

Behavioral biases are temporally stable

Victor Stango and Jonathan Zinman*

December 2024

Abstract

Social scientists often consider temporal stability when assessing a construct and its measures, but whether behavioral biases display such stability is relatively unknown. We estimate stability for 17 biases, using lab-style elicitations in online surveys repeated three years apart. Biases are stable in the aggregate and at least moderately stable within-person. Within-person stability estimates tend to meet or exceed those for standard quantitative measures of patience and risk aversion, despite various pieces of evidence suggesting that classical measurement error substantially attenuates them. We consider implications for micro and macro modeling with potentially behavioral agents, and for survey and elicitation design.

Keywords: test-retest reliability; behavioral economics; multiple biases; construct validity; elicitation design

JEL: C36, C81, D90, E70

* Stango: UC Davis Graduate School of Management, vstango@ucdavis.edu; Zinman: Dartmouth College, IPA, J-PAL, and NBER, jzinman@dartmouth.edu. Thanks to Hannah Trachtman and Sucitro Dwijayana Sidharta for outstanding research assistance; the Sloan/Sage Working Group on Behavioral Economics and Consumer Finance, the Roybal Center (grant # 3P30AG024962), the National University of Singapore, the Michigan Retirement Research Center, and the Pension Research Council/TIAA for funding and patience; Shachar Kariv and Dan Silverman for helping us implement their (with Choi and Muller) interface for measuring choice consistency; Charlie Sprenger for help with choosing the certainty premium elicitation tasks and with adapting the convex time budget tasks; Georg Weizsacker for help in adapting one of the questions we use to measure narrow bracketing; Julian Jamison for advice on measuring ambiguity aversion; Doug Staiger, Josh Schwartzstein, and Dmitry Taubinsky for many conversations; John List for advice on addressing external validity; and Erik Snowberg for helpful discussions regarding measurement error. Special thanks to Joanne Yoong for collaboration on the Round 1 survey design and implementation.

Temporal stability is a key criterion for organizing the study of individual characteristics like cognitive skills, personality, and preferences.¹ Throughout the social sciences, stability is often considered a necessary condition for an individual characteristic to be a meaningful construct, as Stigler and Becker (1977) famously argue in the case of preferences. A lack of stability makes a characteristic more difficult to model, measure and link to behavior.

Relatively little is known about the temporal stability of behavioral biases, a.k.a. behavioral regularities or anomalies, perhaps because they are relatively recent additions to the taxonomy of individual characteristics and differences. Understanding their stability at the population level—for example, the degree of stability in the share of consumers with various biases—is important for macroeconomic modeling, which increasingly focuses on how persistent aspects of consumer decision making can help explain key aggregates and reshape optimal approaches to stabilization. Understanding their temporal stability at the person level has implications for whether and how to allow for biases when modeling consumer choice. Understanding how measurement error affects estimates of temporal stability in data from stylized tasks and survey questions has implications for almost any project that might use a behavioral bias measure for description, targeting or theory-testing.

We develop new empirical evidence on the magnitude and measurement of temporal stability for 17 biases that feature prominently in behavioral economics and are amenable to compact online elicitation. We define and measure biases as departures from clear classical benchmarks (e.g., time-consistent discounting, dominance avoidance, mean-zero expectation error). The data come from two repeated elicitations of each bias, administered in 2014 and 2017 to the same nationally representative sample of 845 U.S. adults.

Our first key finding is that sample-wide average bias prevalence, measured discretely for each bias as biased or unbiased relative to its classical benchmark, is remarkably stable across our three year gap. Across our many biases, the average absolute value change in the share of biased individuals is two percentage points, with the biggest change being nine percentage points (on a base of 83 percent for that bias). The sample share of biased observations remains roughly constant across the three years, starting and ending at 42 percent. These results support macro models that

¹ See, for example, Mata et al.'s (2018) review of work from psychology on risk attitudes.

are microfounded on deep behavioral tendencies, and/or assume somewhat fixed aggregate shares of behavioral and non-behavioral agents over time (e.g., Gabaix 2020; Lian forthcoming; Laibson, Maxted, and Moll forthcoming; Pfauti, Seyrich, and Zinman 2024; Vines and Wills 2018).

Our second key finding is that biases are moderately stable within-person. The sample share of panelists with an unchanged bias indicator across rounds ranges from [0.57 to 0.99] across our many biases, with a mean of 0.72. This suggests that micro theories with persistent biases are on reasonably firm empirical ground. Next, incorporating information on the intensive margins of deviations from classical benchmarks, we find rank stability of biases similar to that for standard measures of patience and risk aversion.² Across our 17 biases, round-to-round rank correlations range from 0.01 to 0.60 (with standard errors mostly in the 0.02 to 0.04 range), averaging 0.31. Estimated correlations for the most comparable standard measures of patience and risk aversion are around 0.30 and 0.40, both in our data and in prior work (Chuang and Schechter 2015; Mata et al. 2018).

Biases thus have estimated stability on par with parameters that are staples of modeling consumer choice over time and under risk. But both sets of variables have temporal stability well below 1. Their estimated stability is also well below that of summary measures of cognitive skills and personality in our data and elsewhere (e.g., Jensen 1998; Anusic and Schimmack 2016), prompting the question of whether this is due to true differential stability across characteristics or to differential measurement error.

Measurement error could, in principle, statistically bias stability estimates in either direction. One concern is attenuation from noise. For example, if respondents speed through an elicitation and pick some answers at random, or “guess” when faced with cognitively taxing questions, their answers will spuriously vary over time even if the object of measurement is truly stable. Another concern is that respondents may exhibit spurious (i.e., artificial) stability. For example, if

² Psychometricians tend to prefer rank stability over level stability as a criterion for evaluating constructs. For example, Schildberg-Hörisch’s (2018) classically-focused review of risk preference elicitation explains that in contrast to level stability, rank stability “accommodates evidence on systematic changes in risk preferences over the life cycle, due to exogenous shocks such as economic crises or natural catastrophes, and due to temporary changes in self-control resources, emotions, or stress” (p. 136). Economists tend to be more focused on parameter values and hence, to the extent that they consider stability, on level stability. Most of our data is insufficiently granular to permit estimation of level parameter value stability, so we focus on rank stability and on the extensive margin of level stability.

respondents speed through an elicitation by providing salient responses, or consistently use the same heuristic when faced with cognitively taxing questions, their answers will indicate spurious stability even if object of measurement is truly unstable.

Our third key finding is that there is little support for a spurious stability confound in our data. As highlighted above, our estimates of behavioral bias stability within-persons are only moderately high. If respondents were spuriously selecting the same response, three years apart, for whatever reason—gravitating toward salient response options, a preference for appearing consistent, etc.—we would expect to find indicator and rank stability estimates much closer to 1. A related point is that estimated stability varies widely, both across the 17 biases and for biases compared to measures of other constructs including qualitative risk aversion and quantitative cognitive skills. If spurious stability were important, one might expect to see more compression in stability estimates across different variables, measures, and constructs. We also discard low-quality respondents and responses per several standard measures of survey effort and comprehension. This should decrease stability estimates if spurious stability is empirically important, but that is not what we find.

In contrast, several findings are consistent with attenuation pushing stability estimates downward. First, the combination of near-perfect stability of bias indicators at the sample level with substantially lesser stability at the person level suggests that measured within-person instability is the artifact of noise around a constant true level. Second, although most respondents exhibit *some* change in most biases across the two rounds, *large* changes are the exception not the rule. For example, among our more granular elicitations, the median change in bias response category is 18%. Again, this is consistent with respondents exhibiting a trembling hand around a true, stable value. Third, we find higher stability for the four clusters of empirically- and conceptually related biases identified in Stango and Zinman (2023) than for single biases, and especially for the two clusters that explain the most variance in the 17 bias measures. Since each cluster is comprised of multiple biases that are strongly correlated with each other, it could well be that the “different” biases within each cluster provide multiple, relatively strong signals of true stability that cancel out some noise.³ Fourth, bias stability magnitudes are similar to those for

³ A similar pattern exists in more standard psychometric measures, like those for personality, as discussed in Meier and Sprenger (2015, fn 21). We have also, in a prior working paper (Stango and Zinman 2020), pursued an explicit bias-level adjustment for measurement error, using a “measurement error instrumental

standard, quantitative measures of presumed-classical patience and risk aversion in our data. Given that the latter two are commonly presumed by economists to be temporally stable, at least qualitatively speaking, this suggests that elicited estimates of preferences, expectations, and decision rules, whether they explicitly allow for behavioral biases or not, are attenuated by classical measurement error.

Another approach to inferring how much classical measurement error attenuates stability estimates is to look for signatures of we call “true instability”: instability that can be linked to identifiable correlates. The more evidence of true instability, the less classical measurement error likely contributes to measured instability. The question of whether and why basic decision inputs are unstable is of substantive interest as well, with many recent studies focusing on how adverse changes in financial condition, traumatic events, and other stressors can increase risk aversion, decrease patience, and reduce cognitive skills (e.g., Mani et al. 2013; A. N. Meier 2022). The few studies considering behavioral biases have focused on whether predictable, recurring changes in financial resources (Carvalho, Meier, and Wang 2016; Lichand and Mani 2020) and exposure to violence and other stressors (Callen et al. 2014; Haushofer et al. 2021) change measured behavioral biases in low-income samples, finding mixed results. Our data allows us to flag respondents who report large changes in finances or subjective well-being, and other major and likely stressful life events, over our three-year horizon. Yet we find little evidence that such respondents exhibit less (or more) stability in their behavioral bias measures than the rest of our sample. Although we lack the sort of plausibly exogenous variation in liquidity (D. Fehr, Fink, and Jack 2022) or exposure to natural disasters and other trauma (e.g., Beine et al. 2020; Callen et al. 2014) that has been the focus of much related literature, it seems likely that any endogeneity confounding stability estimates would push toward differential stability across respondents with more vs. less stable measured (financial) lives and subjective well-being-- and against our findings of homogeneous stability across such sub-samples.

We contribute most directly to three literatures on the measurement and study of behavioral biases.

variable” approach that instruments for each bias with other, correlated biases. That approach leads to significantly higher estimates of bias-level stability.

One is work on the temporal stability of behavioral biases *per se*. We find comparable within-person estimates for only 4 of the 17 biases we measure, and an aggregate stability estimate for only one.⁴ Altogether, we increase the body of empirical evidence on the temporal stability of behavioral biases by roughly an order of magnitude: we quadruple the number of biases studied, estimate both level and rank stability, and provide estimates for clusters of related biases as well as for each of the 17 biases in isolation.

Two, we contribute to literatures in economics and other disciplines measuring the temporal stability of risk and time preferences/attitudes, by distinguishing beliefs from preferences in most of our elicitations, and using elicitations that allow for behavioral deviations from classical expected utility maximization under rational expectations.⁵ We also provide novel within-sample comparisons of the temporal stability of behavioral biases vs. other individual characteristics with longer research traditions.

A third key contribution is to work on the construct validity of behavioral biases. Specifically, we add to recent work highlighting the importance of accounting for measurement error when making inferences about how well behavioral biases satisfy each of the three key psychometric criteria for construct validity in individual differences; that is, for measured cross-sectional heterogeneity in a characteristic to be meaningful. We consider *temporal stability* in this paper. Chapman et al. (2023) and Stango and Zinman (2023) find evidence of *convergent validity*: “how well different measures of a psychological construct capture a common underlying characteristic

⁴ Please see Section 3 for our discussion of Meier and Sprenger (2015), Tasoff and Zhang (2022), Chapman et. al. (2021), Chapman, Snowberg, Wang, and Camerer (2019; 2022), and other related work. Our work also relates to literatures on the stability of inferences across studies, including replication and external validity issues; see, for example, DellaVigna and Pope (2022).

⁵ For example, for risk aversion, the most temporally-stable measures come from “stated” or “self-report” elicitations (Mata et al. 2018) like the Dohmen et al. measure we use: “... Are you generally a person who is fully prepared to take financial risks or do you try to avoid taking financial risks? Please indicate your answer on a scale from 0-100...” As has been well-documented (e.g., Falk et al. 2018), this sort of elicitation provides a very useful reduced-form summary of what is more accurately described as risk attitudes instead of risk preferences. But it does not help researchers unpack the extent to which the reported risk-taking, and stability thereof, is driven by: (a) beliefs/perceptions about risk vs. preferences regarding uncertainty *per se*; or (b) any of the many behavioral biases hypothesized to affect beliefs and preferences over risk. Several of our elicitations allow for such distinctions. We allow for the possibility of biased misperceptions of probability like non-belief in the law of large numbers, and for the possibility of biased perceptions of intertemporal prices like exponential growth biases that could affect perceived returns to risk-taking. We also allow for the possibility of behavioral preferences *per se* like ambiguity aversion, loss aversion or small-stakes risk aversion, preference for certainty, etc.

or trait” (Mata et al. 2018, p. 156).⁶ Stango and Zinman (2023) find evidence of *predictive validity*: “the extent to which a psychological trait has power in forecasting behavior” (Mata et al. 2018, p. 157). It may be that other studies reporting failures of lab-style measures to satisfy these criteria stem from the failure of researchers to account for measurement error.

