JOINT	Positive autocorrelation in rain and 50% marginal. Probability of rain two days in a row can be what?	Hard	0.151	-	-
Q9	JOINT-PROB	Chance it rains in your town and Paris are both 50% and independent. Probability of raining in both cities?	Hard	0.136	-	0.142
Q10	MEAN-REVERT-A	Fair coin comes up head 3 times. Probability of next one being tail?	Medium	0.677	-	0.718
Q11	AUTOCORR-COND	Chance of sunny day is 80% and positive autocorr. If sunny today, what can be prob of sunny tomorrow?	Hard	-	0.377	-
Q12	INDEPENDENCE	Chance it rains in your town and Paris are both 10% and independent. If rains in your town, what is prob of raining in Paris?	Medium	-	0.644	-
Q13	MEAN-REVERT-B	Fair coin comes up head. Probability next is tail?	Easy	-	0.865	-
		Mean, by waves		0.590	0.699	0.693
		N		2475	2305	2308

Table 4. Factor analysis of the 11 probability numeracy items, N=2040

Factors	Eigenvalue	Proportion of explained variance
Factor1	4.20414	0.8392
Factor2	0.76219	0.1521
Factor3	0.42182	0.0842
Factor4	0.22065	0.044
Factor5	0.07167	0.0143
Factor6	-0.02054	-0.0041
Factor7	-0.03218	-0.0064
Factor8	-0.10318	-0.0206
Factor9	-0.10379	-0.0207
Factor10	-0.13587	-0.0271
Factor11	-0.27536	-0.055

*The sample consists of individuals who answered all numeracy items except for 1 & 4 (which had very high correlations with items 2 & 5); only the first answers are used for those who answered the items in multiple ALP waves

Table 5. Predictors of latent probability numeracy and of alternative simulated numeracy scores, ALP, unweighted

	Model output	Total score		Simulated 4-item scores		
	demog	demog		adjusted	demog	no demog
	W1-W3	W1-W3	W1	W1	W1	W1
	[1]	[2]	[3]	[4]	[5]	[6]
Female	-0.378	-0.38	-0.374	-0.388	-0.4	-0.244
	[0.038]***	[0.027]***	[0.029]***	[0.023]***	[0.024]***	[0.028]***
Non-white	-0.601	-0.603	-0.619	-0.571	-0.588	-0.366
	[0.048]***	[0.037]***	[0.042]***	[0.034]***	[0.035]***	[0.042]***
Hispanic	-0.322	-0.322	-0.326	-0.309	-0.319	-0.198
	[0.053]***	[0.040]***	[0.045]***	[0.038]***	[0.039]***	[0.047]***
High school dropout	-0.572	-0.573	-0.555	-0.549	-0.566	-0.411
	[0.097]***	[0.081]***	[0.097]***	[0.077]***	[0.079]***	[0.095]***
High school graduate	ref.	ref.	ref.	ref.	ref.	ref.
Some college	0.343	0.344	0.316	0.322	0.332	0.215
	[0.052]***	[0.039]***	[0.041]***	[0.034]***	[0.035]***	[0.044]***
College	0.883	0.884	0.869	0.865	0.892	0.542
	[0.055]***	[0.039]***	[0.042]***	[0.034]***	[0.035]***	[0.042]***
Age 21-30	ref.	ref.	ref.	ref.	ref.	ref.
Age 31-40	-0.198	-0.195	-0.237	-0.226	-0.233	-0.157
	[0.079]**	[0.056]***	[0.064]***	[0.052]***	[0.053]***	[0.063]**
Age 41-50	-0.219	-0.219	-0.298	-0.251	-0.259	-0.179
	[0.080]***	[0.057]***	[0.064]***	[0.051]***	[0.052]***	[0.062]***
Age 51-60	-0.249	-0.244	-0.306	-0.291	-0.3	-0.204
	[0.076]***	[0.054]***	[0.060]***	[0.048]***	[0.050]***	[0.058]***
Age 61-70	-0.305	-0.302	-0.352	-0.342	-0.352	-0.238
	[0.077]***	[0.053]***	[0.059]***	[0.048]***	[0.049]***	[0.058]***
Age 71+	-0.529	-0.526	-0.59	-0.595	-0.613	-0.411
	[0.083]***	[0.060]***	[0.066]***	[0.053]***	[0.055]***	[0.064]***
Single	-0.187	-0.178	-0.19	-0.164	-0.169	-0.096
	[0.038]***	[0.028]***	[0.030]***	[0.024]***	[0.025]***	[0.029]***
Constant	0.260	0.254	0.337	0.416	0.323	0.224
	-	[0.061]***	[0.068]***	[0.054]***	[0.056]***	[0.066]***
R-squared	-	0.41	0.402	0.498	0.498	0.215
N	2,878	2878	2475	2475	2475	2475

*** p<0.01, ** p<0.05, * p<0.1. The probability numeracy measures use all items from all waves except the redundant items 1 & 4. Columns 1 is based on the model's output. The standard error of the constant is not estimated, because the constant is normalized so that the score has zero mean in the ALP. Columns 2-6 are based on simulated probability numeracy scores: the expected value of probability numeracy conditional on answers to numeracy items (11 items from all waves or 4 items from wave 58) and with or without basic demographic covariates. Column 4 adjusts the score to be standardized on the CPS.

Table 6. Average partial effects of probability numeracy and question format on the probability of correct answers to the individual numeracy items, ALP

	Q2	Q3	Q5	Q6	Q7	Q8
	10-WHITE-BALLS-B	7-WHITE-BALLS-A	7-WHITE-BALLS-C	INVERSE-PROB	SUBSET-EVENT	AUTOCORR-JOINT
Probability numeracy	0.226	0.123	0.292	0.165	0.041	0.003
	[0.007]***	[0.008]***	[0.008]***	[0.010]***	[0.010]***	[0.008]
Battery was placed early	0.048	0.021	0.055	0.045	-0.004	-0.014
	[0.008]***	[0.009]**	[0.008]***	[0.009]***	[0.018]	[0.014]
DK option was offered	-0.024	-0.049	-0.026	-0.022	-0.050	-0.014
	[0.010]**	[0.008]***	[0.010]***	[0.011]*	[0.017]***	[0.014]
Saw the question earlier	0.011	0.002	0.015	0.006	-	-
	[0.008]	[0.008]	[0.009]*	[0.010]	-	-
Observations	2,878	2,878	2,878	2,878	2,878	2,878

	Q9	Q10	Q11	Q12	Q13
	JOINT-PROB	MEAN-REVERT-A	AUTOCORR-COND	INDEPENDENCE	MEAN-REVERT-B
Probability numeracy	0.131	0.204	0.119	0.195	0.120
	[0.009]***	[0.008]***	[0.011]***	[0.009]***	[0.008]***
Battery was placed early	0.017	0.042	-	-	-
	[0.012]	[0.015]***	-	-	-
DK option was offered	0.013	-0.077	-0.023	-0.128	-0.054
	[0.011]	[0.017]***	[0.020]	[0.018]***	[0.011]***
Saw the question earlier	-0.007	0.060	-	-	-
	[0.010]	[0.013]***	-	-	-
Observations	2,878	2,878	2,878	2,878	2,878

1. *** p<0.01, ** p<0.05, * p<0.1. For the continuous probability numeracy variable, the table shows the effect of a one standard deviation increase in numeracy on the probability of correct answers. For the dummy variables it indicates the effect of a discrete jump from zero to one. "Battery was placed early" indicates that the numeracy items appeared in the middle of the survey as opposed to at the end; "DK option was offered" indicates that "do not know" option was offered at the question; and "Saw the question earlier" indicates that the person had already been asked the particular numeracy question at an earlier wave. The probability numeracy measure is based on our preferred model with basic demographic variables included in numeracy and it uses all items from all waves except the redundant items 1 & 4.

Table 7. OLS regressions of the fraction of inconsistent subjective probability answers and the within person standard deviation of subjective survival probabilities, comparing the predictive power of full 11-item numeracy score and simple scores based on only four numeracy items, ALP, weighted

