Does Overconfidence Affect Financial Behaviors? Evidence from Retirement Readiness, Precautionary Savings, and Financial Market Participation

Minghao Yang

May 27, 2020

Abstract

This paper investigates whether overconfidence in financial literacy affects financial behaviors of households. To measure overconfidence of households, I utilize National Financial Capability Study (NFCS) data to construct several machine learning classifiers which learn from demographic characteristics, perceived financial literacy, and true financial literacy of households. Support Vector Machine (SVM) classifier and random forest classifier outperform the others in terms of cross-validation mean squared errors (MSE). Using the probability of overconfidence predicted by SVM and random forest, I find that overconfidence in financial literacy has a positive effect on retirement readiness, precautionary savings, and financial market participation.

1 Data

The data of this paper come from the National Financial Capability Study (NFCS). I use the 2018 tracking data which contain all the observations from the 2009, 2012, 2015, and 2018 studies. I drop the observations in the 2009 study because the education data cannot be obtained. After that, there are 80,164 observations left in the final dataset. The following sections describe the key variables of financial behaviors, demographic characteristics, perceived financial literacy, and true financial literacy.

1.1 Financial Behaviors

The NFCS covers a series of questions regarding financial behaviors of households. This paper makes uses of the questions related to retirement readiness, precautionary savings, and financial market participation. The questions are as follow:

- Have you ever tried (Before you retired, did you try) to figure out how much you need(ed) to save for retirement?

- Have you set aside emergency or rainy day funds that would cover your expenses for 3 months, in case of sickness, job loss, economic downturn, or other emergencies?

- Not including retirement accounts, do you have any investments in stocks, bonds, mutual funds, or other securities?

The questions are all “Yes or No” type, while ”Don’t know (DK)” and ”Prefer not to say” are allowed. For each question, I construct an indicator which equals one if the answer is “Yes”, so that the three indicators could reflect whether a household is ready for retirement, has precautionary savings, or participates in the financial market.

1.2 Demographic Characteristics

The NFCS also covers a rich set of demographic characteristics. For this paper I mainly use age, gender, race, income, education, marital status, and residential state data. For age and income, the NFCS only gives a range for each household, so I use the group mean as the imputed age or income. For education, I construct two dummies representing high school graduates and college graduates. For residential state, I create 51 dummies representing the 51 states included in the

NFCS. Table 1 provides the weighted summary statistics for financial behaviors and demographic characteristics.

Table 1: Summary statistics: Financial behaviors and demographic characteristics

Variables	10^{th} pct	Median	90^{th} pct	Mean	S.D.	#Obs.
Retirement Readiness	0	0	1	0.309	0.462	80164
Precautionary Savings	0	0	1	0.449	0.497	80164
Financial Market Participation	0	0	1	0.314	0.464	80164
Female	0	1	1	0.514	0.500	80164
Age	20	50	70	46.34	16.52	80164
Nonwhite	0	0	1	0.350	0.477	80164
Married	0	1	1	0.523	0.499	80164
Income	7500	42500	125000	62054.3	49231.7	80164
High School	1	1	1	0.954	0.210	80164
College	0	0	1	0.355	0.479	80164

Notes: The variables are extracted from the 2012, 2015, and 2018 NFCS. The sample weights in the NFCS are used to calculate the statistics.

1.3 Perceived Financial Literacy

The NFCS asks the following two questions to capture perceived financial literacy of each households:

- How would you assess your overall financial knowledge?
- How strongly do you agree or disagree with the following statements? - I am pretty good at math.

These two questions indicate self-assessed financial knowledge and math capability of each households, which are two important aspects of financial literacy. The answers are scaled from 1 to 7, where 1 means very low or strongly disagree and 7 means very high or strongly agree.

1.4 True Financial Literacy

Lusardi and Mitchell (2014) summaries the “Big Three” questions regarding interest rate, inflation and risk diversification. Other than the “Big Three” questions, the NFCS further asks two questions regarding mortgage payment and bond price. These five questions can reflect the true financial literacy of households and are referred to as the “Big Five” questions. The households are allowed to reply "Don't know (DK)" or "Prefer not to say (R)" when they answer the questions. The questions are summarized in Table 2, with the correct answers in bold.