We also contribute to several other literatures referenced above, including those on macro modeling with behavioral agents and potential sources of instability of basic decision inputs.

Altogether, our findings suggest three key takeaways about behavioral biases and their measures. One, they are extremely stable in the aggregate. Two, they are at least moderately stable within-person. Measured stability is on par with or exceeds standard quantitative measures of patience and risk aversion, depending on whether one focuses more on single biases or clusters of them, and various pieces of evidence suggest that classical measurement error substantially attenuates within-person stability estimates downward from true stability (we discuss other benchmarks, and their relationship to assumptions about constructs and measurement error, in Section 3). Three, our reliance on circumstantial evidence for inference about measurement error highlights many opportunities for additional work on measurement. We consider caveats to these inferences in the body of the paper and then summarize them, with an eye towards opportunities for future research, in the Conclusion.

⁶ Related literatures on convergent validity consider relationships among measures of a classically-defined construct for different domains, with risk aversion being perhaps the most prominent example in economics. For example, there are literatures that impose expected utility maximization on data from stylized lab-style tasks (see e.g., Mata et al. (2018) for a review) or field choices (e.g., Barseghyan et al. (2011); Einav et al. (2012)) to examine how strongly related choices are across different types of financial products and/or risky behaviors like smoking, self-employment, etc.

1. Research design⁷

A. Variables overview

We focus on two sets of consumer characteristics in this paper. Set one contains behavioral biases, each defined against a clear classical benchmark (as detailed in Section 1-E below and Data Appendix Section 1). Set two contains other characteristics that do not have a clear classical benchmark: demographics, some standard measures of time and risk preferences/attitudes, cognitive skills, and personality traits (Section 1-D and Data Appendix Section 2). We also use data on financial condition, subjective well-being, response times, and response quality, as described in Sections 2-C and 2-D.

B. The American Life Panel

We administered our surveys through the RAND American Life Panel (ALP). The ALP is an online survey panel established in 2003. RAND regularly offers panel members opportunities to participate in surveys designed by researchers for purposes spanning the range of social sciences. Over 400 survey modules have been administered in the ALP, and RAND makes data publicly available after a period of initial embargo.

The ALP takes great pains to obtain a nationally representative sample, combining standard sampling techniques with offers of hardware and a broadband connection to potential participants who lack adequate Internet access. ALP sampling weights match the distribution of age, gender, ethnicity, and income to the Current Population Survey.

C. Research design and sample

Three principles guide our research design. First, measure the richest set of individual characteristics possible, to allow for exploration of relationships among and between behavioral biases and other characteristics. Second, use standard elicitations and survey questions wherever possible, although in many cases we shorten lab-style elicitations to free up time and budget for measuring a broader set of characteristics. Third, take repeated measurements at different points in time, to describe the temporal stability of behavioral biases.⁸

⁷ Portions of this section also appear in Stango and Zinman (2023).

⁸ Repeated measurements also help account for measurement error when addressing other research questions, as in Gillen et al. (2019) and Stango and Zinman (2023).

To those ends, we designed a round of about an hour's worth of elicitations and survey questions. We then split this round into two modules designed to be administered roughly two weeks apart, to minimize survey fatigue and per the ALP's advice regarding module length and invitation sequencing. After extensive piloting and refinements, the ALP fielded our two Round 1 modules starting in November 2014. We targeted 1,500 working-age (age 18 to 60) respondents, sending 2,103 initial invitations, and ultimately received 1,514 responses to Round 1 Module 1 (ALP #315), with 1,426 subsequently completing Round 1 Module 2 (ALP #352). 95% of respondents completing both modules did so by the end of February 2015.

We undertook Round 2 in October 2017, by inviting the universe of 1,308 panelists who completed both Round 1 modules and remained empaneled to take ALP #474, which is a replica of Round 1 Module 1 (but was not advertised as such). We received 967 responses and then invited those panelists to take the second module. ALP #472 is a replica of Round 1 Module 2 with some additional questions added at the end (but was not advertised as such). We received 845 responses to this second module, creating our sample of 845 panelists who responded to both modules in both rounds. Data Appendix Section 3 compares this sample to our initial Round 1 sample of 1,426 and finds little evidence of differential attrition. Moreover, the results in Section 2-D suggest that any drivers or correlates of attrition are unlikely to be correlated with stability.⁹

In refining our elicitations for biases and other consumer characteristics, we held in mind that research budgets force tradeoffs between the depth and breadth of measurements, incentives, and sample size. Per standard ALP practice, we paid panelists \$10 per completed module. Beyond that, all but one of our elicitations are unincentivized on the margin (Table 1), as we decided to allocate our scarce research budget to increasing sample size, increasing the number of measured characteristics (online panel pricing is increasing in survey time), and repeating measurements over time.

⁹ Specifically, Data Appendix Table 2 compares sample means for 38 key measures of biases, patience, risk aversion, cognitive skills, and demographics across the Round 1-only sample of 1,426 and our analysis (both-round) sample of 845. These 38 measures of consumer characteristics are, almost without exception, economically and statistically indistinguishable across the two samples. The only potentially economically significant difference is a 5pp (12%) higher prevalence of college attainment in our analysis sample, although with 38 comparisons one would expect to find at least one such difference purely by chance.

Researchers often hypothesize that subjects find stylized tasks unpleasant and hence need marginal incentives to engage with the tasks, but the ALP measures panelist engagement at the end of each module and roughly 90% of our sample reports that our modules are “Very interesting” or “Interesting”.¹⁰ Section 3-A compares our results to the few other studies with comparable estimates of behavioral bias temporal stability, all but one of which do use marginal incentives, and shows that we obtain similar estimates. Our Data Appendix Section 1-C provides a detailed comparison of elicited data distributions, for a case where our elicitation design hewed closely to one implemented with marginal incentives in a different and plausibly nationally representative sample, and shows that the distributions are remarkably similar.

D. Other characteristics: elicitations and measures

The ALP collects demographics such as income, gender, age, education, and family structure when a panelist first registers, refreshes these measures quarterly, and links them to each module.

Our modules also contain standard elicitations of risk attitudes/preferences from the adaptive lifetime income gamble task developed by Barsky et al. (1997) and the financial risk-taking scale from Dohmen et al. (2010; 2011).¹¹ We measure patience using the average savings rate across the 24 choices in our version of the Convex Time Budget task (Andreoni and Sprenger 2012). We refer to these as measures of “standard” or “classical” preferences/attitudes over risky and intertemporal choice because they are not designed to identify biases *per se*.

We measure cognitive skills using standard tests for general/fluid intelligence (McArdle, Fisher, and Kadlec 2007), numeracy (Banks and Oldfield 2007), cognitive control/executive function (MacLeod 1991; Miyake and Friedman 2012), and financial literacy (Lusardi and Mitchell 2014).¹² Pairwise correlations between these four test scores range from 0.16 to 0.42. In our second round of surveying we add elicitations of the Big Five personality traits to the end of our second module.¹³

¹⁰ Only about 3% reply “Uninteresting” or “Very uninteresting”, with the remaining 7% reporting that our modules are “Neither interesting nor uninteresting”.

¹¹ These Barsky and Dohmen et al. measures are correlated 0.14 in our sample. We also elicit Dohmen et al.’s general risk-taking scale, which is correlated 0.68 with the financial scale.

¹² The Data Appendix Section 2 provides details on each of these elicitations and measures.

¹³ Specifically, we use the validated 10-item version of the Big Five personality trait inventory (Rammstedt and John 2007). We initially decided against eliciting personality measures, given our resource constraints

E. Behavioral bias elicitations and measures

We measure 17 potential biases. Table 1 summarizes our measures and elicitations, with the Data Appendix Section 1 providing details, including granular data descriptions and comparisons of data quality indicators and descriptive statistics to prior work.

Two potential biases relate to discounting, over consumption (snacks) and money (Read and van Leeuwen 1998; Andreoni and Sprenger 2012). Three biases relate to inconsistent and/or dominated choices, including inconsistency with the Generalized Axiom of Revealed Preference (Choi et al. 2014), and narrow bracketing (Rabin and Weizsäcker 2009). Four biases capture risk attitudes: loss aversion/small-stakes risk aversion (E. Fehr and Goette 2007), preference for certainty (Callen et al. 2014) and ambiguity aversion (Dimmock et al. 2016). Another four biases capture three varieties of overconfidence: about absolute performance, relative performance and forecast precision (Moore and Healy 2008). Four biases in mathematical operations include two sources of statistical fallacies—non-belief in a process known to be i.i.d. (which can manifest as either the gambler’s or hot hand fallacies), and non-belief in the law of large numbers (Dohmen et al. 2009; D. Benjamin, Moore, and Rabin 2017; D. Benjamin, Rabin, and Raymond 2016)—and exponential growth biases over borrowing costs and investment returns (Stango and Zinman 2009; Levy and Tasoff 2016). We also elicit limited attention and limited memory, in ways that are motivated by behavioral models of inattention (Ericson 2011; Bronchetti et al. 2023).¹⁴ Many of these biases are correlated with each other within-person, and in Section 2-C we also consider the temporal stability of the clusters of biases identified in Stango and Zinman (2023).

Eight of the 17 bias elicitations identify potential deviations from the classical benchmark in two directions (Table 1 Column 3). For example, discounting can be either present-biased or future-biased relative to time consistency; someone can be either over- or under-confident; and so on. A small subset of our results, on extensive margin stability, considers the different directions separately, for reasons we discuss in Section 2-A and footnote 18. The other nine elicitations identify potential deviations from the classical benchmark in a single direction, either by constraint

and the lack of prior evidence of correlations between them and behavioral biases; for example, Becker et al.’s (2012) review article does not mention biases.

¹⁴ Following a common delineation in behavioral economics, we do not measure social preferences. See Dean and Ortoleva (2019) and Chapman et al. (2023) for evidence on relationships between and among other behavioral biases and social preferences.

or per the underlying construct. For example, our elicitations identify whether someone is ambiguity averse or not, and whether they make choices consistent with classical utility maximization or not.

Some of our elicitations are relatively rich and produce granular information on bias intensity: the extent to which someone deviates from a classic benchmark. Others are relatively coarse, with two producing just a 0/1 indicator of whether someone is biased or not: consumption discounting and limited prospective memory. The median number of distinct values produced by our elicitations, across our 17 biases, is 17 (Table 1 Column 4).

For estimating level stability, we use the extensive margin of each bias in each round: 1 if biased, 0 otherwise. Measuring the extensive margin is feasible in a wide range of settings because it does not require an elaborate elicitation or granular bias measures. It is desirable substantively because bias indicators map well into models with a mix of biased and unbiased agents, and methodologically because it facilitates cross-bias comparisons by providing pertinent and standardized units.¹⁵

For stability estimates that include the intensive margin of deviations from a rational benchmark, we measure the respondent's cross-sectional percentile rank for each bias in each round, on a uni-dimensional scale.¹⁶ As with our measure of bias level, rank is easily comparable across biases.¹⁷ Rank has the added benefits of capturing both the extensive and intensive margins, and of conforming to the psychometric standard for measuring stability (as discussed above in footnote 2).

¹⁵ A complementary approach to estimating level stability would be to do so for parametric measures of bias intensity. But many of our elicitations are too coarse to produce informative parametric bias measures.

¹⁶ Our uni-dimensional scales for potentially bi-directional biases place unbiased in the middle, such that someone who is slightly biased in one direction in one round, and slightly biased in the other direction in our other round, is inferred to be relatively stable. In our working paper (Stango and Zinman 2020), we consider each direction separately by setting the directional value to zero if someone does not exhibit that directional bias. For example, if someone is underconfident in a given round we set their value for overconfidence to zero for that round. Our findings are qualitatively and quantitatively similar under both of these approaches.

¹⁷ There are various ways of measuring rank, and they are very highly correlated with each other in our data.

2. Results

This section reports and interprets our results. We then provide detailed comparisons to other work, including various benchmark standards of stability, in Section 3.

A. Biases are stable in the aggregate

Table 2 describes aggregate bias temporal stability across our two survey rounds. Aggregate stability is of course pertinent only for levels not ranks. Another difference from our subsequent main tables, on within-person temporal stability in both levels and ranks, is that here we consider the different directions of potentially bi-directional biases separately. These are particularly interesting in the aggregate because many studies make assumptions about the relative importance of the two directions. For example, to help with identification, a structural model of time-inconsistent discounting may impose an assumption of no future bias, on the grounds that it is not empirically important. Examining future bias and present bias separately helps assess the soundness of such assumptions.