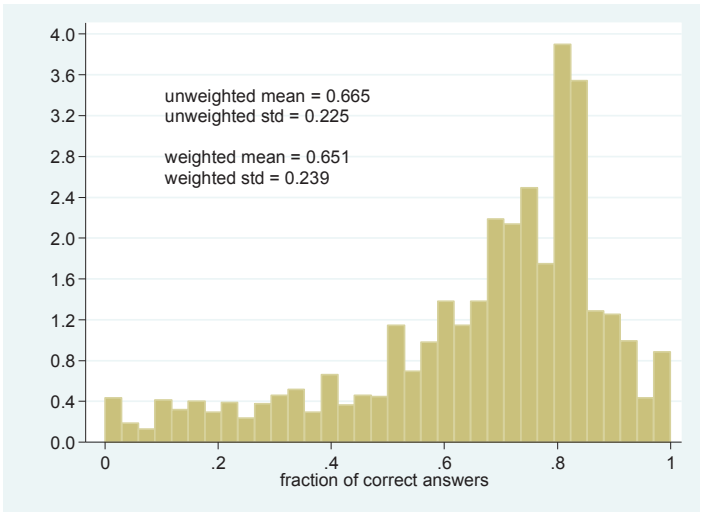
	Inconsistent				Sd(Survive to 75)			
	[1]	[2]	[3]	[4]	[5]	[6]	[7]	[8]
Numeracy, total score	-0.035				-3.022			
	[0.003]***				[0.248]***			
Numeracy, 4-item, with demog, adjust		-0.035				-2.934		
		[0.004]***				[0.313]***		
Numeracy, 4-item, with demog, unadjust			-0.034				-2.848	
			[0.004]***				[0.304]***	
Numeracy, 4-item, no demog				-0.029				-2.386
				[0.004]***				[0.256]***
Female	0.001	0.000	0.000	0.006	-1.647	-1.695	-1.695	-1.140
	[0.004]	[0.005]	[0.005]	[0.004]	[0.358]***	[0.373]***	[0.373]***	[0.357]***
White	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.
Black	0.036	0.040	0.040	0.048	-0.440	-0.086	-0.086	0.696
	[0.010]***	[0.010]***	[0.010]***	[0.010]***	[0.613]	[0.629]	[0.629]	[0.603]
Other race	-0.006	-0.005	-0.005	0.005	0.230	0.264	0.264	1.102
	[0.009]	[0.009]	[0.009]	[0.009]	[0.627]	[0.640]	[0.640]	[0.629]*
Hispanic	0.026	0.027	0.027	0.031	0.378	0.440	0.440	0.856
	[0.008]***	[0.008]***	[0.008]***	[0.008]***	[0.538]	[0.549]	[0.549]	[0.542]
High school dropout	0.020	0.020	0.020	0.027	1.164	1.195	1.195	1.864
	[0.014]	[0.015]	[0.015]	[0.015]*	[0.755]	[0.774]	[0.774]	[0.761]**
High school graduate	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.
Some college	0.001	0.001	0.001	-0.004	0.059	0.106	0.106	-0.335
	[0.005]	[0.006]	[0.006]	[0.006]	[0.448]	[0.460]	[0.460]	[0.452]
College	-0.003	-0.005	-0.005	-0.019	-0.222	-0.441	-0.441	-1.740
	[0.006]	[0.006]	[0.006]	[0.005]***	[0.503]	[0.531]	[0.531]	[0.473]***
Age 21-30	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.
Age 31-40	-0.024	-0.025	-0.025	-0.021	-1.461	-1.429	-1.429	-1.129
	[0.009]***	[0.010]**	[0.010]**	[0.010]**	[0.581]**	[0.592]**	[0.592]**	[0.588]*
Age 41-50	-0.039	-0.039	-0.039	-0.035	-1.536	-1.410	-1.410	-1.085
	[0.010]***	[0.010]***	[0.010]***	[0.010]***	[0.620]**	[0.629]**	[0.629]**	[0.626]*
Age 51-60	-0.038	-0.038	-0.038	-0.034	-2.001	-1.890	-1.890	-1.505
	[0.009]***	[0.009]***	[0.009]***	[0.009]***	[0.621]***	[0.632]***	[0.632]***	[0.627]**
Age 61-70	-0.030	-0.031	-0.031	-0.026	-3.473	-3.396	-3.396	-2.949
	[0.009]***	[0.010]***	[0.010]***	[0.009]***	[0.656]***	[0.668]***	[0.668]***	[0.662]***
Age 71+	-0.025	-0.026	-0.026	-0.017	-	-	-	-
	[0.010]**	[0.010]**	[0.010]**	[0.010]*	-	-	-	-
Married	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.
Divorced	0.004	0.004	0.004	0.007	0.433	0.375	0.375	0.628
	[0.007]	[0.007]	[0.007]	[0.007]	[0.528]	[0.537]	[0.537]	[0.535]
Widowed	0.023	0.023	0.023	0.026	1.441	1.237	1.237	1.494
	[0.012]*	[0.012]*	[0.012]*	[0.012]**	[1.370]	[1.389]	[1.389]	[1.390]
Never married	0.007	0.008	0.008	0.011	-0.161	-0.112	-0.112	0.168
	[0.007]	[0.007]	[0.007]	[0.007]	[0.486]	[0.494]	[0.494]	[0.492]
CESD depression score	0.001	0.001	0.001	0.001	0.654	0.657	0.657	0.657
	[0.003]	[0.003]	[0.003]	[0.003]	[0.197]***	[0.200]***	[0.200]***	[0.200]***
Health excellent	0.005	0.006	0.006	0.006	0.061	0.178	0.178	0.186

	[0.009]	[0.009]	[0.009]	[0.009]	[0.646]	[0.656]	[0.656]	[0.656]
Health very good	0.005	0.004	0.004	0.004	0.025	-0.100	-0.100	-0.104
	[0.005]	[0.005]	[0.005]	[0.005]	[0.414]	[0.420]	[0.420]	[0.420]
Health good	ref.	ref.	ref.	ref.	ref.	ref.	ref.	ref.
Health fair	0.005	0.004	0.004	0.004	0.651	0.522	0.522	0.533
	[0.007]	[0.007]	[0.007]	[0.007]	[0.556]	[0.563]	[0.563]	[0.564]
Health poor	0.000	0.000	0.000	0.000	0.092	0.231	0.231	0.197
	[0.012]	[0.013]	[0.013]	[0.013]	[0.861]	[0.873]	[0.873]	[0.874]
Standard deviation of health					5.710	6.395	6.395	6.394
					[0.801]***	[0.808]***	[0.808]***	[0.808]***
Number of times asked					-0.010	-0.020	-0.020	-0.020
					[0.039]	[0.039]	[0.039]	[0.039]
Constant	0.080	0.084	0.081	0.077	11.763	11.897	11.596	11.212
	[0.012]***	[0.013]***	[0.013]***	[0.013]***	[1.109]***	[1.132]***	[1.127]***	[1.123]***
R-squared	0.227	0.204	0.204	0.205	0.209	0.186	0.186	0.186
N	2420	2420	2420	2420	2021	2021	2021	2021

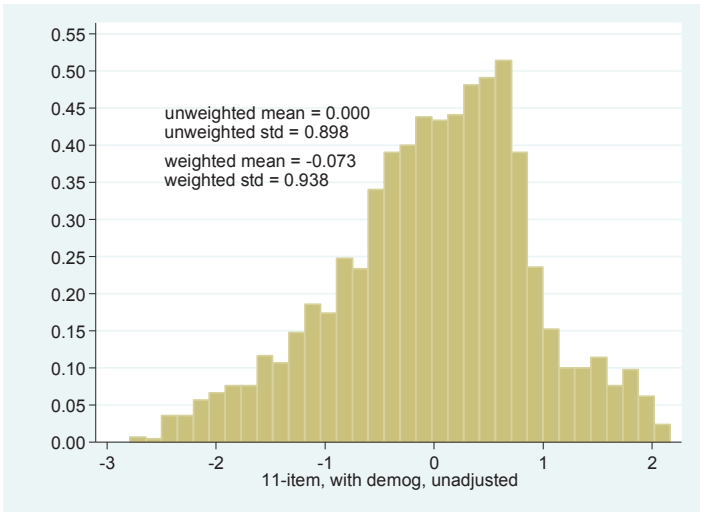
*** p<0.01, ** p<0.05, * p<0.1. The total probability numeracy measure is based on our preferred model with basic demographic variables included in numeracy and it uses all items from all waves except the redundant items 1 & 4. The 4-item batteries use items 2, 5, 6 and 9 from wave 58 only and they either do or do not include basic demographic covariates in numeracy. The sample is restricted to participants in wave 58.

Figures

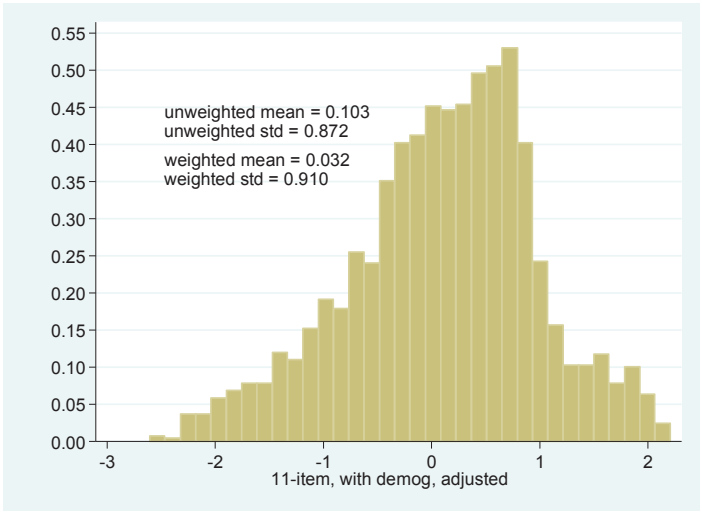
Figure 1. Histograms of probability numeracy: average correct answers and three model-based scores, ALP



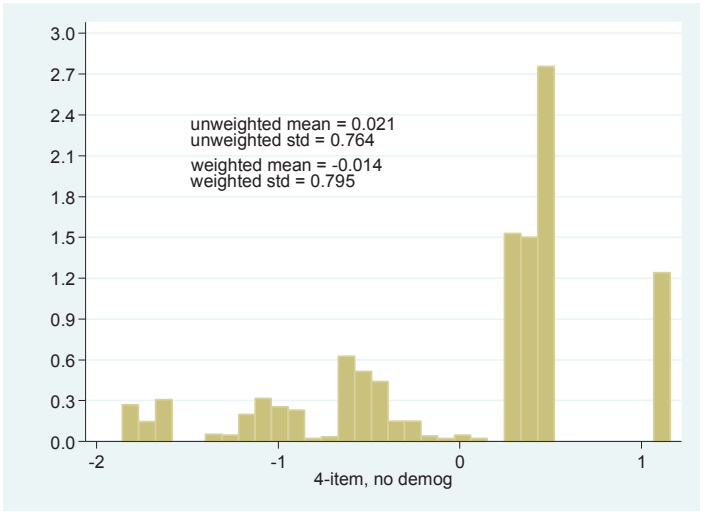
Panel A: Naïve score, average correct answers to all questions