Table 2: Questions regarding true financial literacy

Q1: Suppose you had $100 in a savings account and the interest rate was 2% per year. After 5 years, how much do you think you would have in the account if you left the money to grow?		
More than $102	Exactly $102	Less than $102
Q2: Imagine that the interest rate on your savings account was 1% per year and inflation was 2% per year. After 1 year, how much would you be able to buy with the money in this account?		
More than today	Exactly the same	**Less than today**
Q3: Buying a single company's stock usually provides a safer return than a stock mutual fund.		
True	**False**	
Q4: A 15-year mortgage typically requires higher monthly payments than a 30-year mortgage, but the total interest paid over the life of the loan will be less.		
True	False	
Q5: If interest rates rise, what will typically happen to bond prices?		
They will rise	**They will fall**	They will stay the same

Notes: The questions are from the 2012, 2015, and 2018 NFCS Questionnaires.

2 Methods

2.1 Constructing Overconfidence Measures

To train the machine learning classifiers, I first construct a learning set where the households can be unambiguously defined as overconfident or not. To be specific, the households who give incorrect answers to all the "Big Five" questions but choose six or seven in the two self-assessed questions are hard coded as overconfident. On the other hand, the households who give correct answers to all the "Big Five" questions and choose six or seven to the two self-assessed questions, as well as the households who give incorrect answers to all the " Big Five" questions and choose one or two in the two self-assessed questions are hard coded as not overconfident. The above coding rule yields 858 overconfident households and 7,506 not overconfident households. Then, I use demographic characteristics, answers to the two self-assessed questions, and answers to the "Big Five" questions as the inputs of the classifiers. After fitting all the classifiers with the optimal parameters from randomized search cross validation, the mean squared errors (MSE) are calculated for model selection. In the end, I get the out-of-sample predictions of the remaining 71,800 observations for the best performed classifiers as the overconfidence measures used in the main analyses. Rather than use the predicted class, I use the predicted probability as the overconfidence measure so that it is continuous on range $[0, 1]$. I train six classifiers, namely logistic regression classifier, random forest classifier, Support Vector Machine (SVM) classifier, Naive Bayes (NB) classifier, K-nearest Neighbors (KNN) classifier, and Multi-layer Perception (MLP) classifier. The following sections describe how I train the Random Forest and SVM classifier in detail because they yield the smallest MSEs (see Results section). The description of other classifiers are shown in Appendix A.1.

2.1.1 Random Forest Classifier

The building block of any random forest classifier is the decision tree classifier. It divides the feature space into multiple sub-spaces and uses the mean value of each sub-space as the prediction for that space. Mathematically, let R denotes the feature space. The decision tree classifier divides it into J sub-spaces. Then for household i whose features are in space R_j, the probability of overconfidence is given by

$$\Pr(y_i = 1|\mathbf{X_i} \in R_j) = \frac{1}{\#R_j} \sum_{y_j \in R_j} y_j \tag{1}$$

The random forest classifier is an ensemble method that combines multiple decision trees by bootstrap aggregation and feature randomness. Therefore, the prediction would become more robust than any single tree. For this paper each decision tree takes random features from feature space and make a prediction on whether a household is overconfident. After that, the random forest classifiers select the prediction that most trees give as the final output. The number of trees, the maximum depth of each tree, the minimum number of samples required to split an internal node, the minimum number of samples required to be at a leaf node, and the number of features to consider when looking for the best split are tuned in randomized search cross validation.

2.1.2 Support Vector Machine Classifier

The binary Support Vector Machine (SVM) classifier divides the observations into two classes by constructing a surface in the feature space such that the margin (the distance between the surface and the closest data point) is maximized. Mathematically, the SVM classifier solves the following problem:

$$\begin{aligned} \max_{\beta_1,\beta_2,...\beta_P,M} \quad & M \\ \text{s.t} \quad & \sum_{j=1}^{P} \beta_j^2 = 1 \\ & y_i \left(\beta_0 + \beta_1 X_{1,i} + \beta_2 X_{2,i} + ... + \beta_P X_{P,i}\right) \geq M \quad \forall i = 1, ...N \end{aligned} \tag{2}$$

where M is the maximized margin, y_i is the overconfidence indicator, and $X_{p,i}$ is the pth feature of household i. Empirically, I tune the regulation parameter C in randomized search cross validation. I utilize the RBF kernel to run the model because of its high accuracy and speed.