We examine aggregate stability by estimating whether and how the sample or population proportion exhibiting each (directional) bias changes across our two rounds. Columns 1 and 2 does this for our sample (unweighted estimates), while Columns 3-5 does this for the U.S. working-age population using sampling probability weights provided by the ALP. Results are similar with and without weights. Although bias prevalence varies substantially across biases (reading down the rows in Panel A), bias stability does not (whether we compare Column 1 to Column 2, or Column 3 to Column 4).

The key inference here is that biases are remarkably stable, in the aggregate, over our three year horizon. Focusing on Column 5, the hypothesis of perfect stability— of equality in proportions across the two rounds-- is clearly rejected for only two of the 24 (directional) biases with estimates here. This is about what one would expect to get by chance: we should expect one or two false rejections out of 24 draws. Another way to address multiple hypothesis testing is to aggregate results across biases, as we do in Panel B. This exercise is further motivated by the fact that most biases are correlated with each other within-person (Stango and Zinman 2023), suggesting that each bias measure provides a signal about some underlying construct (see also Section 2-C). Estimated mean bias prevalence is essentially unchanged across the three years, at

0.42 or 0.43 in both rounds. The mean absolute value change is only 2 percentage points. The maximum absolute value change is 9 percentage points, but even this change is modest in percentage terms given high prevalence for that bias: $9/83 \Rightarrow 11\%$.

These results provide support for macroeconomic models assuming fixed (sub-)populations of behavioral agents. The mean-zero change across biases is also noteworthy for inferences about stability at the micro level, as it is consistent with classical measurement error in consumer-level bias measures (see also Section 2-C).

B. Initial estimates suggest that biases are moderately stable within-person

We now turn from aggregate to within-person bias stability. Table 3 provides estimates for both level (Columns 1-5) and rank stability (Column 6). Here we do not examine the different directions of potentially bi-directional biases separately, in large part because uni-dimensional scales, with unbiased in the middle, are likely more informative for rank stability (please refer back to footnote 16 for details).¹⁸ We present only sample probability-weighted estimates in the main table, as a concession to table readability. Since best practice is arguably to present unweighted estimates as well (Solon, Haider, and Wooldridge 2015), as we do in Table 2, Appendix Table 2 is an unweighted version of Table 3. Results are very similar with and without weights.

Starting with our measures of level temporal stability, Table 3 Columns 1-4 present the estimated sample shares for each cell in the 2x2 transition matrix for the extensive margin of each bias: (biased or unbiased in Round 1) x (biased or unbiased three years later).¹⁹ Summing the first (next) two columns gives the proportion of individuals who are stable (unstable) across the two rounds. As an example, 70% of our sample exhibits ambiguity aversion in both rounds (57%) or neither round (13%). The range across biases is [0.57, 0.99]. The cross-bias mean is 0.72. We interpret this as indicating moderate stability, given that 0.50 would suggest no stability and 1.00 perfect stability. Column 5 takes a complementary approach to estimating magnitudes, using the

¹⁸ Appendix Table 1 presents level stability results separately by bias direction. This exercise reveals that directional biases more widely accepted and studied by behavioral scientists—“standard direction” biases, in our table’s parlance (e.g., overconfidence not underconfidence)—do tend to be substantially more stable than their “non-standard” directional counterparts, with only one exception out of eight directional pairs (Column 6).

¹⁹ As the estimates in Columns 1-4 are of sample proportions, they are each precisely estimated, with standard errors of 0.02 or lower.

tetrachoric correlation estimator, and similarly suggests moderate stability. The identifiable point estimates range from 0.10 to 0.79 (see also Figure 1),²⁰ with standard errors [0.03, 0.13]. The mean cross-bias point estimate is 0.43 or 0.46, depending on an identifying assumption detailed in the table notes.

Turning to rank stability, Table 3 Column 6 reports polychoric (i.e., Pearson or tetrachoric) correlation estimates of rank stability across our two survey rounds.²¹ The point estimates range from 0.10 to 0.60 across the 17 biases (see also Figure 1), with the exception of one outlier discussed in the next paragraph. Standard errors range from 0.02 to 0.07. The mean estimated rank correlation across biases is 0.31. Tetrachoric correlation estimates employ particular statistical assumptions,²² but Figure 1 illustrates that biases we estimate to be more stable in levels also tend to be more stable in rank, and the bias-level correlation between the tetrachoric and rank correlations is 0.77 (SE=0.18, N=15).

The one outlier among the rank correlation estimates is money discounting, at 0.01 (SE=0.03). We caution against confidently inferring that money discounting biases are unstable from this result, as one might expect to obtain one false zero out of 17 by chance. Moreover, the extensive margin results in Columns 1-5 suggest at least some degree of stability, albeit less so than all but two of the other biases. Further research seems warranted.

Overall, Table 3 suggests that biases tend to be moderately stable within-person over time. We next explore whether and how measurement error changes this inference.

C. Measurement error: Spurious stability vs. attenuation

As noted at the outset, measurement error could, in principle, statistically bias stability estimates in either direction. One concern is attenuation from noise. For example, if respondents speed through the survey and pick some answers at random, or guess when faced with cognitively

²⁰ Point estimates are not identified for two of the biases, as detailed in the table notes.

²¹ The polychoric estimator optimally chooses tetrachoric or Pearson based on the degree of smoothness in the measure (see Kolenikov and Angeles (2009) for more details on its statistical properties and uses in economics). One can think of the Pearson rank correlations for more continuous variables as being similar to a Spearman correlation, which is also based on rank values, although the “ranks” for percentiles have a cardinal component when there are fewer than 100 response categories.

²² The tetrachoric correlation assumes a continuous latent trait, jointly normally distributed, as well as independent and normally distributed measurement errors across observations.

taxing questions, their answers will spuriously vary over time even if the object of measurement is truly stable. Another concern is that respondents may exhibit spurious stability. For example, if respondents speed through the survey by providing salient responses, or consistently use the same heuristic when faced with cognitively taxing questions, their answers will indicate spurious stability even if object of measurement is truly unstable.

In practice, we find little support for a spurious stability confound in our data in three sets of results.

First, as highlighted above, our estimates of behavioral bias stability within-persons are only moderately high. If respondents were spuriously selecting the same response, three years apart, for whatever reason—gravitating toward salient response options, a preference for appearing consistent, etc.—we would expect to find indicator and rank stability estimates much closer to 1. Indeed, looking at the raw data, only 36% of bias response pairs are exactly the same across the three years, and this falls to 23% for our more granular elicitations (Table 4, Panel B, Column 1).

A second and related point is that estimated stability varies widely, both across the 17 biases (Tables 2-4) and for biases compared to measures of other constructs including qualitative risk aversion and quantitative cognitive skills (Table 5). If spurious stability were important, one might expect to see more compression in stability estimates across different variables, measures, and constructs.

Third, we discard low-quality respondents and responses per several standard measures of survey effort and comprehension. This should decrease stability estimates if spurious stability is empirically important, but that is not what we find in Table 6.²³ We start with our full sample estimates in Column 1 (reproduced from Table 3 Column 6) and then in Columns 2-6 discard respondents with unusually fast responses in both rounds or low conscientiousness (Dohmen and Jagelka, forthcoming), those who report finding the survey uninteresting, or those who perform poorly on what are likely the two most effortful tasks we use to measure cognitive skills. Comparing the average difference between the full sample and those five presumably higher-

²³ Table 6 focuses on rank stability because we are interested in the importance of non-classical vs. classical measurement error; conversely, any measurement error in indicators is non-classical by definition.

quality sub-samples (Column 7), we find a minimum difference across the 17 biases of only -3 percentage points, and a mean difference of -1 (Panel B).²⁴

In contrast, several findings are consistent with attenuation pushing stability estimates downward.

First, the combination of near-perfect stability of bias indicators at the sample level (Table 2) with substantially lesser stability of those same indicators at the person level (Table 3) suggests that measured within-person instability is the artifact of noise around a constant true level.

Second, although most respondents exhibit *some* change in most biases across the two rounds, *large* changes seem to us more the exception not the rule. Table 4 describes this based on the raw data. Across all 17 biases, the mean number of response categories is 30 (Column 3), and the mean change in responses across our two rounds is 5 (Column 2). Across biases with median or greater granularity (i.e., with ≥ 17 response categories), the mean category change is 9 on a base of 52, for an estimated 18% change. The (in)frequency of “large” changes is in the eye of the beholder here, but to us the finding that less than half of the sample changes their responses by more than 20%, together with the moderate degree of stability found in Table 3, again suggests a trembling hand around a true, stable value.

Third, we find higher stability for the four behavioral common factors (BCFs) identifying empirically- and conceptually-related biases (Stango and Zinman 2023) than for single biases (Table 7). Estimated within-person intertemporal correlations in the BCF factor scores range from 0.22 to 0.65 across the four BCFs and two specifications, with standard errors of 0.01 or 0.02. Correlations are highest for the two BCFs that explain the most variance in the 17 bias measures (see Section 2-E for a discussion of what can(not) be learned from this about which biases are more stable than others). Since each BCF is comprised of a cluster of multiple biases that are

²⁴ Not shown in Table 6 due to space constraints are some additional checks of whether lower-quality responses or respondents spuriously inflate estimated stability. For example, we drop respondents with any internally inconsistent response, in any of the bias elicitation that allow for identification of internal inconsistency (not including the GARP consistency elicitation, where inconsistency is the object of interest), and find little if any evidence that stability estimates decrease. We also estimate rank stability separately by categories of “respondent quality” common factor scores extracted from flexibly parameterized response times, measures of internal consistency and item non-response, and respondent characteristics. The most important common factor loads only on survey time, so unsurprisingly (in light of Table 6 Column 2), we do not find strong evidence that stability estimates vary with respondent quality.

strongly correlated with each other—more specifically, of biases that “load” relatively strongly on that BCF, in factor analysis parlance-- it could well be that the “different” biases within each cluster provide multiple, relatively strong signals of true stability that cancel out some noise.²⁵ A similar empirical pattern and interpretation exists for more standard psychometric measures, like those for personality, as discussed in, for example, Meier and Sprenger (2015, fn 21). In a related vein, in a previous draft of this paper we pursued an explicit bias-level adjustment for measurement error, using a “measurement error instrumental variable” approach that instruments for each bias with other, correlated biases. That approach also leads to significantly higher estimates of bias-level stability (Stango and Zinman 2020).

Fourth, bias stability magnitudes are similar to those for standard, quantitative measures of presumed-classical patience and risk aversion in our data, as detailed below in Section 3-B. Given that the latter two are commonly presumed by economists to be temporally stable, at least qualitatively speaking, this suggests that elicited estimates of preferences, expectations, and decision rules, whether they explicitly allow for behavioral biases or not, are attenuated by classical measurement error.

One set of findings potentially tempers our inference that classical measurement error is likely important: dropping low-quality respondents, as we do in Table 6, does not substantially increase stability estimates as one might expect. This may be because the ALP pre-screens for good respondents—sanctioning people with low-effort response patterns on previous surveys, and/or removing them from the Panel—but not good responses.

D. True instability: Our data does not rule it out, but we do not we find clear evidence for it

Another approach to inferring how much classical measurement error attenuates stability estimates is to look for signatures of we call “true instability”: instability that can be linked to identifiable causes or correlates (Table 8). The more evidence of true instability, the less classical measurement error likely contributes to measured instability. True instability is also substantively interesting, as we motivate in the Introduction.

²⁵ An alternative or complementary hypothesis is that the latent factor (here, a BCF) is truly more stable than its emergent components (here, single biases).

The existing literature on behavioral biases has focused on identifying causes of any true stability, using quasi-experimental or lab-experimental variation in predictable, recurring changes in financial resources (Carvalho, Meier, and Wang 2016; Lichand and Mani 2020) or exposure to violence and other stressors (Callen et al. 2014; Haushofer et al. 2021) test whether these events change measured behavioral biases. Fehr et al. (2022) also examine cross-sectional variation in wealth, and randomized access to consumption loans. Those studies have focused on low-income samples and find mixed results.

Relative to the existing literature, our setup offers advantages and disadvantages for identifying true stability. On the plus side, we have good and an unusually broad set of measures of the sort of major, potentially stressful events or indicators that the literature has focused on: we can capture households with major (adverse) changes in household composition, employment status, income, and objective measures of financial distress. We can also measure changes in subjective well-being that might be caused by, or correlated with, those sorts of major changes in life circumstances. There is substantial heterogeneity in each of these changes across households in our sample, suggesting that we are capturing idiosyncratic shocks and not common ones that should leave rank stability unchanged. On the minus side, we lack plausibly exogenous sources variation in the events or stressors of interest. Nevertheless, observational relationships between major life changes/stressors and bias changes are informative for our purposes. Specifically, if one grants that any endogeneity would likely push toward finding differential bias stability across sub-samples respondents with more vs. less stable (financial) lives and subjective well-being, then finding similar stability across such samples is inconsistent with true instability being empirically important.²⁶ We do indeed find similar stability, as detailed next. In other words, if the life changes we measure were correlated with bias changes through some unidentified causal mechanism, then that would likely push against the empirical pattern shown in Table 8.