Panel B: 11-item score with basic demographics



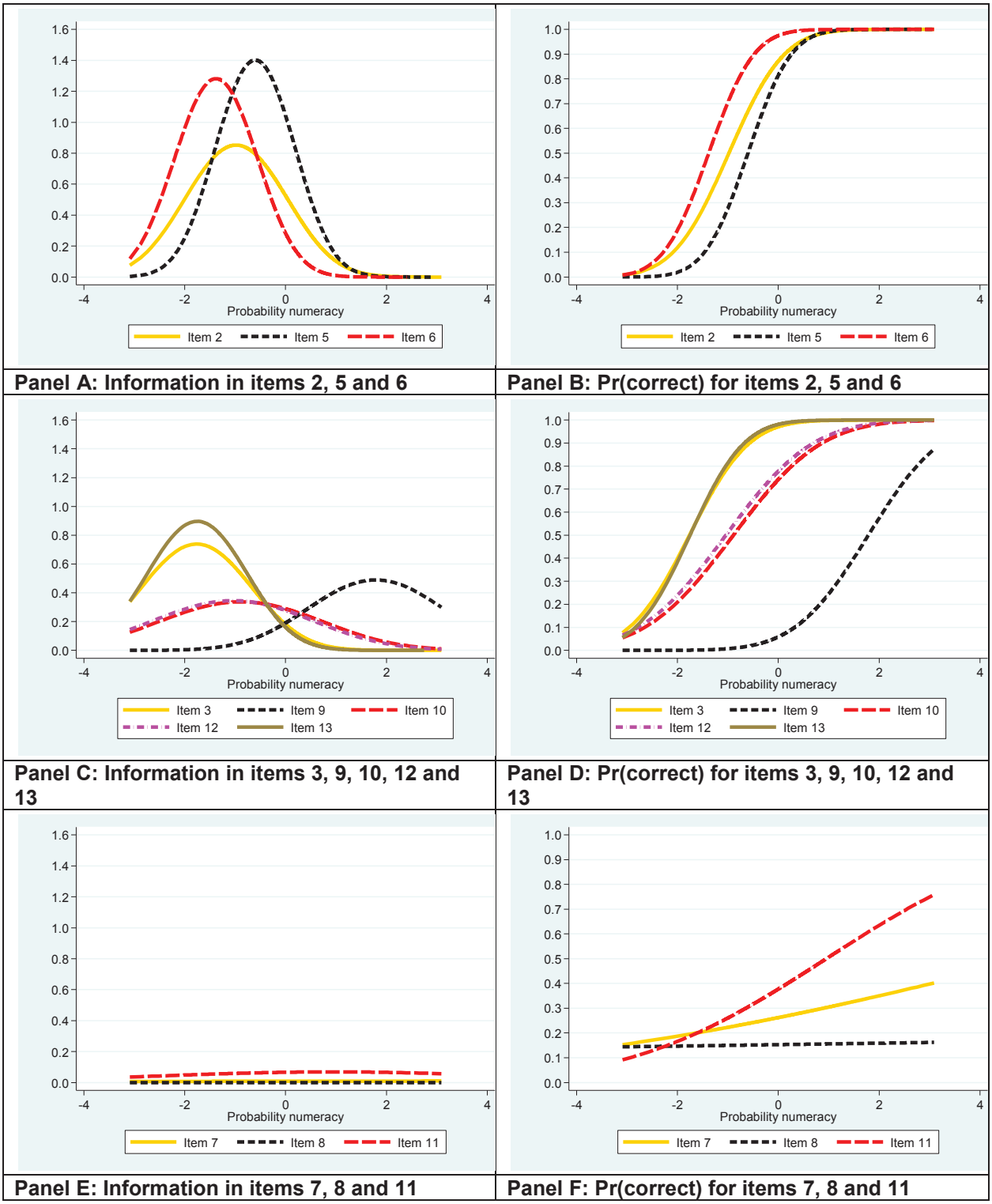
Panel C: CPS adjusted 11-item score with basic demographics



Panel D: 4-item score, no demographics, wave 58

*The scores in panels B and C are based on the same model: basic demographic variables are included in numeracy and all items from all waves are used except the redundant items 1 & 4. The score in panel B does not adjust the score to the CPS and it represents the ALP population. The score in panel C is adjusted to the CPS and represents the U.S. population. The score in panel D is estimated using a similar model, with the exception that demographic information is ignored. When the score is simulated after estimation, only 4 items (#2,5,6,9) from wave 58 are used.

Figure 2. Information functions and probabilities of correct answers to each numeracy items as a function of probability numeracy, ALP



Panel A: Information in items 2, 5 and 6

Panel B: Pr(correct) for items 2, 5 and 6

Panel C: Information in items 3, 9, 10, 12 and 13

Panel D: Pr(correct) for items 3, 9, 10, 12 and 13

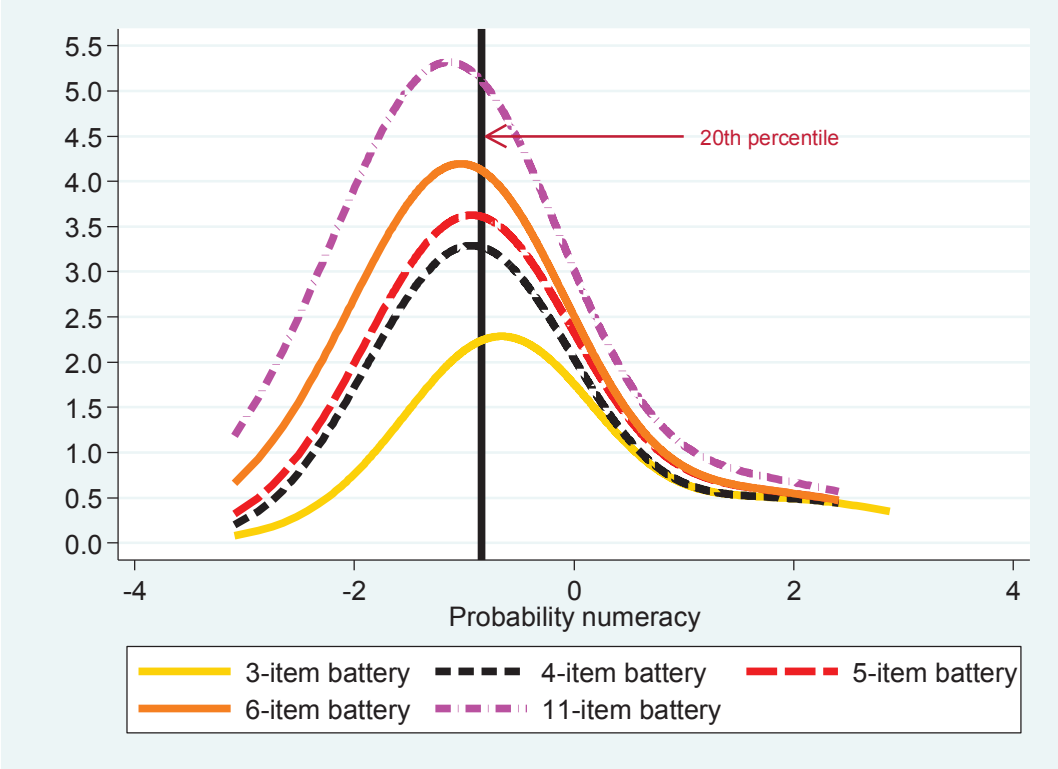
Panel E: Information in items 7, 8 and 11

Panel F: Pr(correct) for items 7, 8 and 11

* The information function shows the Fisher information in the numeracy items. Intuitively it identifies the part of the numeracy distribution where the particular items are most useful for revealing the numeracy of individuals. High values indicate more usefulness. The probability numeracy measure is based on our preferred model with basic

demographic variables included in numeracy and it uses all items from all waves except the redundant items 1 & 4.

Figure 3. Test information functions using alternative sets of items, ALP



* The test information function is the sum of the individual information functions, and it shows which part of the numeracy distribution is best identified by the test. The 3-item battery uses items 2, 5 and 9; the 4-item battery adds item 6; the 5-item battery adds item 10; the 6-item battery adds item 3; and the 11-item battery uses all items except the redundant items 1 & 4. The black vertical line shows the 20th percentile of probability numeracy in the unweighted ALP sample.

Appendix A: Adding the probability score to other surveys: a step-by-step approach

To facilitate research on probability numeracy our project website includes Stata data files containing the probability numeracy scores that can be merged (on the survey answers and potentially demographics) to other datasets. The website includes alternative versions of the scores, summarized in Table A5.

This appendix provides a step-by-step guide to use our questions and our scoring. We note that our questions have only been tested and validated on internet surveys. Numeracy questions may be more difficult to answer in telephone and face-to-face surveys unless respondents receive visual aids.

