We are all behavioral, more or less:
Measuring and using consumer-level behavioral summary statistics

Victor Stango and Jonathan Zinman[*]

July 2019

Can a behavioral summary statistic empirically capture cross-consumer variation in behavioral tendencies and help identify whether behavioral biases, taken together, are linked to material consumer welfare losses? Our answer is yes. We construct simple consumer-level behavioral summary statistics—"B-counts"—by eliciting seventeen potential behavioral biases per person, in a nationally representative panel, in two separate rounds nearly three years apart. B-counts aggregate information on behavioral biases within-person. Nearly all consumers exhibit multiple biases, with substantial variation across people. B-counts are relatively stable within-consumer over time, and that stability helps to address measurement error when using B-counts to model the relationship between biases and consumer welfare. Conditional on classical inputs—risk aversion and patience, life-cycle factors and other demographics, cognitive and non-cognitive skills, and financial resources—B-counts strongly negatively correlate with both objective and subjective aspects of utility. The results hold in much lower-dimensional models with fewer covariates and/or B-counts based on a handful of biases, illuminating lower-cost ways to use behavioral summary statistics to help capture the combined influence of multiple behavioral biases for a wide range of research questions and applications.

JEL Nos. C83, D1, D6, D9, E7, G4

[*] 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 and Josh Schwartzstein for many conversations; Hunt Allcott, Raj Chetty, Xavier Gabaix, and Dmitry Taubinsky for guidance on modeling using behavioral summary statistics; Keith Ericson and Erik Snowberg for detailed comments; and many conference and seminar participants (especially the NBER Public Economics group) for comments on this and related papers. Special thanks to Joanne Yoong for collaboration on the Round 1 survey design and implementation.

> *"A common criticism of behavioral economics is that it does not offer a single unified framework as an alternative to the neoclassical model."* (Chetty 2015, p. 25)

> *"… behavioral models… do not integrate well with basic microeconomic theory because they do not develop a general procedure for the basic economic operation of simplifying reality and acting using that simplified model."* (Gabaix 2014, p. 1662)

Behavioral social scientists have identified myriad biases in decision making that could reduce consumer welfare. But it is challenging to capture myriad influences in a portable model—a "general procedure" per Gabaix, or a "single unified framework" per Chetty.

Some researchers are responding to this dimensionality challenge with models where one or two summary or sufficient statistics capture consumer-level behavioral tendencies and their links to decisions/outcomes.[1] But those models lack empirical validation: most empirical work in behavioral economics examines only one or two biases at a time and hence is silent on how to capture the many potential biases.[2]

An empirically useful consumer-level behavioral summary statistic has many applications. It could help address currently-unexplained heterogeneity in various outcome domains.[3] Other applications are theory- and welfare-driven. Several recent papers show that if a summary statistic captures cross-consumer behavioral decision making tendencies, then that statistic becomes a powerful input for intervention design and welfare analysis.[4] Other work finds that behavioral

[1] For reviews see, e.g., the reduced-form sufficient statistic models in Chetty (2009, 2015) and Mullainathan, Schwartzstein, and Congdon (2012), and the sparsity/behavioral inattention models in Gabaix (2019). We refer to "summary" statistics to span these classes of models.

[2] Chapman et al. (2018a, 2018b), Dean and Ortoleva (2018), and Gillen et al.(forthcoming) also measure a large set of potential behavioral biases per person and examine relationships among biases. But they do not link their biases to field outcomes or develop behavioral summary statistics. The only papers we know of that measure consumer-level summary behavioral statistics, and potential heterogeneity therein, are Taubinsky and Rees-Jones (2018) and Allcott, Lockwood, and Taubinsky (forthcoming). Those papers, while seminal in both theory and empirical implementation, focus on a much smaller set of potential biases and their implications for a much narrower set of decisions.

[3] To take one example, much of the cross-sectional distribution of wealth remains unexplained, despite several decades of work trying to identify its determinants (Poterba 2014; Campbell 2016).