2.2 Constructing True Financial Literacy Measure

The measure is constructed following Lusardi and Mitchell (2017). In order to combine the answers of the "Big Five" questions into a single measure, I generate an indicator for each question which equals one if the household answers the question correctly. After that, I perform a factor analysis on these five indicators using the principal component factor method. The factor score is then calculated and normalized to range $[0, 1]$ so that the scale is the same as the overconfidence measure. The normalized factor score is used as the measure for true financial literacy. The factor loads and uniqueness as well as the summary statistics are provided in Appendix A.3.

2.3 Investigating the Effect of Overconfidence

To see whether overconfidence in financial literacy influence financial behaviors of households, especially those with similar true financial literacy, I run the following logit regression:

$$\begin{aligned}&\Pr(y_i = 1|\mathrm{Overconfidence_i}, \mathrm{True\_Literacy_i}, \mathbf{X}_\mathrm{i}^\mathrm{D}, \mathbf{X}_\mathrm{i}^\mathrm{YS}, \varepsilon_i, \beta_0, \beta_1, \beta_2, \beta_3, \beta_4)\\ &= F(\beta_0 + \beta_1 \mathrm{Overconfidence_i} + \beta_2 \mathrm{True\_Literacy_i} + \mathbf{X}_\mathrm{i}^\mathrm{D}\beta_3 + \mathbf{X}_\mathrm{i}^\mathrm{YS}\beta_4 + \varepsilon_i)\end{aligned} \tag{3}$$

where $F(x) = e^x/(1 + e^x)$. y_i represents the indicators for retirement readiness, precautionary savings, and financial market participation; $\mathrm{Overconfidence_i}$ is the overconfidence measure generated by machine learning classifiers; $\mathrm{True\_Literacy_i}$ is the true financial literacy measure generated by factor analysis; $\mathbf{X}_\mathrm{i}^\mathrm{D}$ is a matrix of demographic characteristics including age, age squared, log income, log income squared, gender, race, marital status, and education; $\mathbf{X}_\mathrm{i}^\mathrm{YS}$ is a matrix of state and year dummies.

3 Results

3.1 Overconfidence Measures from Different Classifiers

The MSEs of different classifiers are presented in Figure 1. It is shown that the SVM and random forest classifiers have much smaller MSE, so their out-of-sample predictions for the remaining 71,800 observations are used in the main analyses. Table 3 presents the weighted summary statistics for the two overconfidence measures over the full sample. The summary statistics for other measures could be found in Appendix A.2. It is shown that the measure from SVM has both a larger mean and a larger variance.

Figure 1: MSEs of different classifiers

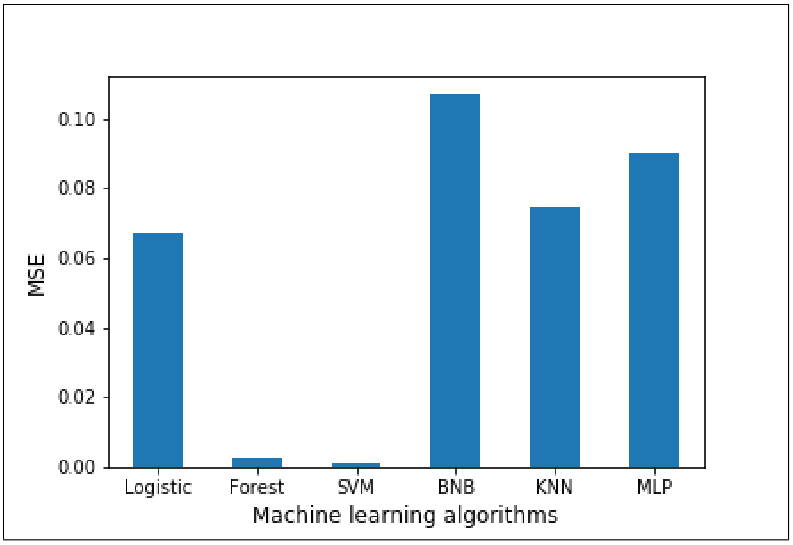

Table 3: Summary statistics: Overconfidence measures from SVM and forest

Classifiers	10^{th} pct	Median	90^{th} pct	Mean	S.D.	#Obs.
SVM	4.37e-05	0.133	1.000	0.392	0.426	80164
Forest	0.029	0.203	0.455	0.234	0.170	80164

Notes: The overconfidence measures are predicted by SVM and random forest classifiers as described in the Methods section. The sample weights in the NFCS are used to calculate the statistics.