Step 1: How many items to include?

We find a 4-item battery is sufficient to capture the most important variation in numeracy, and it is relatively cheap to add to surveys. The median length of the introduction and the four questions were 95 seconds on the ALP. We also include 5- and 6-item batteries for those who seek additional precision. While additional questions are available, they have shortcomings, as explained in the main text, and provide little information that cannot be discerned from the six recommended items.

Step 2: Alternative question formats

We do not recommend offering "Don't know" (DK) options for the questions, but we do recommend allowing respondents to skip questions if they wish. Our website only provides scores for batteries that do not include DK options.

We recommend placing the battery early in the questionnaire because these cognitively demanding questions are easier to answer with a fresher mind. Our scoring assumes that the questions were included at about 10 minutes into the survey. We also provide scores that assume a later placement (about 30 minutes). These scores are a bit higher to compensate individuals for the added difficulty.

Step 3: With or without demographics

Each score comes in two forms: one with and one without including basic demographic information (gender, education, race, marital status) in the scores. We showed that scores including such covariates are more precise and less biased in regression analysis, but they may only work in a population similar enough to the ALP sample.

We are concerned about using demographic information for scores among respondents outside the United States. Education systems and racial composition of populations differ by country, and the gender and age differences in probability numeracy may differ as well. Using demographics for scores outside the United States would bias estimated demographic differences in numeracy towards the U.S. differentials. We therefore recommend using the scores without demographics in other countries. An interesting extension of our work could be to test whether demographic differences in the particular

countries are similar to those in the United States. If they are, then using the score with demographics may be justified.

Step 4: Standardizing the score

Our preferred score is assumed to be standard normal, with a zero mean and a standard deviation of one in the ALP sample. We showed that the weighted average of numeracy is negative in the ALP, on average, because the ALP sample is more numerate than the general U.S. population.

We derived and implemented a procedure that assumed that numeracy is standard normal in the U.S. population. It has some important drawbacks:

- The procedure was somewhat complicated
- It requires using demographic covariates in the scores, which may not be advised for other reasons (see Step 3)
- It matters very little in regression analysis, because the adjustment is basically a shifter (to the right).

We therefore do not recommend using this adjustment. Instead, our weighted results should be used for comparisons, with notation that the score is standard normal in the ALP sample, which is somewhat more affluent than the general U.S. population.

Our website does, however, include scores with the CPS adjustment.

Step 5: Using the score in regressions

Probability numeracy is a latent variable not directly observed (or even observable) in surveys. The probability numeracy *scores* we provide are only estimates of the latent variable, and they are subject to prediction errors. Section 3.2. briefly discussed the effect of prediction error on the consistency of regression coefficients. The coefficients may or may not be consistent and it is not possible to come up with a scoring that works in all cases. Hyslop and Imbens (2001) and Kimball et al. (2008) discuss these issues in more detail. They also discuss how the standard errors of the coefficients can be adjusted for prediction error.

Appendix B: Details of the ALP Financial Crisis Survey and sample definitions

Basic demographic information is collected in each ALP survey wave. A small fraction of the sample has inconsistent demographics from one wave to the next, and for an even smaller fraction, demographics is missing from certain waves.

We investigated these cases in detail. We looked at seven dimensions: gender, birth year, race (aggregated to white, black, or other), Hispanic ethnicity, U.S. birth, birth state, and education (aggregated to less than high school, high school, some college, or college). Inconsistency in education was defined as any decrease in the level of education based on the four aggregate levels. Inconsistency in the other dimensions was defined as any change.

We dropped 35 persons (out of 4,795, 0.7% of all cases) who had inconsistencies in more than one dimension, reducing the sample size to 4,760 individuals and 134,077 person-year observations. By looking at their answers, we judged most of these cases highly suspicious: it is likely that the wrong person filled out the survey at least once, or the person identifiers got mixed up.

Observations with inconsistencies in only one dimension were kept in the sample. The most frequent change was in education, but there have been many other smaller discrepancies. For example, Puerto Ricans sometimes stated that they were born in the United States, sometimes stated that they were born abroad; or people changed their race from white to other. We defined their real demographic values as the mode (their most frequent answer). From multiple modes we chose the more recent one. Even though an increase in the level of education is not inconsistent per se, we replaced education with its mode as well; 298 persons had at least one change in the level of education, and 170 persons had at least one change in any of the other dimensions. Lastly, we dropped 19 persons for whom demographic information was completely missing in at least one dimension. Our final sample had 4,741 people and 133,983 person-year observations.

Appendix C: MCMC estimation of the probability numeracy score

C.1. The logic of MCMC

MCMC is a Bayesian estimation method that is particularly useful for estimating models with latent variables in a hierarchical structure. Gelman et al. (2013) provide a thorough overview of MCMC and other similar techniques. MCMC is similar, in some ways, to maximum simulated likelihood. It requires a fully specified model, but it can be used in more complex cases when simulating the likelihood function would be too tedious or too slow. The unusual feature of MCMC is that it assumes that the parameters of the model being estimated are random, such as the person effects in a random-effect panel regression model. The data-generating process is described by the c.d.f. $F(y,\alpha)$, where *y* represents the data and α represents the unknown parameters to be estimated.

The goal of the procedure is to derive the posterior distribution of the parameters, $F(\alpha \mid y)$. Once the posterior is estimated, one can use its mean as the estimate, and its standard deviation as the standard error of the estimate:

$$\hat{\alpha} \equiv E_F\left[\alpha \mid y\right] \tag{20}$$

$$se(\hat{\alpha}) \equiv std_F\left[\alpha \mid y\right] \tag{21}$$

Alternatively, one can directly define a 95% "confidence interval" by taking the 2.5th and 97.5th percentiles of $F(\alpha \mid y)$, which is sometimes called the 95% *credible interval*. We used the means and standard errors from (20) and (21).

The posterior distributions of the parameters are simulated with the following procedure: First, take an initial guess at all parameters α_j. The set of parameters includes all latent variables, such as the probability numeracy of individuals, as well.

Second, define a prior distribution for each parameter, $f_0(\alpha_j)$. For regular parameters researchers usually choose so-called non-informative priors, such as normal distributions with mean zero and variance 1000. For latent variables the prior distribution comes from the model. For example, if probability numeracy had a standard normal distribution, our prior for each individual's numeracy would be the standard normal distribution itself (the person can be anywhere in the population).

Third, draw new values of the parameters from their posterior distribution conditional on their prior, the data, and all other parameters, $F\left(\alpha_j \mid y, \alpha_{-j}, f_0(\alpha_j)\right)$. One can update the parameters individually, jointly, or in blocks.

Fourth, continue updating the parameters a large number of times, N^S. According to theory, under mild regularity conditions these simulation draws will converge in distribution to the true posterior distribution $F(\alpha \mid y)$.

Finally, drop the first N^B burn-in draws, and use the rest of the $N^S - N^B$ draws for inference.

The most challenging part of the estimation is to derive (or approximate) the posterior distributions of all parameters, but in standard cases there are readily available formulas. In particular, there are formulas for the posterior of the means and variances of normally distributed variables, and for coefficients of linear regressions with normally distributed errors. When all posterior distributions can be derived analytically, the simulation procedure is called the Gibbs sampler. When at least one posterior distribution is not possible to derive (or not efficient to do so), one can simulate the posterior instead, using, for example, the Metropolis-Hastings algorithm. In this project we use the regular Gibbs sampler.

C.2. Our Gibbs sampler

Now we describe the steps of our Gibbs sampler. We first slightly rewrote the model to the following form:

$$p_i^* = \beta' x_i + u_i, \tag{22}$$

$$u_i \sim N(0,1), \tag{23}$$

Note that in the original model we assumed that p_i^* has zero mean and variance 1, while here we assume the same about the residual. This alternative model is easier to estimate, and after estimation we can renormalize the model to the original form.

Our prior about each individual's numeracy is just this distribution. Numeracy predicts performance on each numeracy question according to

$$\Pr(q_{ijt} = 1) = \Phi(a_j + b_j p_i^* + \gamma_j' z_{it}), \tag{24}$$

As standard in the literature, we augment this model with another set of latent variables, q_{ijt}^* in the following way:

$$q_{ijt}^* = a_j + b_j p_i^* + \gamma_j' z_{it} + \varepsilon_{ijt} \tag{25}$$

$$\varepsilon_{ijt} \sim N(0,1) \tag{26}$$

$$q_{ijt} = \begin{cases} 0 & \text{if } q_{ijt}^* < 0 \\ 1 & \text{if } q_{ijt}^* \geq 0 \end{cases} \tag{27}$$

Now we discuss how to derive the posterior distributions of all these parameters. The posterior of the latent q^*_{ijt} variables (conditional on everything else) is a censored normal distribution. It is censored on the left if the person answered the particular question correctly, and it is censored on the right if he answered incorrectly. To update q^*_{ijt}, we simply draw a random value from these censored normal distributions.