Table 8, which adopts the same setup as Table 6, shows that we find no evident empirical signatures of true instability. Specifically, we compare the full sample rank correlation point estimate for each bias (reproduced in Column 1) with six sub-samples that drop the most unstable households in terms of household structure, employment, income, financial distress, or subjective

²⁶ Conversely, if one finds differential bias stability across such sub-samples, that could be due to endogeneity and/or to true stability.

well-being in the financial or life satisfaction domains in Columns 2-7. In other words, we compare the full sample to sub-samples of relatively stable households, with varying sample sizes and along various pertinent dimensions, to check whether doing so changes estimated bias stability. There is little if any evidence of differential stability. Column 8 reports the difference between the full sample point estimate and the mean of the sub-sample point estimates, within each bias. This difference ranges from -0.03 to 0.05 (i.e., from -3 to 1 percentage point(s)), with a mean of 1 pp.

In summary, we look for empirical signatures of true instability in biases and do not find them. It could be that we are missing some key margin of heterogeneity that is correlated with bias instability but not instability in the life circumstances or subjective well-being proxies we do measure. This strikes us as unlikely—for example, an unobserved disaster or trauma would likely affect family structure, finances, and/or subjective well-being— but merits continued research. It could be that true bias instability is homogenous across households. We are not aware of any theory or empirical evidence supporting that conjecture but posit that it too is worthy of investigation.

Together with the constellation of evidence considered in Section 2-C, the lack of compelling evidence for true instability reinforces our speculative inference that classical measurement error is the main reason why estimated stability is $\ll 1$. We consider the sorts of additional work needed to sharpen or change that inference in the Conclusion.

E. Are some biases more stable than others?

Figure 1 and Table 7 suggest—but only very speculatively-- that some biases are more stable than others. Specifically, it appears that biases related to beliefs and decision quality, which Stango and Zinman (2023) link to processing-based theories of behavioral biases, are more stable than those related to time or risk preferences. It could be that our measurements of belief and decision quality biases capture something more fundamental and stable about consumer decision making and limitations therein. Or this could be a mirage, in whole or in part, if measurement error attenuates stability estimates and that attenuation varies across biases. In other words, it might well be the case that some biases have higher estimated stability because they are measured with less noise (see Sections 3-B through 3-D for related discussion comparing bias measurement to that of cognitive skills and personality). We suspect that making progress on this question will require randomized variation in elicitation design, and/or new measurement error instrumental variables strategies, and discuss some possibilities in the Conclusion.

3. Benchmarking stability estimates

We now consider our estimates of behavioral bias stability in comparison to benchmarks from our data on other individual characteristics (Table 5) and various literatures. It will be useful to keep in mind that our mean rank stability estimate across single biases is 0.31 (first shown in Table Table 3, Panel B, Column 6), and that our mean factor score stability estimate across clusters of related biases is 0.46 (= mean of the eight point estimates in Table 7).

A. Other work on behavioral biases

For the aggregate prevalence estimates in our Table 2, the only other estimate we know of for the temporal stability of a bias sample or population share can be inferred from Chapman et al. (2022) Figure 4, which suggests that the aggregate share of loss averse consumers is almost perfectly stable over their 6-month horizon.

For within-person stability, we are not aware of any work estimating the within-person stability of the extensive margins of behavioral biases, as we do in, for example, Table 3 Columns 1-5. But prior work does provide a point of comparison for our estimates of bias intensity in Column 6, for four of our 17 behavioral biases. For present-biased money discounting, the most comparable estimates are from Tasoff and Zhang (2022), which also estimates rank correlations using data from Convex Time Budget elicitation in the ALP, in their case for a short-run discounting parameter. Their estimate is 0.03 over about 16 months, as compared to our estimate 0.01 over about three years. For choice inconsistency, Chapman et al. (2019) find a 6-month correlation 0.24 for a structurally estimated parameter, as compared to our three-year rank correlation estimates of 0.17 and 0.32 for our two measures of inconsistency with GARP. For loss aversion, Chapman et al. (2022) find a 6-month within-person correlation of 0.40 for a structurally estimated parameter, as compared to 0.49 for our relatively coarse measure. For ambiguity aversion, Chapman et al. (2021) find a 6-month within-person correlation of 0.21, as compared to our 0.27.²⁷

Overall, our estimates bias stability are quite similar to those found in other studies. Each of those studies use marginal incentives, except for Tasoff and Zhang.

²⁷ Callen et al. (2014) examine how exposure to violence changes another one of the biases we consider, preference for certainty, but they do so only in a cross-section.

B. Stability estimates for other decision inputs

As detailed in Section 1-D, our data permits in-sample comparisons between behavioral biases and more commonly measured decision inputs: presumed-classical patience and risk aversion, and cognitive skills (Table 5). We also compare to the literatures on those other inputs and to personality measures.

For patience, our estimated intertemporal rank correlation of 0.31 is most comparable to Meier and Sprenger's (2015) estimated long-run discount factor stability of 0.25 (using multiple price lists elicited one year apart), and to Chapman et al.'s (2022) estimated discount factor stability of 0.47 (using Dynamically Optimized Sequential Experimentation elicited 6-months apart, and finding greater stability with DOSE than with multiple price lists).²⁸ For risk aversion, we estimate correlations of 0.28 from the Barsky et al. (1997) income gamble elicitation, and 0.59 from the Dohmen et al. (2010; 2011) qualitative financial risk-taking scale. Mata et al.'s (2018) meta-analysis of risk aversion suggests 3-year stability of around 0.2 for measures gleaned from lottery-choice elicitations like Barsky et al.'s, +/- about 0.1.²⁹ Mata et al.'s meta-analysis of measures gleaned from self-reported risk-taking elicitations like Dohmen et al.'s suggests precisely estimated 3-year stability of about 0.5.

Since our behavioral bias elicitations are nearly all quantitative, the most apt comparisons here are to quantitative elicitations of patience and risk aversion. As discussed in the previous paragraph, these tend to produce temporal stability estimates centered in the 0.2 to 0.4 range, both in our sample (0.31 for patience and 0.28 for large-stakes risk aversion) and in prior work. Keeping in mind that our cross-bias mean rank stability estimate is 0.31, single biases thus have measured stability on par with parameters that are economic staples of modeling consumer choice over time and under risk. Groups of related biases have greater measured stability (as Table 5 shows, the mean of our eight estimates of bias-cluster stability in Table 7 is 0.46).

²⁸ See Chuang and Schecter (2015, Table 2) for a review of 7 other studies, mostly on convenience samples and all covering shorter horizons than ours, and for some evidence of their own from rural Paraguay.

²⁹ Subsequent to Mata et al.'s meta-analysis, Chapman et al. (2022) find 6-month stability of 0.44 for their DOSE-elicited lottery choice elicitation, with DOSE yielding greater stability than MPLs. And Tasoff and Zhang (2022) find risk aversion parameter stability, estimated from Barsky et al. elicitation in the ALP administered at varying time intervals, ranging from 0.08 to 0.58. They find that other lottery-choice methods in the ALP have within-person and -elicitation stability ranging from 0.27 to 0.53, again over various horizons.

Turning to cognitive skills, our four standard measures have precisely estimated intertemporal rank correlations ranging from 0.48 to 0.78. Standard measures of fluid intelligence-- like those based on our number series elicitation, or IQ tests-- have long been shown to be quite stable (Jensen 1998). For personality measures, we only elicit them in Round 2 and therefore lack in-sample stability estimates, but meta-analysis of the literature finds a precisely estimated 3-year within-person rank correlation of about 0.6 (Anusic and Schimmack 2016). As discussed in Meier and Sprenger (2015, fn 21), intertemporal correlations for the individual components of personality measures—that is, for each of the single questions that researchers combine to create measures of personality characteristics—tend to be much lower, in the 0.2 to 0.3 range.

C. Other benchmarks from psychometrics literatures

Work in psychometrics also offers benchmarks for adequately-sized test-retest correlations, although these are typically applied in cases where the tester has a strong prior that the construct they seek to measure is essentially stable. These standards are basically measurement error tolerances, and as such they vary across fields, settings, and use cases (Joint Committee 2014; Lance, Butts, and Michels 2006). We have found thresholds as low as 0.4 (Cicchetti 1994), with 0.7 commonly applied (Lance, Butts, and Michels 2006; Litwin 2003). But any such threshold begs the question of whether low estimated stability is due to a faulty measure (one that produces too much measurement error) and/or to true instability in the latent construct.

D. Summarizing our results in light of benchmarks

Overall, we find that behavioral bias measures, and measures of patience and risk aversion that map best into most economic models, tend to have estimated intertemporal stability substantially below that of standard summary measures of cognitive skills and personality. We highlight two explanations that strike us as plausible, considering related literatures and our findings on the likely importance of attenuation in behavioral bias stability estimates (Sections 2-C and 2-D). First, cognitive skills and personality have much longer and richer histories of measurement research, and so we should expect measures of them to be more refined and less laden with error. Second, common conceptual frameworks of traits and construct validity likely imply that measures are selected in part based on stability. For example, since personality is widely assumed to be largely stable, a measure of some component of personality will become standard only if it is found to be

stable.³⁰ More research is needed on these and other potential explanations, and we suggest several related lines of inquiry below.

4. Conclusion

We examine the temporal stability of 17 behavioral biases, each measured using the same elicitation completed by the same subjects in 2014 and 2017 in a nationally representative online panel survey of U.S. adults. We find that behavioral biases are extremely stable in the aggregate and at least moderately stable within-person. Our estimates of within-person stability for single biases are on par with those for standard quantitative measures of patience and risk aversion. Comparable estimates for clusters of related biases exceed those for standard quantitative measures of patience and risk aversion. We find little evidence of cross-person heterogeneity in bias stability, which is perhaps surprising given the likely empirical importance of cross-person heterogeneity in the biases themselves (Stango and Zinman 2023). One potential explanation is that biases truly are more or less uniformly stable, and indeed various pieces of evidence here suggest that classical measurement error substantially attenuates within-person stability estimates.

Despite greatly expanding the evidence base on the temporal stability and construct validity of behavioral biases and related constructs, our work here has several limitations that highlight opportunities for future research. One is that we measure only a subset of the biases that interest social scientists and policymakers. Another data limitation leads us to consider level stability only on the extensive margin, and not of behavioral parameter estimates. Measuring stability over longer horizons than three years will be useful, particularly for investigations of learning about biases and how to manage them.³¹ Nor do our data and design allow us to infer much about cross-

³⁰ Thanks to one of our referees for this point and example about measure selection.

³¹ It may be the case that learning about one's biases is limited (e.g., Gagnon-Bartsch, Rabin, and Schwartzstein 2023), and context- and behavior-specific (e.g., Allcott et al. 2022; Heidhues, Koszegi, and Strack 2023; Le Yaouanq and Schwardmann 2022), and/or that learning leaves underlying biases largely unchanged but affects behavior by increasing sophistication about how to manage them or in “model-free” ways (e.g., Laibson 2018; Barberis and Jin 2023). A related point is that our evidence of stable biases does not rule out dynamics in macroeconomic expectation forecast errors and biases; we do not elicit forecasts but speculate that they emerge from interactions between market forces and the sort of persistent tendencies toward bias that we do measure. This is but one of many classes of theories of belief distortions; for a recent discussion see, for example, Bianchi, Ludvigson, and Ma (2022).

bias heterogeneity in biases (although our bias cluster results are suggestive), for reasons detailed next.

Perhaps most strikingly, our reliance on circumstantial evidence for inference about how measurement error affects stability estimates highlights many opportunities for additional work on measurement.

One opportunity is examining data-quality returns to elicitation granularity. Our results suggest that elicitations using more questions produce inferences of lower stability, but we lack the requisite variation *within* measured object, in granularity or other elicitation design features, that is required for producing sharp inferences. Both standard practice and some recent evidence suggest that there are meaningful returns to repetition, at least to a point, particularly when it is customized (e.g., Chapman, et al. 2019) and/or probing (e.g., D. J. Benjamin, Fontana, and Kimball 2021) based on previous responses.

A related line of inquiry that is seemingly, but not necessarily, in conflict with repetition is returns to elicitation simplification (Belzil and Jagelka, 2024; Enke and Shubatt 2024). It may be that the optimal design is to ask repeated, adaptive, but relatively simple questions designed to measure the same object, subject of course to constraints on researcher budgets and subjects' time and attention.