3.2 The Effects of Overconfidence on Financial Behaviors

In this section I only use the overconfidence measures given by SVM and random forest to run regression (3). The regression results with other overconfidence measures given by different classifiers are provided in Appendix A.4.

Table 4 shows the results from regression (3) using the SVM based overconfidence measure, where the coefficients are the *average marginal effects* rather than the *log odds ratios*. Demographic characteristics and year dummies are controlled in all columns, while column (2), (4), and (6) further control state dummies. In column (1) and (2), the dependent variable is the indicator for retirement readiness; in column (3) and (4), the dependent variable is the indicator for precautionary savings; in column (5) and (6), the dependent variable is the indicator for financial market participation. The average marginal effects of overconfidence on retirement readiness, precautionary savings, and financial market participation are all positive, which suggests that overconfidence does affect financial behaviors of households. Concretely, a standard deviation increase (0.43 in Table 3) in overconfidence will increase the probability of retirement readiness, precautionary savings, and financial market participation by 6 - 6.5 percent, with all others being equal. The increases in probabilities are quite decent given that only 30.9% households prepare for retirement, only 44.9% households have precautionary savings, and only 31.4% households participate in financial market. (The numbers are from Table 1.)

Table 5 shows the results from regression (3) using the random forest based overconfidence measure. The setting of Table 5 is exactly the same as Table 4. The average marginal effects of overconfidence are still positive, but with larger scales. This is because the standard deviation of the random forest based overconfidence measure is much smaller. Again from standard deviation perspective, a standard deviation increase (0.17 in Table 3) in overconfidence will increase the probability of retirement readiness, precautionary savings, and financial market participation by 7.8 - 8.1 percent, with all others being equal. Hence, the effects of overconfidence on financial behaviors become even stronger.

Table 4: Logit regression on overconfidence (SVM based) and true financial literacy

Dependent Variables	(1) Readiness	(2) Readiness	(3) Precaution	(4) Precaution	(5) Participation	(6) Participation
Overconfidence	0.141***	0.142***	0.151***	0.152***	0.139***	0.152***
	(0.00534)	(0.00534)	(0.00541)	(0.00542)	(0.00547)	(0.00542)
True Literacy	0.348***	0.344***	0.311***	0.313***	0.376***	0.313***
	(0.00833)	(0.00834)	(0.00835)	(0.00835)	(0.00854)	(0.00835)
#Obs.	80,164	80,164	80,164	80,164	80,164	80,164
Demo. chars.	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
State dummies	No	Yes	No	Yes	No	Yes
Pseudo R-squared	0.142	0.143	0.160	0.161	0.191	0.161

Notes: The results are from regression (3). Overconfidence measure is predicted by the SVM classifier. True financial literacy is calculated via factor analysis. Demographic characteristics and year dummies are controlled in all columns. Column (2), (4), and (6) further control state dummies. In column (1) and (2), the dependent variable is the indicator for retirement readiness; in column (3) and (4), the dependent variable is the indicator for precautionary savings; in column (5) and (6), the dependent variable is the indicator for financial market participation. Observations are weighted by the NFCS sample weights. Standard errors are in the parentheses. The symbols *, **, and *** denote significance at the 10%, 5% and 1% levels respectively.

Table 5: Logit regression on overconfidence (random forest based) and true financial literacy

Dependent Variables	(1) Readiness	(2) Readiness	(3) Precaution	(4) Precaution	(5) Participation	(6) Participation
Overconfidence	0.473***	0.477***	0.463***	0.459***	0.481***	0.459***
	(0.0209)	(0.0210)	(0.0219)	(0.0219)	(0.0212)	(0.0219)
True Literacy	0.443***	0.441***	0.390***	0.389***	0.475***	0.389***
	(0.0126)	(0.0126)	(0.0128)	(0.0129)	(0.0128)	(0.0129)
#Obs.	80,164	80,164	80,164	80,164	80,164	80,164
Demo. chars.	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
State dummies	No	Yes	No	Yes	No	Yes
Pseudo R-squared	0.140	0.141	0.157	0.158	0.190	0.158

Notes: The results are from regression (3). Overconfidence is predicted by the random forest classifier. True financial literacy is calculated via factor analysis. Demographic characteristics and year dummies are controlled in all columns. Column (2), (4), and (6) further control state dummies. In column (1) and (2), the dependent variable is the indicator for retirement readiness; in column (3) and (4), the dependent variable is the indicator for precautionary savings; in column (5) and (6), the dependent variable is the indicator for financial market participation. Observations are weighted by the NFCS sample weights. Standard errors are in the parentheses. The symbols *, **, and *** denote significance at the 10%, 5% and 1% levels respectively.