Now it is easy to update the parameters a_j, b_j and γ_j (conditional on everything else), because we just need to run a Bayesian linear regression of the latent q^*_{ijt} on probability numeracy and z_{it}, which is standard, and then draw random values from the posterior multivariate normal distribution. We can then update the probability numeracy score of each individual. The prior is given by (22) and (23), the data are the individual's answers for the numeracy questions. We can rewrite (25) as

$$c_{ijt} \equiv \frac{q^*_{ijt} - a_j - \gamma_j' z_{it}}{b_j} = p^*_i + \upsilon_{ijt} \tag{28}$$

$$\upsilon_{ijt} \sim N\left(0, \frac{1}{b_j^2}\right) \tag{29}$$

To update probability numeracy of individual i, we use (28). The left hand side is the data (it depends on observed variables and simulated parameters) and the right hand side is numeracy plus normally distributed error terms. The posterior distribution is in a standard form again, as numeracy is the mean of normally distributed variables. Finally, to update the predictors of numeracy (β) we run a Bayesian regression of p^*_i on xi.

In our preferred model we used 3,000,000 simulation draws and we dropped the first 300,000 burn-in draws.

C.3. Recovering the probability numeracy score

Our goal is to recover our best guesses for the probability numeracy of individuals, defined as

$$p^s_i \equiv E\left[p^*_i \mid q_{ijt}, x_i \right] \tag{30}$$

The simulation draws of p^*_i are converging in distribution to the distribution of p^*_i. Thus p^s_i can be estimated by saving all simulation draws of p^*_i and then taking the average of them for each individual separately. Alternatively, one can simulate the expected values outside the MCMC procedure using the model estimates and the data.

C.4. Renormalizing the model

After estimating the model we renormalize all coefficients so that probability numeracy has zero mean and variance 1 using the following formulas:

$$e_x \equiv E[\beta' x_i] \tag{31}$$

$$v_x \equiv V[\beta' x_i] \tag{32}$$

$$p_i^{*,final} = \frac{p_i^* - e_x}{\sqrt{1+v_x}} \tag{33}$$

$$\beta_0^{final} = \frac{-e_x}{\sqrt{1+v_x}} \tag{34}$$

$$\beta_{-0}^{final} = \frac{\beta_{-0}}{\sqrt{1+v_x}} \tag{35}$$

$$b_j^{final} = b_j\sqrt{1+v_x} \tag{36}$$

$$a_j^{final} = \alpha_j + b_j e_x \tag{37}$$

Appendix D: Additional figures and tables

Table A1. Raw output of the effect of probability numeracy and question format indicators on the probability of correct answers to the numeracy questions, model using basic demographic covariates in numeracy, ALP

	Q2	Q3	Q5	Q6	Q7
	10-WHITE-BALLS-B	7-WHITE-BALLS-A	7-WHITE-BALLS-C	INVERSE-PROB	SUBSET-EVENT
Probability numeracy	1.157***	1.077***	1.484***	1.419***	0.126***
	[0.043]	[0.052]	[0.056]	[0.072]	[0.031]
Battery was placed early	0.263***	0.197**	0.295***	0.447***	-0.014
	[0.046]	[0.084]	[0.046]	[0.093]	[0.056]
DK option was offered	-0.119**	-0.363***	-0.131***	-0.179*	-0.164***
	[0.048]	[0.062]	[0.049]	[0.093]	[0.056]
Saw the question earlier	0.055	0.021	0.075*	-	-
	[0.044]	[0.070]	[0.044]	-	-
Constant	1.135***	1.898***	0.893***	1.958***	-0.636***
	[0.110]	[0.120]	[0.136]	[0.156]	[0.049]
Observations	2,878	2,878	2,878	2,878	2,878

	Q9	Q10	Q11	Q12	Q13
	JOINT-PROB	MEAN-REVERT-A	AUTOCORR-COND	INDEPENDENCE	MEAN-REVERT-B
Probability numeracy	0.877***	0.728***	0.328***	0.737***	1.187***
	[0.045]	[0.033]	[0.032]	[0.044]	[0.077]
Battery was placed early	0.110	0.157***	-	-	-
	[0.075]	[0.057]	-	-	-
DK option was offered	0.086	-0.262***	-0.065	-0.439***	-0.444***
	[0.074]	[0.057]	[0.055]	[0.062]	[0.094]
Saw the question earlier	-0.048	0.227***	-	-	-
	[0.068]	[0.055]	-	-	-
Constant	-1.567***	0.648***	-0.313***	0.765***	2.080***
	[0.104]	[0.080]	[0.048]	[0.078]	[0.142]
Observations	2,878	2,878	2,878	2,878	2,878

Table A2. Pairwise correlations between the correctness of the numeracy answers, ALP wave 58, 60, and 61, unweighted

	W1										W2						W3						
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q2	Q3	Q5	Q11	Q12	Q13	Q1	Q2	Q4	Q5	Q6	Q9	Q10
Q1W1	1.00																						
Q2W1	0.93	1.00																					
Q3W1	0.45	0.44	1.00																				
Q4W1	0.48	0.46	0.46	1.00																			
Q5W1	0.49	0.47	0.46	0.97	1.00																		
Q6W1	0.41	0.40	0.42	0.41	0.42	1.00																	
Q7W1	0.04	0.04	0.03	0.02	0.03	0.03	1.00																
Q8W1	-0.03	-0.03	-0.03	-0.03	-0.04	-0.01	0.11	1.00															
Q9W1	0.13	0.14	0.09	0.16	0.16	0.11	0.13	0.15	1.00														
Q10W1	0.27	0.23	0.24	0.25	0.26	0.28	0.03	-0.03	0.14	1.00													
Q2W2	0.38	0.35	0.28	0.33	0.34	0.34	0.02	-0.03	0.10	0.22	1.00												
Q3W3	0.27	0.24	0.23	0.26	0.26	0.27	0.01	-0.04	0.07	0.17	0.45	1.00											
Q5W2	0.35	0.33	0.29	0.47	0.48	0.36	0.03	-0.05	0.15	0.24	0.48	0.46	1.00										
Q11W2	0.14	0.13	0.08	0.12	0.13	0.09	0.02	-0.01	0.13	0.06	0.14	0.12	0.14	1.00									
Q12W2	0.25	0.23	0.22	0.28	0.29	0.25	0.00	-0.02	0.11	0.16	0.33	0.27	0.33	0.11	1.00								
Q13W2	0.26	0.22	0.25	0.32	0.33	0.34	-0.02	-0.04	0.05	0.28	0.39	0.36	0.38	0.10	0.34	1.00							
Q1W3	0.44	0.42	0.35	0.35	0.35	0.33	0.02	-0.01	0.10	0.19	0.42	0.27	0.36	0.14	0.28	0.33	1.00						
Q2W3	0.43	0.42	0.34	0.34	0.34	0.33	0.04	-0.02	0.10	0.18	0.40	0.26	0.36	0.14	0.27	0.32	0.93	1.00					
Q4W3	0.34	0.32	0.30	0.50	0.50	0.33	0.08	-0.04	0.10	0.20	0.31	0.27	0.45	0.11	0.27	0.31	0.47	0.46	1.00				
Q5W3	0.35	0.34	0.31	0.49	0.49	0.34	0.08	-0.05	0.11	0.19	0.32	0.28	0.46	0.10	0.28	0.32	0.50	0.49	0.96	1.00			
Q6W3	0.34	0.32	0.33	0.36	0.36	0.43	0.02	-0.06	0.08	0.19	0.32	0.32	0.38	0.09	0.27	0.33	0.48	0.46	0.45	0.45	1.00		
Q9W3	0.16	0.15	0.09	0.19	0.19	0.12	0.11	0.10	0.58	0.16	0.15	0.10	0.20	0.16	0.15	0.11	0.15	0.16	0.18	0.17	0.11	1.00	
Q10W3	0.25	0.24	0.19	0.27	0.27	0.25	0.01	-0.03	0.12	0.42	0.31	0.19	0.31	0.11	0.19	0.36	0.34	0.32	0.29	0.29	0.29	0.17	1.00

* Questions are number as Q1, Q2, etc. The three waves are denoted by W1, W2 and W3.

Table A3. Simulated correlations between the numeracy answers based on a model that assumes conditional independence between the items*