Another design feature commonly associated with response quality is paying some marginal financial incentive. Our results could be read as questioning the conventional wisdom that paying marginal incentives reduces measurement error-- three of the four comparable stability estimates from prior studies are based on elicitations using marginal incentives, which our elicitations do not use, yet our estimates are similar to theirs. But we again emphasize that we lack the within-object randomized variation in incentives, and other design features, that would facilitate sharper inferences about how elicitation design affects measurement error and inferred stability.

Some intriguing new approaches confront measurement error with respondent-level information about self-knowledge and self-reports. Falk et al. (2021) structurally estimate self-knowledge from survey data. Dohmen and Jagelka (forthcoming) directly ask about how well a respondent's answers reflect their true characteristics. These approaches could be complements or substitutes for measurement error instrumental variables approaches that rely on the stability of

measured objects (e.g., Gillen, Snowberg, and Yariv 2019) but are challenging to use when attempting to make inferences about the magnitude of stability itself (Stango and Zinman 2020).

Overall, we infer that continued work on survey and elicitation design, in concert with the continued development and application of statistical techniques for dealing with measurement error, holds great promise for improving the accuracy of measurements and temporal stability estimates.

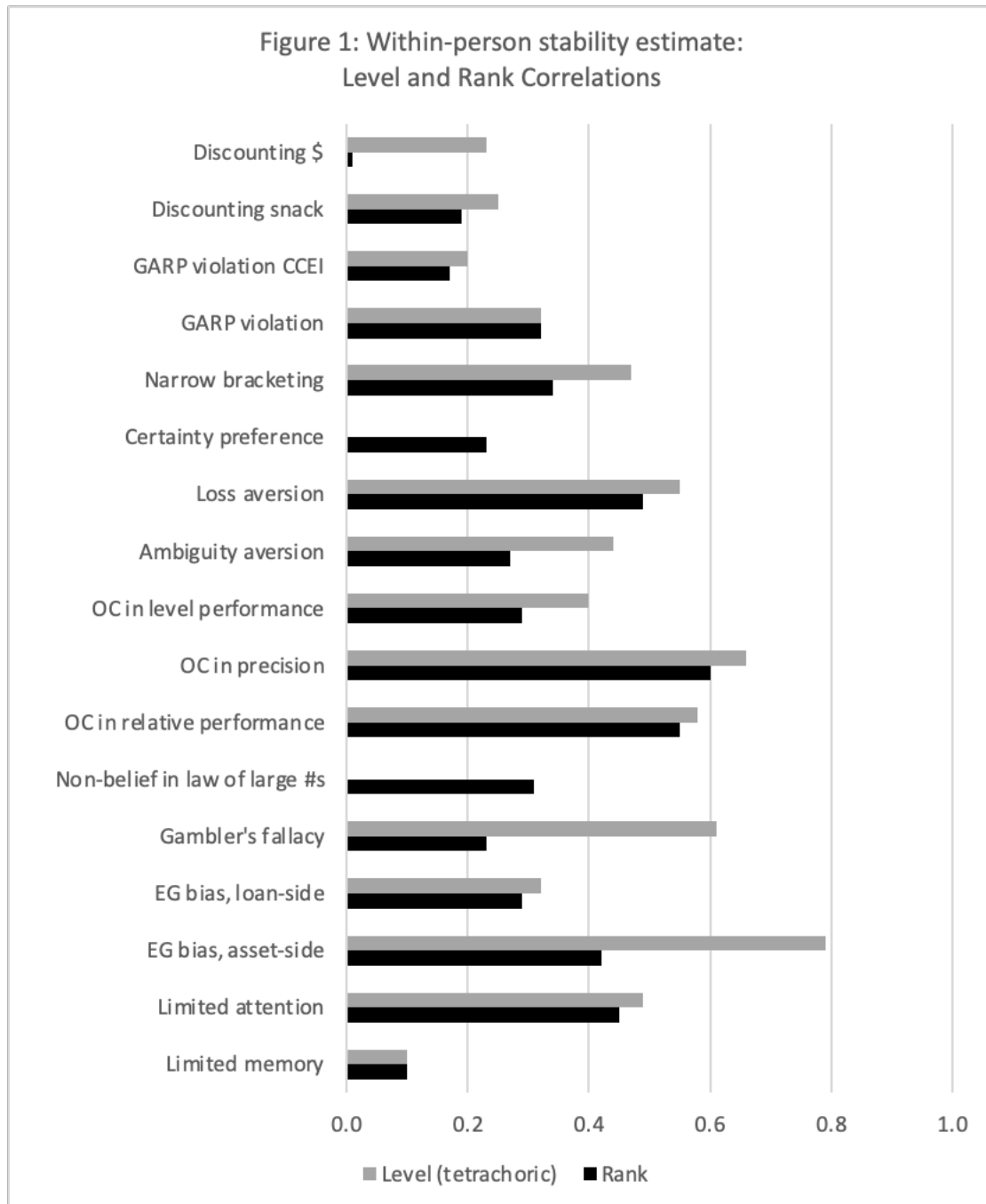
References

- Allcott, Hunt, Joshua Kim, Dmitry Taubinsky, and Jonathan Zinman. 2022. "Are High-Interest Loans Predatory? Theory and Evidence from Payday Lending." *Review of Economic Studies* 89 (3): 1041–84.
- Andreoni, James, and Charles Sprenger. 2012. "Estimating Time Preferences from Convex Budgets." *The American Economic Review* 102 (7): 3333–56.
- Anusic, Ivana, and Ulrich Schimmack. 2016. "Stability and Change of Personality Traits, Self-Esteem, and Well-Being: Introducing the Meta-Analytic Stability and Change Model of Retest Correlations." *Journal of Personality and Social Psychology* 110 (5): 766–81.
- Banks, J., and Z. Oldfield. 2007. "Understanding Pensions: Cognitive Function, Numerical Ability, and Retirement Saving." *Fiscal Studies* 28 (2): 143–70.
- Barberis, Nicholas, and Lawrence Jin. 2023. "Model-Free and Model-Based Learning as Joint Drivers of Investor Behavior."
- Barseghyan, Levon, Jeffrey Prince, and Joshua C. Teitelbaum. 2011. "Are Risk Preferences Stable across Contexts? Evidence from Insurance Data." *American Economic Review* 101 (2): 591–631.
- Barsky, Robert B, F. Thomas Juster, Miles S Kimball, and Matthew D Shapiro. 1997. "Preference Parameters and Behavioral Heterogeneity; An Experimental Approach in the Health and Retirement Study." *Quarterly Journal of Economics* 112 (2): 537–79.
- Becker, Anke, Thomas Deckers, Thomas Dohmen, Armin Falk, and Fabian Kosse. 2012. "The Relationship Between Economic Preferences and Psychological Personality Measures." *Annual Review of Economics* 4 (1): 453–78.
- Beine, Michel, Gary Charness, Arnaud Dupuy, and Majlinda Joxhe. 2020. "Shaking Things up: On the Stability of Risk and Time Preferences."
- Belzil, Christian, and Tomáš Jagelka. 2024. "Separating Preferences, Skills, and Other Latent Personal Attributes from Endogenous Effort and Cognitive Noise."
- Benjamin, Daniel J., Mark Fontana, and Miles S. Kimball. 2021. "Reconsidering Risk Aversion."
- Benjamin, Daniel, Don Moore, and Matthew Rabin. 2017. "Biased Beliefs about Random Samples: Evidence from Two Integrated Experiments."
- Benjamin, Daniel, Matthew Rabin, and Collin Raymond. 2016. "A Model of Nonbelief in the Law of Large Numbers." *Journal of the European Economic Association* 14 (2): 515–44.
- Bianchi, Francesco, Sydney Ludvigson, and Sai Ma. 2022. "Belief Distortions and Macroeconomic Fluctuations." *American Economic Review* 112 (7): 2269–2315.
- Bronchetti, Erin, Judd Kessler, Ellen Magenheimer, Dmitry Taubinsky, and Eric Zwick. 2023. "Is Attention Produced Optimally?" *Econometrica* 91 (2): 669–707.
- Callen, Michael, Mohammad Isaqzadeh, James D Long, and Charles Sprenger. 2014. "Violence and Risk Preference: Experimental Evidence from Afghanistan." *The American Economic Review* 104 (1): 123–48.
- Carvalho, Leandro, Stephan Meier, and Stephanie Wang. 2016. "Poverty and Economic Decision-Making: Evidence from Changes in Financial Resources at Payday." *American Economic Review* 106 (2): 260–84.
- Chapman, Jonathan, Mark Dean, Pietro Ortoleva, Erik Snowberg, and Colin Camerer. 2021. "On the Relation between Willingness to Accept and Willingness to Pay."
———. 2023. "Econographics." *Journal of Political Economy Microeconomics* 1 (1): 115–61.
- Chapman, Jonathan, Erik Snowberg, Stephanie Wang, and Colin Camerer. 2019. "Loss Attitudes in the U.S. Population: Evidence from Dynamically Optimized Sequential Experimentation (DOSE)."
- Chapman, Jonathan, Erik Snowberg, Stephanie W. Wang, and Colin Camerer. 2022. "Looming Large or Seeming Small? Attitudes towards Losses in a Representative Sample."

- Choi, Syngjoo, Shachar Kariv, Wieland Müller, and Dan Silverman. 2014. "Who Is (More) Rational?" *American Economic Review* 104 (6): 1518–50.
- Chuang, Yating, and Laura Schechter. 2015. "Stability of Experimental and Survey Measures of Risk, Time, and Social Preferences: A Review and Some New Results." *Journal of Development Economics* 117 (November):151–70.
- Cicchetti, Domenic. 1994. "Guidelines, Criteria, and Rules of Thumb for Evaluating Normed and Standardized Assessment Instruments in Psychology." *Psychological Assessment* 6 (4): 284–90.
- Dean, Mark, and Pietro Ortoleva. 2019. "The Empirical Relationship between Non-Standard Economic Behaviors." *Proceedings of the National Academy of Science* 116 (33): 16262–67.
- DellaVigna, Stefano, and Devin Pope. 2022. "Stability of Experimental Results: Forecasts and Evidence." *American Economic Journal: Microeconomics* 14 (3): 889–925.
- Dimmock, Stephen, Roy Kouwenberg, Olivia S. Mitchell, and Kim Peijnenburg. 2016. "Ambiguity Aversion and Household Portfolio Choice Puzzles: Empirical Evidence." *Journal of Financial Economics* 119 (3): 559–77.
- Dohmen, Thomas, Armin Falk, David Huffman, Felix Marklein, and Uwe Sunde. 2009. "Biased Probability Judgment: Evidence of Incidence and Relationship to Economic Outcomes from a Representative Sample." *Journal of Economic Behavior & Organization* 72 (3): 903–15.
- Dohmen, Thomas, Armin Falk, David Huffman, and Uwe Sunde. 2010. "Are Risk Aversion and Impatience Related to Cognitive Ability?" *American Economic Review* 100 (3): 1238–60.
- Dohmen, Thomas, Armin Falk, David Huffman, Uwe Sunde, Jürgen Schupp, and Gert G. Wagner. 2011. "Individual Risk Attitudes: Measurement, Determinants, and Behavioral Consequences." *Journal of the European Economic Association* 9 (3): 522–50.
- Dohmen, Thomas, and Tomáš Jagelka,. forthcoming. "Accounting for Individual-Specific Reliability of Self-Assessed Measures of Economic Preferences and Personality Traits." *Journal of Political Economy Microeconomics*.
- Einav, Liran, Amy Finkelstein, Iuliana Pascu, and Mark R. Cullen. 2012. "How General Are Risk Preferences? Choices under Uncertainty in Different Domains." *American Economic Review* 102 (6): 2606–38.
- Enke, Benjamin, and Cassidy Shubatt. 2024. "Quantifying Lottery Choice Complexity."
- Ericson, Keith Marzilli. 2011. "Forgetting We Forget: Overconfidence and Memory." *Journal of the European Economic Association* 9 (1): 43–60.
- Falk, Armin, Anke Becker, Thomas Dohmen, Benjamin Enke, David Huffman, and Uwe Sunde. 2018. "Global Evidence on Economic Preferences." *The Quarterly Journal of Economics* 133 (4): 1645–92.
- Falk, Armin, Thomas Neuber, and Philipp Strack. 2021. "Limited Self-Knowledge and Survey Response Behavior."
- Fehr, Dietmar, Günther Fink, and B. Kelsey Jack. 2022. "Poor and Rational: Decision-Making under Scarcity." *Journal of Political Economy* 130 (11): 2862–97.
- Fehr, Ernst, and Lorenz Goette. 2007. "Do Workers Work More If Wages Are High? Evidence from a Randomized Field Experiment." *American Economic Review* 97 (1): 298–317.
- Gabaix, Xavier. 2020. "A Behavioral New Keynesian Model." *American Economic Review* 110 (8): 2271–2327.
- Gagnon-Bartsch, Tristan, Matthew Rabin, and Joshua Schwartzstein. 2023. "Channeled Attention and Stable Errors."
- Gillen, Ben, Erik Snowberg, and Leeat Yariv. 2019. "Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study." *Journal of Political Economy* 127 (4): 1826–63.