References

Lusardi, Annamaria and Olivia S Mitchell, "The economic importance of financial literacy: Theory and evidence," *Journal of economic literature*, 2014, *52* (1), 5–44.

_ **and** _ , "How ordinary consumers make complex economic decisions: Financial literacy and retirement readiness," *Quarterly Journal of Finance*, 2017, *7* (03), 1750008.

A Appendix

A.1 Constructing Overconfidence Measures (Con't)

This section describes the other machine learning classifiers that I used to construct the overconfidence measure. However, because of the large MSEs, the out-of-sample prediction are not used in the main analyses.

A.1.1 Logistic Regression Classifier

The logistic classifier fits the linear regression in a sigmoid function such that the probability will not exceed the range $[0, 1]$. Formally,

$$\Pr(y_i = 1|\mathbf{X_i}, \varepsilon_i, \beta_0, \beta_1) = F(\beta_0 + \mathbf{X_i}\beta_1 + \varepsilon_i) \tag{4}$$

where $F(x) = e^x/(1+e^x)$, y_i is an indicator of overconfidence, and $\mathbf{X_i}$ represents the feature matrix. In this paper, the inverse of regulation strength C is tuned in randomized search cross validation.

A.1.2 Naive Bayes Classifier

The Naive Bayes (NB) classifier is an application of Bayes rule. Given overconfidence indicator $y \in 0, 1$, features X_1 through X_P, and the naive conditional independence assumption, Bayes rule gives:

$$\Pr(y|X_1, \cdots, X_P) = \frac{\Pr(y)\prod_{p=1}^{P}\Pr(X_p|y)}{\Pr(X_1, \cdots, X_P)} \tag{5}$$

Since $\Pr(X_1, \cdots, X_P)$ is constant, we can use Maximum A Posteriori (MAP) estimation to estimate the probability of household i to be overconfident, which is

$$y_i = \underset{y\in\{0,1\}}{\arg\max} \Pr(y) \prod_{p=1}^{P} \Pr(X_{p,i}|y) \tag{6}$$

Since most features in this paper is binary or categorical, I use a Bernoulli kernel to estimate the model. The additive smoothing parameter α is tuned in randomized search cross validation.

A.1.3 K-nearest Neighbors Classifier

The K-nearest Neighbors (KNN) classifier is a non-parametric model, which simply uses the data in the neighborhood of each data point to predict type. Concretely, the probability of household i to be overconfident is given by

$$\Pr(y_i = 1|\mathbf{X_i}) = \frac{1}{K}\sum_{k\in\mathcal{N}_0} I\left(y_k = 1\right) \tag{7}$$

where $\mathcal{N}_0$ is the set of K nearest neighbors. In this paper I use Euclidean distance to find the nearest neighbors. The number of neighbors K is tuned in randomized search cross validation. To avoid overfitting, I set K to be larger than 50.

A.1.4 Multi-layer Perceptron Classifier

The Multi-layer Perceptron (MLP) classifier is a neural network model with multiple hidden layers and nodes. It ensembles nonlinear functions of linear functions of features. To get the nodes at layer j, the MLP classifier estimates the following function:

$$Z_{m,j} = f_j(\alpha_j + \sum_{k \in \mathcal{N}_{j-1}} \beta_{k,j} Z_{k,j-1}) \tag{8}$$

where $Z_{m,j}$ denotes the mth node at layer j, $Z_{k,j-1}$ denotes the kth node at layer $j-1$, $\mathcal{N}_{j-1}$ denotes the set of nodes at layer $j-1$, and $f_j(\cdot)$ denotes the nonlinear activation function. In this paper I use rectified linear unit (reLU) function to estimate the model. The hidden layer sizes and the L2 penalty parameter α are tuned in randomized search cross validation. Given that the learning set is quite large, I set the initial learning rate at 0.02 so that the classifier does not always give the same prediction for every observation.