	W1										W2						W3						
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q2	Q3	Q5	Q11	Q12	Q13	Q1	Q2	Q4	Q5	Q6	Q9	Q10
Q1W1	1.00																						
Q2W1	0.41	1.00																					
Q3W1	0.37	0.28	1.00																				
Q4W1	0.53	0.51	0.37	1.00																			
Q5W1	0.51	0.49	0.33	0.61	1.00																		
Q6W1	0.38	0.37	0.28	0.44	0.41	1.00																	
Q7W1	0.04	0.02	0.01	0.05	0.06	0.04	1.00																
Q8W1	0.00	-0.01	0.01	0.00	0.00	-0.01	-0.01	1.00															
Q9W1	0.14	0.12	0.09	0.17	0.16	0.10	0.04	0.03	1.00														
Q10W1	0.27	0.23	0.19	0.30	0.29	0.21	0.05	-0.01	0.10	1.00													
Q2W2	0.45	0.38	0.32	0.44	0.43	0.33	0.06	0.00	0.12	0.21	1.00												
Q3W3	0.34	0.28	0.22	0.35	0.35	0.27	0.01	0.02	0.08	0.15	0.30	1.00											
Q5W2	0.50	0.49	0.37	0.61	0.56	0.37	0.04	0.01	0.16	0.26	0.46	0.36	1.00										
Q11W2	0.11	0.11	0.07	0.14	0.13	0.07	0.05	-0.02	0.06	0.07	0.15	0.11	0.21	1.00									
Q12W2	0.25	0.21	0.15	0.30	0.28	0.19	0.01	0.02	0.11	0.12	0.26	0.23	0.33	0.15	1.00								
Q13W2	0.38	0.31	0.25	0.40	0.38	0.27	0.04	-0.02	0.11	0.14	0.35	0.28	0.39	0.14	0.25	1.00							
Q1W3	0.46	0.40	0.33	0.54	0.50	0.34	0.04	-0.04	0.12	0.21	0.43	0.30	0.52	0.14	0.26	0.36	1.00						
Q2W3	0.42	0.38	0.32	0.46	0.43	0.31	0.07	0.00	0.12	0.23	0.32	0.29	0.45	0.12	0.22	0.31	0.43	1.00					
Q4W3	0.51	0.46	0.38	0.62	0.60	0.40	0.02	0.02	0.15	0.29	0.49	0.38	0.59	0.16	0.28	0.39	0.54	0.48	1.00				
Q5W3	0.51	0.45	0.35	0.59	0.55	0.40	0.04	0.00	0.15	0.27	0.44	0.34	0.57	0.16	0.31	0.40	0.48	0.48	0.59	1.00			
Q6W3	0.35	0.31	0.23	0.40	0.38	0.27	0.03	0.00	0.09	0.12	0.33	0.28	0.38	0.09	0.16	0.28	0.36	0.34	0.42	0.40	1.00		
Q9W3	0.15	0.13	0.06	0.18	0.18	0.09	0.00	-0.02	0.06	0.11	0.17	0.15	0.23	0.24	0.17	0.13	0.21	0.21	0.24	0.22	0.15	1.00	
Q10W3	0.27	0.23	0.20	0.28	0.28	0.19	0.05	0.04	0.06	0.13	0.23	0.17	0.28	0.12	0.15	0.17	0.27	0.23	0.29	0.29	0.24	0.22	1.00

* We estimated a model using all items from all waves, based on the assumption that all answers are independent conditional on (latent) probability numeracy. The model is described in detail in Section 3.1. Then we estimated the expected value of probability numeracy for each individual in the sample given his/her answers to the numeracy questions. Finally, we simulated new correct/incorrect answers using the model's estimated parameters and individual's estimated numeracy. We expect a positive correlation between the simulated answers because they are all based on the same numeracy value: those with higher numeracy are more likely to provide correct answers to any questions.

Table A4. Differences between the empirical and the simulated correlations between the numeracy answers

	W1										W2						W3						
	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q2	Q3	Q5	Q11	Q12	Q13	Q1	Q2	Q4	Q5	Q6	Q9	Q10
Q1W1	0.00																						
Q2W1	0.52	0.00																					
Q3W1	0.09	0.15	0.00																				
Q4W1	-0.06	-0.04	0.08	0.00																			
Q5W1	-0.01	-0.01	0.13	0.36	0.00																		
Q6W1	0.03	0.03	0.14	-0.03	0.01	0.00																	
Q7W1	0.00	0.03	0.02	-0.03	-0.04	0.00	0.00																
Q8W1	-0.03	-0.02	-0.04	-0.03	-0.03	0.00	0.12	0.00															
Q9W1	-0.01	0.01	0.01	-0.01	0.00	0.01	0.10	0.11	0.00														
Q10W1	0.00	0.00	0.05	-0.05	-0.04	0.07	-0.02	-0.02	0.04	0.00													
Q2W2	-0.07	-0.03	-0.05	-0.11	-0.09	0.01	-0.04	-0.04	-0.02	0.01	0.00												
Q3W3	-0.07	-0.05	0.01	-0.09	-0.09	0.00	0.00	-0.06	-0.02	0.02	0.14	0.00											
Q5W2	-0.16	-0.17	-0.08	-0.14	-0.08	-0.02	-0.01	-0.06	-0.01	-0.02	0.03	0.10	0.00										
Q11W2	0.03	0.01	0.02	-0.02	0.00	0.03	-0.03	0.01	0.06	0.00	-0.01	0.00	-0.07	0.00									
Q12W2	0.00	0.02	0.07	-0.02	0.01	0.06	-0.01	-0.04	0.00	0.04	0.07	0.04	0.00	-0.04	0.00								
Q13W2	-0.12	-0.09	0.00	-0.08	-0.05	0.07	-0.06	-0.02	-0.06	0.15	0.04	0.08	-0.01	-0.03	0.08	0.00							
Q1W3	-0.02	0.02	0.02	-0.20	-0.15	-0.02	-0.01	0.03	-0.01	-0.02	-0.01	-0.03	-0.16	0.00	0.02	-0.03	0.00						
Q2W3	0.01	0.05	0.02	-0.12	-0.08	0.02	-0.03	-0.01	-0.02	-0.04	0.08	-0.03	-0.09	0.02	0.05	0.01	0.50	0.00					
Q4W3	-0.17	-0.14	-0.07	-0.12	-0.10	-0.07	0.06	-0.06	-0.04	-0.09	-0.17	-0.11	-0.14	-0.06	-0.01	-0.08	-0.07	-0.01	0.00				
Q5W3	-0.16	-0.12	-0.04	-0.10	-0.06	-0.06	0.04	-0.05	-0.04	-0.08	-0.12	-0.06	-0.11	-0.06	-0.03	-0.08	0.02	0.01	0.37	0.00			
Q6W3	-0.01	0.00	0.10	-0.04	-0.02	0.15	-0.01	-0.06	0.00	0.07	-0.02	0.04	0.00	0.00	0.11	0.05	0.12	0.12	0.03	0.06	0.00		
Q9W3	0.01	0.02	0.03	0.02	0.01	0.03	0.11	0.12	0.52	0.05	-0.03	-0.05	-0.03	-0.08	-0.02	-0.02	-0.05	-0.06	-0.06	-0.05	-0.04	0.00	
Q10W3	-0.01	0.00	-0.01	-0.01	-0.01	0.06	-0.03	-0.07	0.06	0.29	0.07	0.02	0.03	-0.01	0.04	0.19	0.07	0.09	-0.01	0.00	0.06	-0.05	0.00

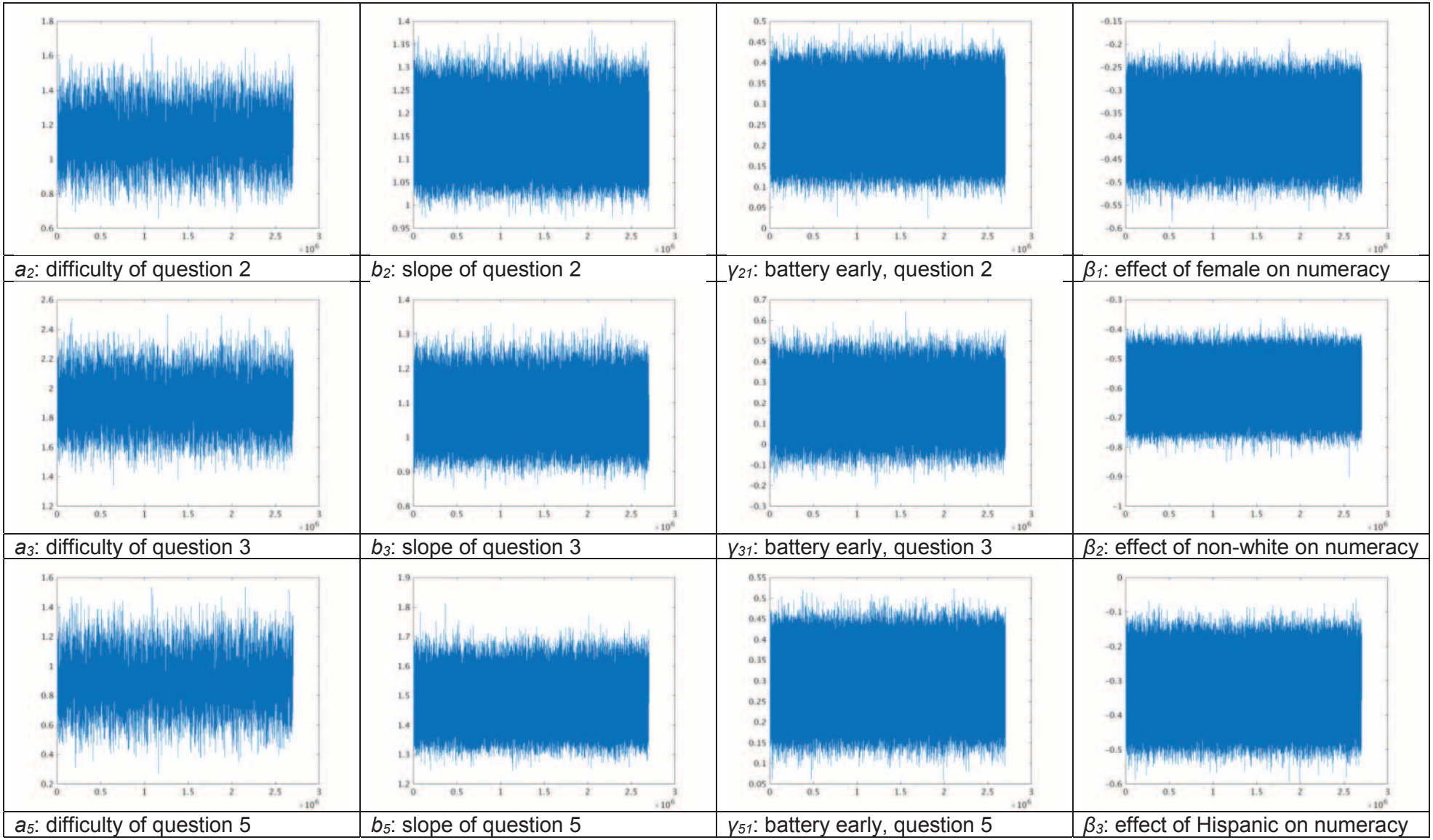
* The simulated correlations are based on a model that assumes independence between the items conditional on (latent) probability numeracy (the table notes under Table A3 provide details). Values further away from zero in this table may indicate violations of this independence assumption. Positive values indicate that the items are more related in reality than what the independence assumption implies; and negative values indicate that they are less related. Dark grey cells indicate that the difference between the empirical and simulated correlations is larger than 0.2 in absolute value, and these cells need thorough investigations. Light grey cells indicate that the difference is between 0.1 and 0.2 in absolute value.