- Haushofer, Johannes, Prachi Jain, Abednego Musau, and David Ndeti. 2021. "Stress May Increase Choice of Sooner Outcomes, but Not Temporal Discounting." *Journal of Economic Behavior & Organization* 183 (March):377–96.
- Heidhues, Paul, Botond Koszegi, and Philipp Strack. 2023. "Misinterpreting Yourself."
- Jensen, Arthur Robert. 1998. *The g Factor: The Science of Mental Ability*. Praeger.
- Joint Committee, of the American Educational Research Association, American Psychological Association, and the National Council on Measurement in Education. 2014. *Standards for Educational and Psychological Testing*. Washington, DC: American Educational Research Association.
- Kolenikov, Stanislav, and Gustavo Angeles. 2009. "Socioeconomic Status Measurement with Discrete Proxy Variables: Is Principal Component Analysis a Reliable Answer?" *Review of Income and Wealth* 55 (1): 128–65. <https://doi.org/10.1111/j.1475-4991.2008.00309.x>.
- Laibson, David. 2018. "Private Paternalism, the Commitment Puzzle, and Model-Free Equilibrium." *AEA Papers and Proceedings* 108 (May):1–21.
- Laibson, David, Peter Maxted, and Benjamin Moll. forthcoming. "Present Bias Amplifies the Household Balance-Sheet Channels of Macroeconomic Policy." *Quarterly Journal of Economics*.
- Lance, Charles E., Marcus M. Butts, and Lawrence C. Michels. 2006. "The Sources of Four Commonly Reported Cutoff Criteria: What Did They Really Say?" *Organizational Research Methods* 9 (2): 202–20.
- Le Yaouanq, Yves, and Peter Schwardmann. 2022. "Learning about One's Self." *Journal of the European Economic Association* 20 (5): 1791–1828.
- Levy, Matthew, and Joshua Tasoff. 2016. "Exponential-Growth Bias and Lifecycle Consumption." *Journal of the European Economic Association* 14 (3): 545–83.
- Lian, Chen. forthcoming. "Mistakes in Future Consumption, High MPCs Now." *American Economic Review: Insights*.
- Lichand, Guilherme, and Anandi Mani. 2020. "Cognitive Droughts."
- Litwin, Mark S. 2003. *How To Assess and Interpret Survey Psychometrics*. SAGE.
- Lusardi, Annamaria, and Olivia S. Mitchell. 2014. "The Economic Importance of Financial Literacy: Theory and Evidence." *Journal of Economic Literature* 52 (1): 5–44.
- MacLeod, Colin M. 1991. "Half a Century of Research on the Stroop Effect: An Integrative Review." *Psychological Bulletin* 109 (2): 163.
- Mani, Anandi, Sendhil Mullainathan, Eldar Shafir, and Jiaying Zhao. 2013. "Poverty Impedes Cognitive Function." *Science* 341 (6149): 976–80.
- Mata, Rui, Renato Frey, David Richter, Jürgen Schupp, and Ralph Hertwig. 2018. "Risk Preference: A View from Psychology." *Journal of Economic Perspectives* 32 (2): 155–72.
- McArdle, John J., Gwenith G. Fisher, and Kelly M. Kadlec. 2007. "Latent Variable Analyses of Age Trends of Cognition in the Health and Retirement Study, 1992-2004." *Psychology and Aging* 22 (3): 525–45.
- Meier, Armando N. 2022. "Emotions and Risk Attitudes." *American Economic Journal: Applied Economics* 14 (3): 527–58.
- Meier, Stephan, and Charles D. Sprenger. 2015. "Temporal Stability of Time Preferences." *Review of Economics and Statistics* 97 (2): 273–86.
- Miyake, Akira, and Naomi P. Friedman. 2012. "The Nature and Organization of Individual Differences in Executive Functions: Four General Conclusions." *Current Directions in Psychological Science* 21 (1): 8–14.
- Moore, Don A., and Paul J. Healy. 2008. "The Trouble with Overconfidence." *Psychological Review* 115 (2): 502–17.

- Pfauti, Oliver, Fabian Seyrich, and Jonathan Zinman. 2024. "Bad Luck or Bad Decisions? Macroeconomic Implications of Persistent Heterogeneity in Cognitive Skills and Overconfidence."
- Rabin, Matthew, and Georg Weizsäcker. 2009. "Narrow Bracketing and Dominated Choices." *American Economic Review* 99 (4): 1508–43.
- Rammstedt, Beatrice, and Oliver P. John. 2007. "Measuring Personality in One Minute or Less: A 10-Item Short Version of the Big Five Inventory in English and German." *Journal of Research in Personality* 41 (1): 203–12.
- Read, Daniel, and Barbara van Leeuwen. 1998. "Predicting Hunger: The Effects of Appetite and Delay on Choice." *Organizational Behavior and Human Decision Processes* 76 (2): 189–205.
- Schildberg-Hörisch, Hannah. 2018. "Are Risk Preferences Stable?" *Journal of Economic Perspectives* 32 (2): 135–54.
- Solon, Gary, Steven J. Haider, and Jeffrey M. Wooldridge. 2015. "What Are We Weighting For?" *Journal of Human Resources* 50 (2): 301–16.
- Stango, Victor, and Jonathan Zinman. 2009. "Exponential Growth Bias and Household Finance." *Journal of Finance* 64 (6): 2807–49.
- . 2020. "Behavioral Biases Are Temporally Stable." NBER Working Paper No. 27860.
- . 2023. "We Are All Behavioural, More, or Less: A Taxonomy of Consumer Decision-Making." *The Review of Economic Studies* 90 (3): 1470–98.
- Stigler, George, and Gary Becker. 1977. "De Gustibus Non Est Disputandum." *American Economic Review* 67:76–90.
- Tasoff, Joshua, and Wenjie Zhang. 2022. "The Performance of Time-Preference and Risk-Preference Measures in Surveys." *Management Science* 68 (2): 1149–73.
- Vines, David, and Samuel Wills. 2018. "The Rebuilding Macroeconomic Theory Project: An Analytical Assessment." *Oxford Review of Economic Policy* 34 (1–2): 1–42.



This figure plots the point estimates reported in Table 3 Columns 5 and 6. As detailed in the Table 3 notes, level correlations are not identified for Certainty Preference and NBLLN. The bias-level correlation between the level and rank correlation estimates is 0.77 (SE=0.18, N=15).

Table 1. Behavioral bias elicitations, definitions, and measurements

Potential source of bias:	Elicitation method description	Bias indicator(s)	Bias rank(s): unique values
key antecedent(s) for elicitation design			
(1)	(2)	(3)	(4)
Time inconsistent discounting of money: Andreoni & Sprenger (2012), Barcellos & Carvalho (2014)	Convex Time Budget. 24 decisions allocating 100 tokens each between smaller- sooner and larger-later amounts; decisions pose varying start dates (today vs. 5 weeks from today). delay lengths (5 or 9 weeks) & savings yields.	<i>Present-biased: discounts more when sooner date is today</i> <i>Future-biased: discounts more when sooner date is 5 weeks from today</i>	63
Time inconsistent discounting of consumption: Read & van Leeuwen (1998), Barcellos & Carvalho (2014)	Two decisions between two snacks: healthier/less-delicious vs. less healthy/more delicious. Decisions vary only in date snack is delivered: now, or 5 weeks from now.	<i>Present-biased: choose less healthy now, healthy 5 weeks from now</i> <i>Future-biased: choose healthy now, less healthy 5 weeks from now</i>	3
Violates General Axiom of Revealed Preference (GARP) and dominance avoidance: Choi et al (2014)	Decisions from 11 different linear budget constraints under risk. Subjects choose a point on the line, and then the computer randomly chooses whether to pay the point value of the x-axis or the y-axis.	<i>Violates GARP: potential earnings wasted per CCEI>0</i> <i>Violates GARP and dominance avoidance: potential earnings wasted per combined-CCEI>0</i>	54 97
Narrow bracketing: Rabin & Weizsacker (2009)	Two tasks of two decisions each. Each decision presents the subject with a choice between a certain payoff and a gamble. Each decision pair appears on the same screen. with an instruction to consider the two decisions jointly.	<i>Narrow-bracketing: making a choice that is dominated given implications of an earlier decision, on one or both tasks.</i>	5
Preference for (un)certainity Callen et al (2014)	2 screens of 10 choices each between two lotteries, one a (p, 1-p) gamble over X and Y > X, (p; X, Y), the other a (q, 1-q) gamble over Y and 0, (q; Y, 0). Y=\$450. X=\$150. a €0.1. 1.01. n=0.5 on one screen and 1.0 on the other.	<i>Preference for certainty: certainty premium (CP) >0</i> <i>Cumulative prospect theory: certainty premium (CP)<0</i>	31
Loss aversion/small-stakes risk aversion: Fehr & Goette (2007)	Two choices. Choice 1: between a 50-50 lottery (win \$80 or lose \$50), and \$0. Choice 2: between playing the lottery in Choice 1 six times, and \$0.	<i>Loss aversion: choosing the \$0 payoff in one or more choices.</i>	4
Ambiguity aversion: Dimmock et al. (2016)	Two questions re: a game where win \$500 if pick green ball. 1. Choose between bag with 45 green-55 yellow and bag with unknown mix. 2. If chose 45-55 bag, how many green balls in 45-55 bag would induce switch.	<i>Ambiguity Aversion: prefers bag with 45 green to bag with unknown mix.</i>	17
Over- or under-confidence in performance: Larrick et al (2007), Moore & Healy (2008)	"How many of the last 3 questions (the ones on the disease, the lottery and the savings account) do you think you got correct?"	<i>Overconfidence in perform: self-assessment > actual score</i> <i>Underconfidence in perform: self-assessment < actual score</i>	6
Overconfidence in precision: Larrick et al (2007), Moore & Healy (2008)	Questions about about likelihoods of different numeracy quiz scores and future income increases.	<i>Overconfidence in precision: responds 100% to one or both questions</i>	3
Overconfidence in relative performance: Larrick et al (2007), Moore & Healy (2008)	"... what you think about your intelligence as it would be measured by a standard test. How do you think your performance would rank, relative to all of the other ALP members who have taken the test?"	<i>Greater diff between self-assessed and actual rank indicates more overconfidence.</i> <i>"Overconfident" = overconfidence above median.</i>	78
Non-belief in the law of large numbers (NBLN): Benjamin, Moore, and Rabin (2013)	Question re: percent chances that, among 1,000 coin flips, the # of heads will fall in ranges [0, 480], [481, 519], and [520, 1000]. NBLN = distance between response for [481, 519] and 78.	<i>Underestimates convergence to 50-50: responds with<78%</i> <i>Overestimates convergence to 50-50: responds with>78%</i>	25
Gambler's fallacies: Benjamin, Moore, and Rabin (2013)	"Imagine that we had a computer "flip" a fair coin... 10 times. The first 9 are all heads. What are the chances, in % terms, that the 10th flip will be a head?"	<i>Gambler's (Cold-hand) fallacy: responds with<50%</i> <i>Hot-hand fallacy: responds with>50%</i>	15
Exponential growth bias (EGB), debt-side: Stango & Zinman (2009; 2011)	Survey first elicits monthly payment respondent would expect to pay on a \$10,000, 48 month car loan (this response defines the actual annual percentage rate). Then elicits perceived APR implied by that payment.	<i>Underestimates EG: actual APR>perceived APR</i> <i>Overestimates EG: actual APR<perceived APR</i>	82
Exponential growth bias (EGB), asset-side: Banks et al (2007)	Elicits perceived future value (FV) of \$200, earning 10% annual, after two years.	<i>Underestimates EG: perceived FV<actual FV=\$242</i> <i>Overestimates EG: perceived FV>actual FV=\$242</i>	17
Limited attention: Author-developed	Four questions re: whether subject's finances would improve with more attention given the opportunity cost of attention, with questions varying the types of decisions: day-to-day, medium-run, long-run, or choosing financial products/services.	<i>Limited attention: Indicates regret about paying too little attention given opportunity cost of attention, on one or more of the four questions</i>	5
Limited prospective memory: Ericson (2011)	"The ALP will offer you the opportunity to earn an extra \$10.... This special survey has just a few simple questions but will only be open for 24 hours, starting 24 hours from now.... please tell us now whether you expect to do this special survey."	<i>Limited memory: Says will complete task but does not complete.</i>	2

The Data Appendix Section 1 provides additional details on measuring each behavioral bias. "CCEI" = Critical Cost Efficiency Index. "ALP"=American Life Panel. Unique values is the maximum across our two survey rounds.