A.2 Overconfidence Measures from Other Classifiers

Table A1 shows the weighted summary statistics of overconfidence measures from other machine learning classifiers that are not presented in Table 3.

Table A1: Summary statistics: Overconfidence measures from other classifiers

Classifiers	10^{th} pct	Median	90^{th} pct	Mean	S.D.	#Obs.
Logistic	0.002	0.113	0.789	0.264	0.301	80164
Bernoulli NB	0.010	0.187	0.591	0.247	0.231	80164
KNN	0	0.155	0.464	0.203	0.200	80164
MLP	0.011	0.169	0.353	0.189	0.154	80164

Notes: The overconfidence measures are predicted by logistic regression, Bernoulli NB, KNN, and MLP classifiers as described in Appendix A.1. The sample weights in the NFCS are used to calculate the statistics.

A.3 True Financial Literacy Measure from Factor Analysis

Panel A of Table A2 shows the factor loads and uniqueness of the correct indicator for each "Big Five" question. Panel B presents the weighted summary statistics of the constructed measure.

A.4 The Effects of Overconfidence on Financial Behaviors (Con't)

Table A3 - A6 show the results from regression (3) using different overconfidence measures. Table A3 uses the logistic regression based overconfidence measure; Table A4 uses the Bernoulli NB based overconfidence measure; Table A5 uses the KNN based overconfidence measure; Table A6 uses the MLP based overconfidence measure. The setting of the tables are the same as Table 4. No matter which measure I use, the average marginal effects of overconfidence are always positive and significant, except for that in Table A4 column (5). This is because the NB classifiers always have

Table A2: Measure for true financial literacy: Factor loads and summary statistics

Panel A: Factor loads and uniqueness for the "Big Five" questions						
Question	Interest Rate	Inflation	Risk Diversification	Mortgage Payment	Bond Price	
Loads	0.6435	0.7315	0.4972	0.6508	0.6824	
Uniqueness	0.5859	0.4649	0.7528	0.5765	0.5344	
Panel B: Summary statistics of the true financial literacy measure						
	10^{th} pct	Median	90^{th} pct	Mean	S.D.	#Obs.
True Literacy	.214	.630	1	0.580	0.299	80164

Notes: I perform a factor analysis on the correct indicators of the "Big Five" questions. Panel A displays the factor loads and uniqueness. The measure for true financial literacy is constructed as the normalized factor score. Panel B presents the summary statistics using the sample weights from the NFCS.

a pretty good in-sample fit, but the out-of-sample predictions of them cannot be seriously treated given its easy setup. Overall, the relationship found in the main analyses is robust.

Table A3: Logit regression on overconfidence (logistic regression based) and true financial literacy

Dependent Variables	(1) Readiness	(2) Readiness	(3) Precaution	(4) Precaution	(5) Participation	(6) Participation
Overconfidence	0.294***	0.292***	0.412***	0.413***	0.310***	0.413***
	(0.0113)	(0.0113)	(0.0121)	(0.0121)	(0.0115)	(0.0121)
True Literacy	0.254***	0.250***	0.224***	0.225***	0.278***	0.225***
	(0.00636)	(0.00639)	(0.00638)	(0.00639)	(0.00620)	(0.00639)
#Obs.	80,164	80,164	80,164	80,164	80,164	80,164
Demo. chars.	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
State dummies	No	Yes	No	Yes	No	Yes
Pseudo R-squared	0.142	0.143	0.163	0.164	0.192	0.164

Notes: The results are from regression (3). Overconfidence measure is predicted by the logistic regression classifier. True financial literacy is calculated via factor analysis. Demographic characteristics and year dummies are controlled in all columns. Column (2), (4), and (6) further control state dummies. In column (1) and (2), the dependent variable is the indicator for retirement readiness; in column (3) and (4), the dependent variable is the indicator for precautionary savings; in column (5) and (6), the dependent variable is the indicator for financial market participation. Observations are weighted by the NFCS sample weights. Standard errors are in the parentheses. The symbols *, **, and *** denote significance at the 10%, 5% and 1% levels respectively.