Table A5. 18 versions of the probability numeracy score available on our project website

#	Items and used adjustments	Our recommendation	Other specifications
1	4 items, no demo, no adjust	Preferred score	no DK, early
2	4 items, with demo, no adjust	Preferred score	no DK, early
3	5 items, no demo, no adjust	Extended score	no DK, early
4	5 items, with demo, no adjust	Extended score	no DK, early
5	6 items, no demo, no adjust	Full score	no DK, early
6	6 items, with demo, no adjust	Full score	no DK, early
7	4 items, with demo, adjusted	Not recommended	no DK, early
8	5 items, with demo, adjusted	Not recommended	no DK, early
9	6 items, with demo, adjusted	Not recommended	no DK, early
10	4 items, no demo, no adjust	Not recommended	no DK, late
11	4 items, with demo, no adjust	Not recommended	no DK, late
12	5 items, no demo, no adjust	Not recommended	no DK, late
13	5 items, with demo, no adjust	Not recommended	no DK, late
14	6 items, no demo, no adjust	Not recommended	no DK, late
15	6 items, with demo, no adjust	Not recommended	no DK, late
16	4 items, with demo, adjusted	Not recommended	no DK, late
17	5 items, with demo, adjusted	Not recommended	no DK, late
18	6 items, with demo, adjusted	Not recommended	no DK, late

* “No demo” indicates that the score is only based on the probability numeracy items. “With demo” indicates that the score also makes use of basic demographic information on gender, age, race, education, marital status. “No adjust” indicates that latent numeracy is standardized to have a zero mean and a standard deviation of one on the ALP sample. “Adjusted” indicates that the score is standardized on the general U.S. population in March 2016. “no DK” indicates that “Don’t know” option was not offered in the survey (but individuals could skip the questions). “Early” (vs. “Late”) indicates that the battery was placed relatively early in the questionnaire.

Figure A1. Time series plots of the MCMC simulations, selected parameters



a_2: difficulty of question 2	b_2: slope of question 2	γ_{21}: battery early, question 2	β_1: effect of female on numeracy
a_3: difficulty of question 3	b_3: slope of question 3	γ_{31}: battery early, question 3	β_2: effect of non-white on numeracy
a_5: difficulty of question 5	b_5: slope of question 5	γ_{51}: battery early, question 5	β_3: effect of Hispanic on numeracy

Figure A2. Histograms of the MCMC simulation draws, selected parameters

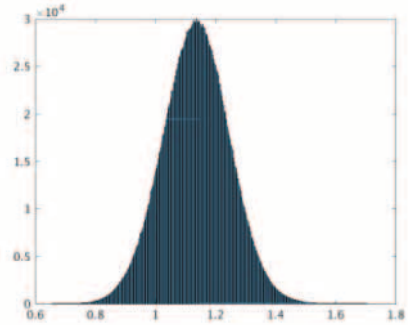
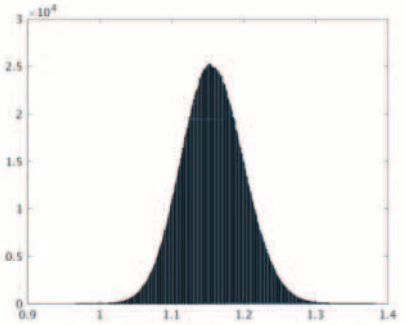
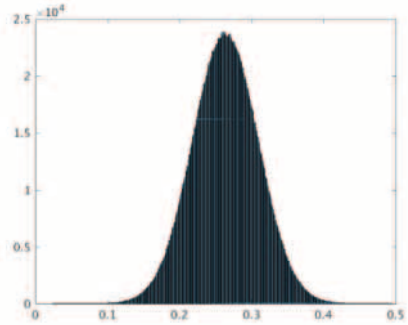
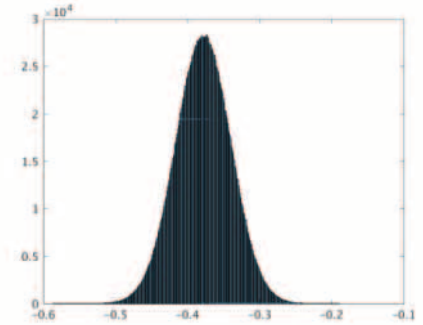
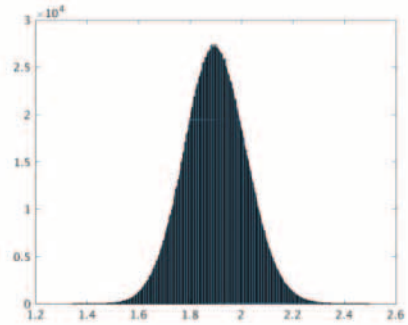
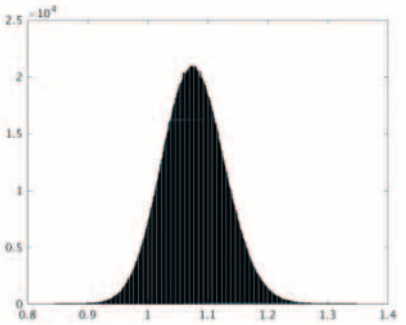
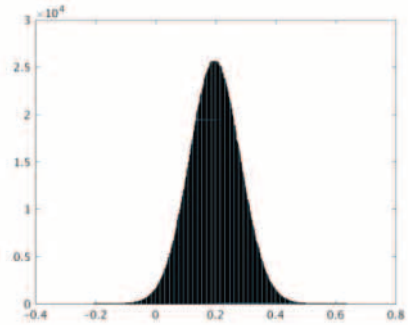
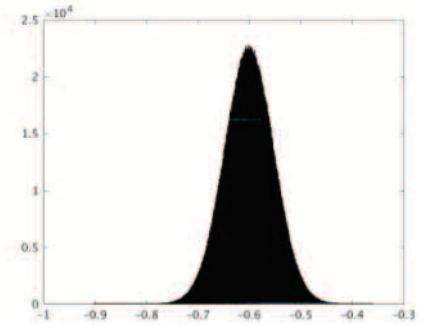
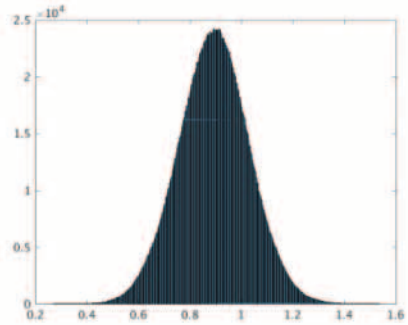
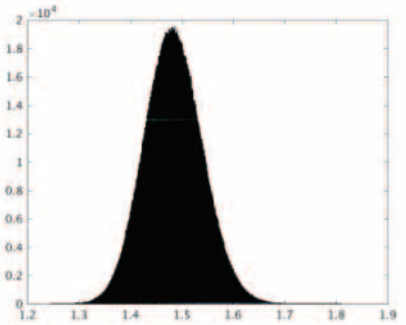
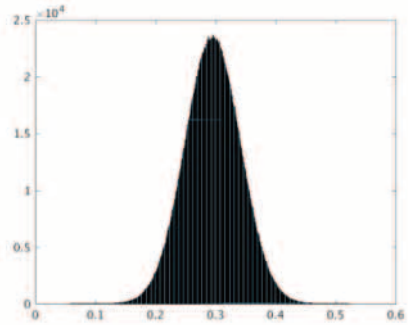
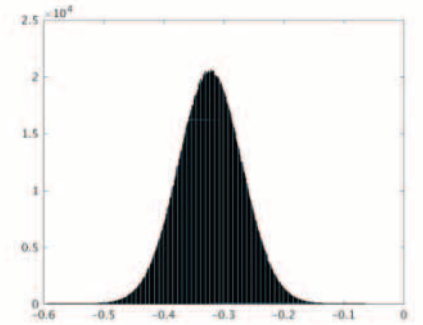
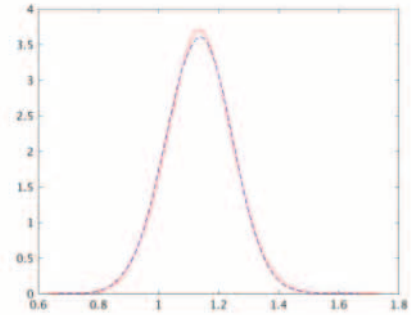
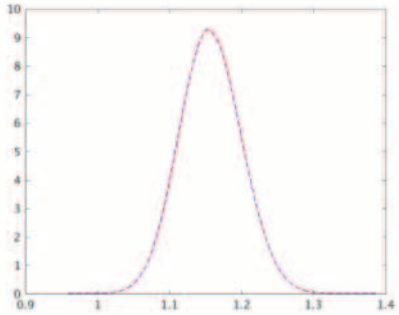
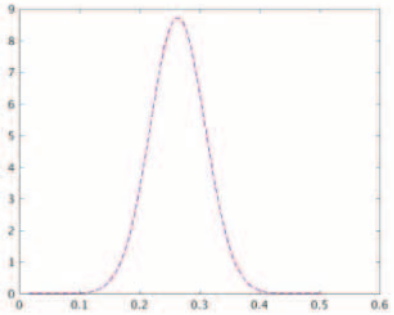
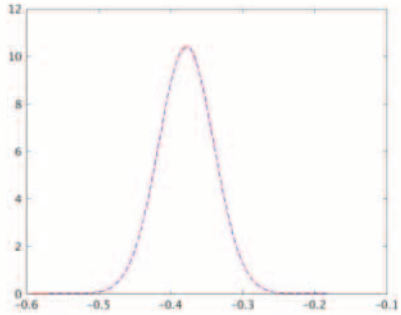
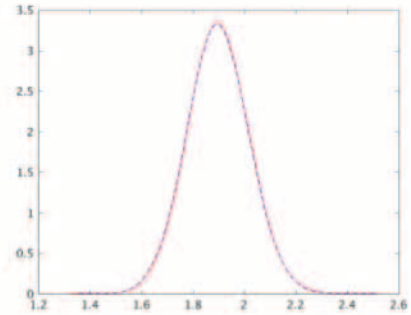
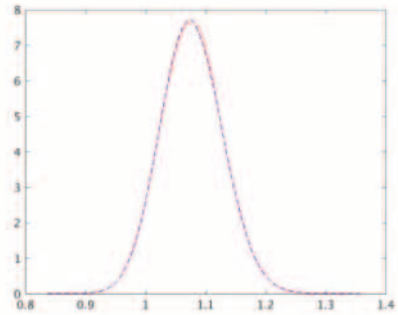
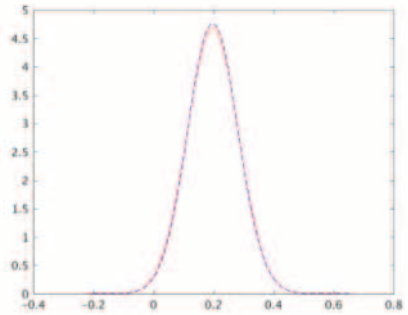
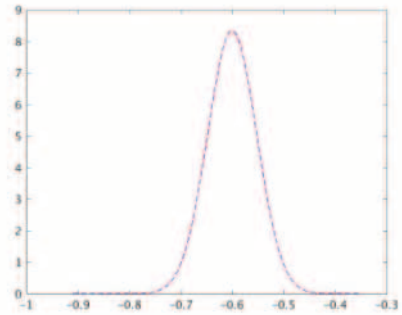
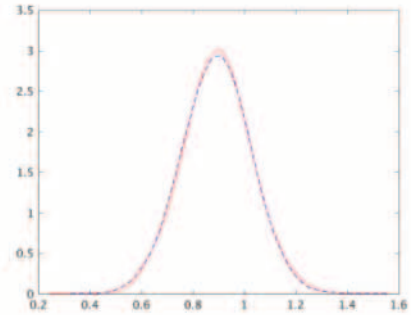
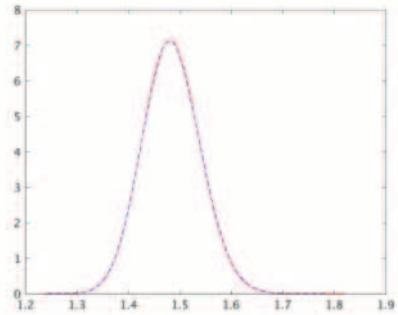
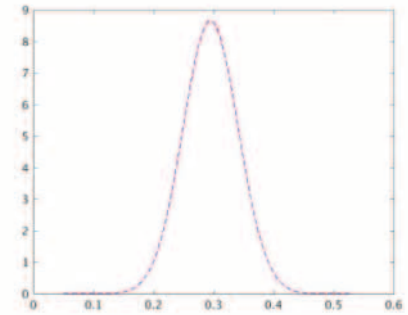
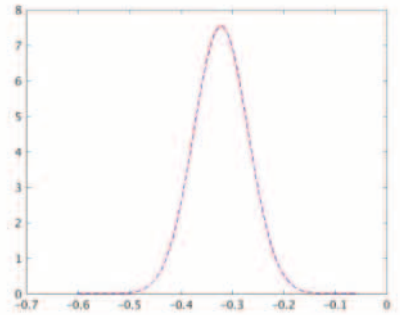
			