Table 2. Behavioral bias stability in the aggregate: Prevalence estimates from the same sample, 3 years apart

Panel A. Each (Directional) Bias	Prevalence (unweighted)		Prevalence (weighted)		p-value: (3) = (4)	N	
	Round 1	Round 2	Round 1	Round 2		Round 1	Round 2
Potential source of bias and/or Bias indicator	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Time-inconsistent discounting money: <i>Present-biased</i>	0.26	0.25	0.24	0.26	0.71	803	819
Time-inconsistent discounting money: <i>Future-biased</i>	0.37	0.37	0.35	0.39	0.33		
Time-inconsistent discounting consumption: <i>Present-biased</i>	0.14	0.17	0.14	0.15	0.56	835	829
Time-inconsistent discounting consumption: <i>Future-biased</i>	0.07	0.06	0.07	0.08	0.68		
<i>Violates GARP (based on CCEI)</i>	0.50	0.50	0.48	0.48	0.88	774	799
<i>Violates GARP (with dominance avoidance)</i>	0.95	0.96	0.94	0.96	0.30		
<i>Narrow-brackets</i>	0.59	0.62	0.59	0.63	0.18	827	837
Preference re certainty: <i>Certainty premium > 0</i>	0.77	0.76	0.75	0.77	0.70	620	620
Preference re certainty: <i>Certainty premium < 0</i>	0.23	0.24	0.25	0.23	0.70		
<i>Loss-averse</i>	0.64	0.60	0.65	0.56	0.00	843	845
<i>Ambiguity-averse</i>	0.74	0.76	0.73	0.73	0.93	842	821
Mis-confidence in level performance: <i>Overconfident</i>	0.36	0.34	0.37	0.39	0.49	829	813
Mis-confidence in level performance: <i>Underconfident</i>	0.12	0.13	0.11	0.14	0.23		
<i>Overconfident in precision</i>	0.45	0.45	0.43	0.39	0.23	793	775
<i>Overconfident in relative performance</i>	N/A	N/A	N/A	N/A	N/A	844	818
Non-belief in the law of large numbers: <i>Underestimates convergence</i>	0.85	0.85	0.85	0.89	0.11	833	819
Non-belief in the law of large numbers: <i>Overestimates convergence</i>	0.15	0.15	0.15	0.11	0.11		
Gambler's fallacy: <i>Cold Hand</i>	0.26	0.24	0.24	0.23	0.63	842	817
Gambler's fallacy: <i>Hot Hand</i>	0.12	0.10	0.11	0.11	0.93		
Exponential growth bias, loan-side: <i>Underestimates APR</i>	0.51	0.53	0.52	0.54	0.54	778	783
Exponential growth bias, loan-side: <i>Overestimates APR</i>	0.34	0.34	0.35	0.34	0.86		
Exponential growth bias, asset-side: <i>Underestimates future value</i>	0.44	0.41	0.45	0.45	0.88	761	735
Exponential growth bias, asset-side: <i>Overestimates future value</i>	0.08	0.06	0.09	0.09	0.83		
<i>Limited attention</i>	0.45	0.44	0.49	0.49	0.93	832	829
<i>Limited memory</i>	0.82	0.76	0.83	0.74	0.00	825	803
Panel B. Across biases							
Mean	0.43	0.42	0.42	0.42	0.53		
Mean difference across rounds	0.00			0.00			
Mean difference across rounds	0.02			0.02			
Max difference across rounds	0.07			0.09			

Columns 3 and 4 use the ALP's sampling weights. Our full sample is 845 respondents, with the lower sample sizes for bias-round measures in Columns 6 and 7 due mostly to item non-response (although the certainty premium can be calculated only for respondents without multiple switch points). Biases measured using the same elicitation have the same sample size within-round. Relative overconfidence level is not identified in our elicitation. Please see Table 1 for summary of bias measures, and the Data Appendix Section 1 for details.

Table 3. Behavioral bias temporal stability within-person.

	Stable		Unstable		Correlations (standard error)			
	Share biased => biased	Share unbiased => unbiased	Share biased => unbiased	Share unbiased => biased	Level	Rank	N	
Panel A. Each Bias	(1)	(2)	(3)	(4)	(5)	(6)	(7)	
Time-inconsistent discounting money	0.41	0.19	0.17	0.23	0.23 (0.06)	0.01 (0.03)	782	
Time-inconsistent discounting consumption	0.07	0.63	0.13	0.16	0.25 (0.06)	0.19 (0.05)	821	
GARP violation (based on CCEI)	0.27	0.30	0.24	0.20	0.20 (0.06)	0.17 (0.03)	739	
GARP violation (with dominance avoidance)	0.91	0.01	0.03	0.05	0.32 (0.13)	0.32 (0.02)		
Narrow-bracketing	0.45	0.22	0.15	0.18	0.47 (0.05)	0.34 (0.04)	820	
Preference re certainty	unidentified	unidentified	unidentified	unidentified	unidentified	0.23 (0.03)	514	
Loss-aversion	0.44	0.25	0.19	0.12	0.55 (0.04)	0.49 (0.03)	843	
Ambiguity-aversion	0.57	0.13	0.13	0.16	0.44 (0.05)	0.27 (0.03)	818	
Mis-confidence in level performance	0.33	0.29	0.18	0.19	0.4 (0.04)	0.29 (0.04)	799	
Overconfidence in precision	0.27	0.46	0.15	0.12	0.66 (0.04)	0.60 (0.03)	748	
Overconfidence in relative performance	0.37	0.33	0.15	0.15	0.58 (0.04)	0.55 (0.02)	817	
Non-belief in the law of large #s	0.99	0.00	0.00	0.00	unidentified	0.31 (0.02)	809	
Gambler's fallacy	0.22	0.51	0.15	0.12	0.61 (0.04)	0.23 (0.02)	815	
Exponential growth bias, loan-side	0.78	0.03	0.09	0.10	0.32 (0.08)	0.29 (0.02)	737	
Exponential growth bias, asset-side	0.45	0.34	0.14	0.08	0.79 (0.03)	0.42 (0.02)	688	
Limited attention	0.33	0.33	0.18	0.15	0.49 (0.05)	0.45 (0.03)	819	
Limited memory	0.62	0.05	0.21	0.11	0.10 (0.07)	0.10 (0.07)	788	
Panel B. Across biases	Mean	0.47	0.25	0.14	0.13	0.43	0.31	772
	Mean stable, unstable	0.72		0.28				
	Min	0.57		0.00	0.10	0.01		
	Max	0.99		0.44	0.79	0.60		

All results here use the ALP sampling weights. Columns 1-4 present the estimated sample shares for each cell in the 2x2 transition matrix for the extensive margin of each bias: (biased or unbiased in Round 1) x (biased or unbiased three years later). SEs for these proportions are each 0.02 or lower and we do not report them above to avoid cluttering the table. Column 5 estimates the serial correlation in the bias indicator across rounds, using the tetrachoric estimator. Level stability is unidentified for certainty premium because the elicitation does not allow for perfectly unbiased responses. Level stability is technically unidentified for NBLLN with the tetrachoric estimator because almost everyone exhibits at least a little bit of bias in both rounds (see Table 4 for evidence on the distribution of changes); if we assume this correlation is 1, the cross-bias mean tetrachoric correlation increases from 0.43 to 0.46. Lacking an absolute measure of overconfidence in relative performance, we define its bias indicator as confidence > median. Rank correlations in Column 6 estimated using the polychoric estimator, which optimally chooses tetrachoric or Pearson based on the degree of smoothness in the measure. Please see Table 1 for summary of bias measures, and the Data Appendix Section 1 for details.

Table 4. Behavioral bias temporal stability within-person: Granular data descriptions

	Cross-round change in bias category		Number of bias categories	% change in bias categories at median
	Proport. with none	median (sd)		
Panel A. Each bias	(1)	(2)	(3)	(4)
<i>Time-inconsistent discounting money</i>	0.11	11 (13)	63	0.17
Time-inconsistent discounting consumption	0.69	0 (1)	3	0.00
<i>GARP violation (based on CCEI)</i>	0.39	12 (15)	54	0.22
<i>GARP violation (with dominance avoidance)</i>	0.15	17 (22)	97	0.18
Narrow-bracketing	0.38	1 (1)	5	0.20
<i>Preference re certainty</i>	0.43	6 (5)	31	0.19
Loss-aversion	0.51	0 (1)	4	0.00
<i>Ambiguity-aversion</i>	0.37	3 (3)	17	0.18
Mis-confidence in level performance	0.51	1 (1)	6	0.17
Overconfidence in precision	0.69	0 (1)	3	0.00
<i>Overconfidence in relative performance</i>	0.07	12 (12)	78	0.15
<i>Non-belief in the law of large #s</i>	0.14	4 (5)	25	0.16
<i>Gambler's fallacy</i>	0.05	3 (2)	17	0.18
<i>Exponential growth bias, loan-side</i>	0.15	13 (17)	82	0.16
Exponential growth bias, asset-side	0.26	2 (2)	17	0.13
Limited attention	0.47	1 (1)	5	0.20
Limited memory	0.69	0 (0)	2	0.00
Panel B. Cross-bias means				
All biases	0.36	5	30	0.13
<i>Biases with >=median # categories</i>	0.23	9	52	0.18

Bias measures with >=median number of categories are italicized. Column 3 reproduces Table 1 Column 4, for reference. Please see Table 1 for summary of bias measures, and the Data Appendix Section 1 for details. All estimates here use the ALP sampling weights.

Table 5. Within-person stability: Biases & other (classically-measured) characteristics

	Rank correlation	N
Single behavioral biases	Table 3 Column (6)	Table 3 Column (7)
Mean correlation across biases	0.31	Table 3 Column (7)
Mean correlation of behavioral common factors	0.46	845
<i>Patience</i>	0.31	782
<i>Risk-aversion: Lifetime income gamble</i>	0.28	821
<i>Risk-aversion: Financial risk-taking scale</i>	0.59	835
<i>Cognitive skills: Number series</i>	0.74	819
<i>Cognitive skills: Numeracy</i>	0.56	802
<i>Cognitive skills: Financial literacy</i>	0.78	821
<i>Cognitive skills: Stroop</i>	0.48	773

Behavioral common factor estimate here is the unweighted mean of the 8 point estimates in Table 7. Other characteristics are italicized (please see Data Appendix Section 2 for definition and measurement details). Their correlations estimated using Pearson, within-person across our two rounds of data collection 3 years apart. P-values all <0.01. We lack temporal stability estimates for personality traits in our data because we only measured them in Round 2, but Anusic and Schimmack's (2016) meta-analysis of personality measures finds a precisely estimated average 3-year rank correlation of about 0.6. All estimates here use the ALP sampling weights.

Table 6. Discarding panelists with fast response times, low conscientiousness, low survey interest or low cognitive skills does not change estimated bias rank stability.

	Full sample (1)	Discard fast responses (2)	Discard low conscientious (3)	Discard survey uninteresting (4)	Discard low number series (5)	Discard low stroop (6)	Diff: Full - mean(sub- samples) (7)
Panel A. Each Bias							
Time-inconsistent discounting money	0.01	0.01	-0.06	0.03	0.03	-0.04	0.02
Time-inconsistent discounting consumption	0.19	0.24	0.14	0.25	0.17	0.22	-0.01
GARP violation (based on CCEI)	0.17	0.22	0.19	0.15	0.13	0.18	0.00
GARP violation (with dominance avoidance)	0.32	0.36	0.30	0.31	0.35	0.37	-0.01
Narrow-bracketing	0.34	0.33	0.33	0.34	0.42	0.38	-0.02
Preference re certainty	0.23	0.20	0.16	0.23	0.26	0.23	0.01
Loss-aversion	0.49	0.48	0.51	0.48	0.56	0.52	-0.02
Ambiguity-aversion	0.27	0.22	0.33	0.28	0.34	0.31	-0.03
Mis-confidence in level performance	0.29	0.34	0.33	0.33	0.24	0.18	0.01
Overconfidence in precision	0.60	0.55	0.62	0.65	0.61	0.59	0.00
Overconfidence in relative performance	0.55	0.58	0.56	0.59	0.51	0.55	-0.01
Non-belief in the law of large #s	0.31	0.30	0.27	0.29	0.32	0.27	0.02
Gambler's fallacy	0.23	0.26	0.24	0.21	0.25	0.25	-0.01
Exponential growth bias, loan-side	0.29	0.26	0.31	0.31	0.26	0.26	0.01
Exponential growth bias, asset-side	0.42	0.43	0.46	0.42	0.36	0.44	0.00
Limited attention	0.45	0.52	0.47	0.45	0.44	0.47	-0.02
Limited memory	0.10	-0.01	0.17	0.14	0.16	0.16	-0.02
Panel B. Across Biases							
Max N	843	591	658	751	704	713	
Min	0.01	-0.01	-0.06	0.03	0.03	-0.04	-0.03
Max	0.60	0.58	0.62	0.65	0.61	0.59	0.02
Mean	0.31	0.31	0.31	0.32	0.32	0.31	-0.01

Column 1 reproduces the full sample Pearson or tetrachoric rank correlation estimates from Table 3, Column 6. "Fast responses" are those in the quickest two deciles, by total time spent on all tasks related to behavioral biases; we discard panelists whose responses in both rounds fall into the quickest two deciles. "Low conscientiousness" discards roughly the lowest quartile (22%) of responses. "Survey uninteresting" discards panelists who report the survey as less than "interesting" or very "interesting" in more than one survey; the criterion discards 5% of observations. "Discard low number series" drops all panelists with a number series score below roughly the 10th percentile in rounds 1 and 2. "Discard low Stroop" drops all panelists below the 10th percentile in rounds 1 and 2. Column 7 reports the difference between the full sample point estimate in Column 1 and the mean of the sub-sample point estimates in Columns 2-6. Please see Table 1 for summary of bias measures, and the Data Appendix Section 1 for details. Data Appendix Section 2-B contains details on the number series and Stroop elicitations and measures. Conscientiousness measured using its two items in the 10-item version of the Big Five inventory (Rammstedt and John 2007). All estimates here use the ALP sampling weights.