Table A4: Logit regression on overconfidence (Bernoulli NB based) and true financial literacy

Dependent Variables	(1) Readiness	(2) Readiness	(3) Precaution	(4) Precaution	(5) Participation	(6) Participation
Overconfidence	0.0588***	0.0828***	0.0572***	0.0456***	-0.0404***	0.0456***
	(0.0124)	(0.0137)	(0.0127)	(0.0143)	(0.0130)	(0.0143)
True Literacy	0.202***	0.200***	0.157***	0.157***	0.214***	0.157***
	(0.00611)	(0.00613)	(0.00623)	(0.00625)	(0.00600)	(0.00625)
#Obs.	80,164	80,164	80,164	80,164	80,164	80,164
Demo. chars.	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
State dummies	No	Yes	No	Yes	No	Yes
Pseudo R-squared	0.135	0.137	0.153	0.154	0.185	0.154

Notes: The results are from regression (3). Overconfidence measure is predicted by the Bernoulli NB classifier. True financial literacy is calculated via factor analysis. Demographic characteristics and year dummies are controlled in all columns. Column (2), (4), and (6) further control state dummies. In column (1) and (2), the dependent variable is the indicator for retirement readiness; in column (3) and (4), the dependent variable is the indicator for precautionary savings; in column (5) and (6), the dependent variable is the indicator for financial market participation. Observations are weighted by the NFCS sample weights. Standard errors are in the parentheses. The symbols *, **, and *** denote significance at the 10%, 5% and 1% levels respectively.

Table A5: Logit regression on overconfidence (KNN based) and true financial literacy

Dependent Variables	(1) Readiness	(2) Readiness	(3) Precaution	(4) Precaution	(5) Participation	(6) Participation
Overconfidence	0.172***	0.171***	0.190***	0.189***	0.210***	0.189***
	(0.0134)	(0.0134)	(0.0140)	(0.0140)	(0.0138)	(0.0140)
True Literacy	0.242***	0.238***	0.201***	0.202***	0.276***	0.202***
	(0.00700)	(0.00702)	(0.00708)	(0.00709)	(0.00692)	(0.00709)
#Obs.	80,164	80,164	80,164	80,164	80,164	80,164
Demo. chars.	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
State dummies	No	Yes	No	Yes	No	Yes
Pseudo R-squared	0.137	0.138	0.154	0.156	0.187	0.156

Notes: The results are from regression (3). Overconfidence measure is predicted by the KNN classifier. True financial literacy is calculated via factor analysis. Demographic characteristics and year dummies are controlled in all columns. Column (2), (4), and (6) further control state dummies. In column (1) and (2), the dependent variable is the indicator for retirement readiness; in column (3) and (4), the dependent variable is the indicator for precautionary savings; in column (5) and (6), the dependent variable is the indicator for financial market participation. Observations are weighted by the NFCS sample weights. Standard errors are in the parentheses. The symbols *, **, and *** denote significance at the 10%, 5% and 1% levels respectively.

Table A6: Logit regression on overconfidence (MLP based) and true financial literacy

Dependent Variables	(1) Readiness	(2) Readiness	(3) Precaution	(4) Precaution	(5) Participation	(6) Participation
Overconfidence	0.251***	0.250***	0.186***	0.185***	0.112***	0.185***
	(0.0166)	(0.0166)	(0.0174)	(0.0175)	(0.0172)	(0.0175)
True Literacy	0.242***	0.238***	0.186***	0.187***	0.240***	0.187***
	(0.00671)	(0.00673)	(0.00690)	(0.00691)	(0.00667)	(0.00691)
#Obs.	80,164	80,164	80,164	80,164	80,164	80,164
Demo. chars.	Yes	Yes	Yes	Yes	Yes	Yes
Year dummies	Yes	Yes	Yes	Yes	Yes	Yes
State dummies	No	Yes	No	Yes	No	Yes
Pseudo R-squared	0.137	0.139	0.154	0.155	0.185	0.155

Notes: The results are from regression (3). Overconfidence measure is predicted by the MLP classifier. True financial literacy is calculated via factor analysis. Demographic characteristics and year dummies are controlled in all columns. Column (2), (4), and (6) further control state dummies. In column (1) and (2), the dependent variable is the indicator for retirement readiness; in column (3) and (4), the dependent variable is the indicator for precautionary savings; in column (5) and (6), the dependent variable is the indicator for financial market participation. Observations are weighted by the NFCS sample weights. Standard errors are in the parentheses. The symbols *, **, and *** denote significance at the 10%, 5% and 1% levels respectively.