a_2: difficulty of question 2	b_2: slope of question 2	γ_{21}: battery early, question 2	β_1: effect of female on numeracy
			
a_3: difficulty of question 3	b_3: slope of question 3	γ_{31}: battery early, question 3	β_2: effect of non-white on numeracy
			
a_5: difficulty of question 5	b_5: slope of question 5	γ_{51}: battery early, question 5	β_3: effect of Hispanic on numeracy

Figure A3. Histograms of the MCMC simulation draws, first 50% (solid red) vs. second 50% (dashed blue), selected parameters

			
a_2: difficulty of question 2	b_2: slope of question 2	γ_{21}: battery early, question 2	β_1: effect of female on numeracy
			
a_3: difficulty of question 3	b_3: slope of question 3	γ_{31}: battery early, question 3	β_2: effect of non-white on numeracy
			
a_5: difficulty of question 5	b_5: slope of question 5	γ_{51}: battery early, question 5	β_3: effect of Hispanic on numeracy

References

Börsch-Supan, A., Elsner, D., Faßbender, H., Kiefer, R., McFadden, D., & Winter, J. (2004). How to Make Internet Surveys Representative: A Case Study of a Two-step Weighting Procedure. *Manuscript, Mannheim Germany: Mannheim University*.

Sriraman, B., & Chernoff, E. J. (Eds.). (2014). *Probabilistic Thinking: Presenting Plural Perspectives*. Springer.

Cowles, M. K., & Carlin, B. P. (1996). Markov Chain Monte Carlo Convergence diagnostics: A Comparative Review. *Journal of the American Statistical Association*, *91*(434), 883-904. doi:10.2307/2291683

Craig, B. M., Hays, R. D., Pickard, A. S., Cella, D., Revicki, D. A., & Reeve, B. B. (2013). Comparison of US Panel Vendors for Online Surveys. *Journal of Medical Internet Research*, *15*(11). http://doi.org/10.2196/jmir.2903

Fox, J. P. (2010). *Bayesian Item Response Modeling: Theory and Applications*. Springer Science & Business Media.

Gelman, A., Stern, H. S., Carlin, J. B., Dunson, D. B., Vehtari, A., & Rubin, D. B. (2013). *Bayesian Data Analysis*. Chapman and Hall/CRC.

Gilovich, T., Griffin, D., & Kahneman, D. (Eds.). (2002). *Heuristics and Biases: The Psychology of Intuitive Judgment*. Cambridge University Press.

Hudomiet P., Hurd, M., & S. Rohwedder (2018). Probability Numeracy: Measurement of Heterogeneity in the Quality of Subjective Probability Data, Unpublished working paper.

Hurd, M. (2009). Subjective Probabilities in Household Surveys. *Annual Review of Economics, 1*, 543-562. doi:10.1146/annurev.economics.050708.142955.

Hyslop, D. R., & Imbens, G. W. (2001). Bias from Classical and Other Forms of Measurement Error. *Journal of Business & Economic Statistics*, *19*(4), 475-481.

Tversky, A., & Kahneman, D. (1974). Judgment Under Uncertainty: Heuristics and Biases. *Science*, *185*(4157), 1124-1131.

Kahneman, D., & Tversky, A. (1972). Subjective Probability: A Judgment of Representativeness. In *The Concept of Probability in Psychological Experiments* (pp. 25-48). Springer, Dordrecht.

Kimball, M.S., Sahm, C. R. & Shapiro, M.D., (2008), Imputing Risk Tolerance from Survey Responses, *Journal of the American Statistical Association*, 103(483), 1028-1038.

van der Linden, W. J., & van der Linden, W. J. (2017). Introduction "Handbook of Item Response Theory, Volume One: Models." *Chapman & Hall/CRC Statistics in the Social and Behavioral Sciences*.

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General Performance on a Numeracy Scale Among Highly Educated Samples. *Medical Decision Making*, 21, 37–44.

Manski, C. (2004). Measuring Expectations. *Econometrica,* 72(5): 1329-1376.

Reyna, V. F., Nelson, W. L., Han, P. K., & Dieckmann, N. F. (2009). How Numeracy Influences Risk Comprehension and Medical Decision Making. *Psychological Bulletin*, 135(6), 943.

Schwartz, L. M., Woloshin, S., Black, W. C., & Welch, H. G. (1997). The Role of Numeracy in Understanding the Benefit of Screening Mammography. *Annals of Internal Medicine*, 127, 966–972.