Table 7. Clusters of related biases are relatively stable

Behavioral Common Factor ("BCF")	Within-person temporal stability of BCF scores:	
	Correlation estimate	
	(1)	(2)
BCF 1: Beliefs	0.65 (0.01)	0.64 (0.01)
BCF 2: Decision quality	0.55 (0.01)	0.54 (0.01)
BCF 3: Time discounting	0.22 (0.02)	0.25 (0.02)
BCF 4: Risk/uncertainty	0.48 (0.01)	0.32 (0.02)
	N	
	845	845

Pearson correlations, with standard errors in parentheses. Factor scores are standardized, measured at the person-round level, and each capture a common factor (CF) underlying a set of behavioral biases that are inter-correlated. BCF 1 captures overconfidence and mathematical operations biases; BCF 2 captures inconsistent/dominated choices and attention/memory; BCF 3 captures present-bias on consumption and dollars as well as classical discounting; and BCF 4 captures the risk-related biases as well as classical risk aversion. Column 1 weights each bias/characteristic in each common factor as in Stango and Zinman (2023 RESTUD, Table 5 and Section 5.3). Column 2 excludes the measures of classical discounting and risk aversion from the common factors by assigning them zero weights in the factor score. All estimates here use the ALP sampling weights.

Table 8. Limiting the sample to panelists with more stable household structure or economic status does not change estimated bias rank stability.

	Full sample (N=845)	Households that are stable across rounds						Difference: Full-mean (sub- samples)	
		Objective measures				Subjective well-being			
		HH members (N=613)	Work status (N=781)	Income cat. (N=409)	Fin. distress (N=658)	Financial (N=703)	Life satis. (N=464)		
Panel A. Each Bias	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	
Time-inconsistent discounting money	0.01	0.02	0.03	0.06	-0.02	0.04	-0.03	-0.01	
Time-inconsistent discounting consumption	0.19	0.08	0.20	0.02	0.33	0.18	0.22	0.02	
GARP violation (based on CCEI)	0.17	0.16	0.18	0.20	0.14	0.22	0.15	-0.01	
GARP violation (with dominance avoidance)	0.32	0.29	0.32	0.20	0.31	0.34	0.27	0.03	
Narrow-bracketing	0.34	0.33	0.31	0.25	0.34	0.31	0.25	0.04	
Preference re certainty	0.23	0.29	0.22	0.30	0.27	0.24	0.20	-0.02	
Loss-aversion	0.49	0.41	0.49	0.44	0.49	0.44	0.38	0.05	
Ambiguity-aversion	0.27	0.09	0.28	0.25	0.29	0.29	0.24	0.03	
Mis-confidence in level performance	0.29	0.27	0.29	0.31	0.24	0.30	0.29	0.01	
Overconfidence in precision	0.60	0.56	0.61	0.61	0.70	0.65	0.63	-0.03	
Overconfidence in relative performance	0.55	0.54	0.54	0.53	0.55	0.57	0.52	0.01	
Non-belief in the law of large #s	0.31	0.29	0.33	0.29	0.34	0.33	0.34	-0.01	
Gambler's fallacy	0.23	0.21	0.22	0.22	0.18	0.23	0.14	0.03	
Exponential growth bias, loan-side	0.29	0.29	0.27	0.34	0.23	0.37	0.29	-0.01	
Exponential growth bias, asset-side	0.42	0.49	0.42	0.44	0.40	0.39	0.42	-0.01	
Limited attention	0.45	0.49	0.49	0.51	0.39	0.44	0.50	-0.02	
Limited memory	0.10	0.15	0.09	0.24	0.12	0.04	0.08	-0.02	
Panel B. Across Biases									
	Min	0.01	0.02	0.03	0.02	-0.02	0.04	-0.03	-0.03
	Max	0.60	0.56	0.61	0.61	0.70	0.65	0.63	0.05
	Mean	0.31	0.29	0.31	0.31	0.31	0.32	0.29	0.01

Column 1 reproduces the full sample Pearson or tetrachoric rank correlation estimates from Table 3, Column 6. Column 2 includes only respondents with the same number of household members in each round. Column 3 includes only those with the same employment status in each round. Column 4 includes only those with the same income category in both rounds (the ALP's measure of income is categorical: 17 bins.). Column 5 includes only those with same value of our severe financial distress indicator-- indicating ≥ 1 of 4 measures of severe measures of financial distress-- in both rounds. Column 6 includes only those whose subjective financial well-being changes by one or fewer of four categories across the two rounds. Column 7 includes only those whose subjective life satisfaction changed by 10 points or fewer on a 100-point scale across the two rounds. Column 8 reports the difference between the full sample point estimate in Column 1 and the mean of the sub-sample point estimates in Columns 2-7. Please see Data Appendix Section 1 for additional details on bias measures. All estimates here use the ALP sampling weights.

Appendix Table 1. Bi-directional bias stability on the extensive margins
(building on Table 3, Columns 1-4, here separately considering the two potential directions for bi-directional biases)

	Stable		Unstable			(Share same bias)/ (Share w/that bias)
	Share biased => same bias	Share unbiased => unbiased	Share biased => unbiased	Share biased => other bias	Share unbiased => biased	
Panel A. Potential source of bias: <i>directional bias indicator</i>	(1)	(2)	(3)	(4)	(5)	(6)
Time-inconsistent discounting money: <i>Present-biased</i>	0.06		0.08	0.10	0.09	0.25
Time-inconsistent discounting money: <i>Future-biased</i>	0.16	0.19	0.10	0.10	0.09	0.44
Time-inconsistent discounting consumption: <i>Present-biased</i>	0.04		0.09	0.00	0.09	0.31
Time-inconsistent discounting consumption: <i>Future-biased</i>	0.01	0.63	0.05	0.02	0.07	0.13
Preference re certainty: <i>Certainty premium > 0</i>	0.61		unidentified	0.16	unidentified	
Preference re certainty: <i>Certainty premium < 0</i>	0.07	unidentified	unidentified	0.16	unidentified	
Mis-confidence in level performance: <i>Overconfident</i>	0.22		0.15	0.04	0.13	0.54
Mis-confidence in level performance: <i>Underconfident</i>	0.04	0.29	0.03	0.04	0.06	0.36
Non-belief in the law of large numbers: <i>Underestimates convergence</i>	0.80		0.00	0.08	0.00	0.91
Non-belief in the law of large numbers: <i>Overestimates convergence</i>	0.04	0.00	0.00	0.09	0.00	0.31
Gambler's fallacy: <i>Cold Hand</i>	0.10		0.10	0.04	0.09	0.42
Gambler's fallacy: <i>Hot Hand</i>	0.04	0.51	0.05	0.04	0.03	0.31
Exponential growth bias, loan-side: <i>Underestimates APR</i>	0.36		0.04	0.13	0.04	0.68
Exponential growth bias, loan-side: <i>Overestimates APR</i>	0.13	0.03	0.05	0.16	0.05	0.38
Exponential growth bias, asset-side: <i>Underestimates future value</i>	0.31		0.12	0.02	0.07	0.69
Exponential growth bias, asset-side: <i>Overestimates future value</i>	0.05	0.34	0.02	0.06	0.02	0.38
Panel B. Across biases and directions						
Cross-bias mean, standard direction						0.56
Cross-bias mean, non-standard direction						0.33

Columns 1-5 present the sample shares for each cell in the transition matrix for the extensive margin of each bidirectional bias. Please see Table 1 for summary of bias measures, and the Data Appendix Section 1 for details on measurement and standard vs. non-standard direction designations. Column 6 compares stability across different directions by conditioning on directional prevalence (i.e., by scaling by directional prevalence in the denominator). All estimates here use the ALP sampling weights.

Appendix Table 2. Behavioral bias temporal stability within-person: Unweighted estimates of extensive margin persistence and rank correlation (same as Table 3, but unweighted)

	Stable		Unstable		Correlations (standard error)		
	Share biased => biased	Share unbiased => unbiased	Share biased => unbiased	Share unbiased => biased	Level	Rank	N
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Panel A. Each Bias							
Time-inconsistent discounting money	0.43	0.19	0.19	0.18	0.34 (0.05)	0.01 (0.03)	782
Time-inconsistent discounting consumption	0.07	0.62	0.15	0.16	0.22 (0.06)	0.18 (0.06)	821
GARP violation (based on CCEI)	0.29	0.29	0.21	0.21	0.25 (0.06)	0.20 (0.03)	739
GARP violation (with dominance avoidance)	0.92	0.01	0.03	0.04	0.45 (0.06)	0.29 (0.03)	
Narrow-bracketing	0.43	0.22	0.16	0.19	0.42 (0.05)	0.33 (0.04)	820
Preference re certainty	unidentified	unidentified	unidentified	unidentified	unidentified	0.20 (0.03)	514
Loss-aversion	0.46	0.22	0.18	0.14	0.51 (0.05)	0.47 (0.04)	843
Ambiguity-aversion	0.61	0.11	0.13	0.15	0.42 (0.05)	0.24 (0.03)	818
Mis-confidence in level performance	0.30	0.35	0.18	0.18	0.44 (0.05)	0.35 (0.04)	799
Overconfidence in precision	0.31	0.41	0.14	0.18	0.64 (0.04)	0.59 (0.03)	748
Overconfidence in relative performance	0.34	0.37	0.13	0.14	0.62 (0.04)	0.57 (0.02)	817
Non-belief in the law of large #s	0.99	0.00	0.00	0.00	unidentified	0.32 (0.02)	809
Gambler's fallacy	0.23	0.52	0.14	0.11	0.67 (0.04)	0.33 (0.02)	815
Exponential growth bias, loan-side	0.76	0.04	0.09	0.11	0.35 (0.08)	0.20 (0.03)	737
Exponential growth bias, asset-side	0.37	0.42	0.13	0.09	0.78 (0.03)	0.47 (0.04)	688
Limited attention	0.29	0.40	0.16	0.15	0.53 (0.04)	0.50 (0.04)	819
Limited memory	0.63	0.06	0.19	0.12	0.17 (0.07)	0.18 (0.07)	788
Panel B. Across biases							
Mean	0.46	0.26	0.14	0.13	0.45	0.32	772
Mean stable, unstable		0.73		0.27			
Min		0.58		0.00	0.17	0.01	
Max		0.99		0.42	0.78	0.59	

Columns 1-4 present the estimated sample shares for each cell in the 2x2 transition matrix for the extensive margin of each bias: (biased or unbiased in Round 1) x (biased or unbiased three years later). SEs for these proportions are each 0.02 or lower and we do not report them above to avoid cluttering the table. Column 5 estimates the serial correlation in the bias indicator across rounds, using the tetrachoric estimator. Level stability is unidentified for certainty premium because the elicitation does not allow for perfectly unbiased responses. Level stability is technically unidentified for NBLLN with the tetrachoric estimator because almost everyone exhibits at least a little bit of bias in both rounds (see Table 4 for evidence on the distribution of changes); if we assume this correlation is 1, the cross-bias mean tetrachoric correlation increases from 0.45 to 0.49. Lacking an absolute measure of overconfidence in relative performance, we define its bias indicator as confidence>median. Rank correlations in Column 6 estimated using the polychoric estimator, which optimally chooses tetrachoric or Pearson based on the degree of smoothness in the measure. Please see Table 1 for summary of bias measures, and the Data Appendix Section 1 for details.