[4] See especially Allcott, Lockwood, and Taubinsky; Allcott and Taubinsky (2015); Baicker, Mullainathan, and Schwartzstein (2015); Farhi and Gabaix (2018); and Taubinsky and Rees-Jones; as we detail below, all of these models account for behavioral heterogeneity that our evidence suggests is empirically important.

heterogeneity *per se* can lead to social welfare loss, from product misallocation, exceeding that from the average level of bias in the population (Taubinsky and Rees-Jones 2018). Policymakers increasingly formulate high-level strategy and specific regulations based on assumptions about how multiple biases, taken together, are prevalent and affect decision-making.[5] The "nudge units" proliferating in public, private, and nonprofit sectors also presume that multiple behavioral biases combine to affect behavior.[6]

We develop empirical behavioral summary statistics ("B-counts") that permit examination of the key assumptions maintained by such models and interventions. B-counts can be used to help identify welfare consequences of multiple behavioral biases, diagnose when intervention is warranted, guide policy development, and evaluate interventions and their targeting.

B-counts are consumer-level statistics that capture behavioral tendencies by aggregating information on multiple biases within-person. We adapt standard lab-style elicitation methods to measure 17 potential sources of behavioral biases per consumer (Table 1), in about 30 minutes of online survey/task time.[7] We then administer those elicitations twice, in two separate rounds of data collection about three years apart, to the same representative sample of 845 U.S. consumers from the American Life Panel. We also collect rich data on outcomes (various measures of objective and subjective well-being in financial and other domains), "classical decision inputs" (cognitive and non-cognitive skills, presumed-classical preferences, life-cycle factors and other demographics), and survey effort.

B-counts summarize behavioral tendencies at the consumer-level, simply: they count how many biases (deviations from the classical benchmark) a consumer exhibits. In addition to a "Full" B-count based on all 17 potential biases we measure, we construct a number of "B-sub-counts" motivated by other classes of models, including: "Sparsity B-counts" focusing on limited attention/memory and related phenomena following Gabaix; a "Preference B-count" that includes inconsistency with revealed preference, loss aversion, etc., but not biased beliefs or problem-

[5] Recent examples include the Department of Energy, Consumer Financial Protection Bureau, and the SEC in the U.S., the Financial Conduct Authority in the U.K., the World Bank, and the United Nations.
[6] See, e.g. Afif et al. (2018) and Guntner et al. (2019).
[7] We chose the 17 based on prior work linking biases to consumer decisions, particularly in the financial domain, and on practical considerations and constraints. See Section 3-A for discussion.

solving approaches; and a “Math B-count” including only exponential growth biases and statistical fallacies.

Empirical validation of B-counts requires that they usefully describe consumer heterogeneity in behavioral tendencies, and that they explain economically meaningful variation in consumer welfare. We show that B-counts do both.

First, we show the typical consumer exhibits multiple behavioral biases, but with substantial heterogeneity across people: we are all behavioral, more or less. A consumer at the 10th percentile has a B-count of 7 out of a possible 17 in each of our two survey rounds; the 50th percentile is 10, and the 90th is 13. Critically, several results indicate that B-counts are empirically distinct from other individual characteristics and differences. Equally critically, B-counts are relatively stable within-consumer when measured twice over our three-year horizon. The within-person cross-round correlation in the B-count is 0.44— a high number relative to prior work estimating temporal stability in behavioral biases or presumed-classical preferences. This stability both shows that the B-count is not noise, and helps us address measurement error both informally, when estimating B-count variance (as an input for behavioral sufficient statistic modeling), and formally, when estimating links between B-counts and outcomes.

In practical terms, we show that one can add B-counts to many research designs with simple, quick elicitations. We construct a “Narrow Sparsity” B-count, a subset of biases guided by theory *a la* Gabaix (and models of more haphazard behavioral inattention), which takes less than two minutes to elicit and also correlates strongly with outcomes. Even B-counts constructed using randomly selected subsets of our 17 behavioral biases deliver similar conditional correlations with outcomes to the Full B-count. (This exercise also suggests that no single bias exerts an outsized influence on the results.) Altogether, our results suggest that the B-count and sub-counts can usefully capture consumer-level behavioral heterogeneity.

Second, and fundamentally for summary statistic models, B-counts strongly and negatively correlate with outcomes—with measures of various “aspects” of utility—conditional on our rich sets of covariates. We offer a simple framework, drawing especially on Allcott, Lockwood, and Taubinsky (forthcoming), Benjamin et al. (2014) and Benjamin et al. (2014), that clarifies conditions under which the B-count captures a systematic wedge between the “normative utility” of a classical consumer and the “decision utility” of a behavioral consumer.

Our repeated elicitations help when estimating conditional correlations between outcomes and B-counts, allowing us to instrument for one round's B-count with the other ("standard IV"), or to employ the more efficient Obviously Related Instrumental Variables method (Gillen, Snowberg, and Yariv forthcoming). Those techniques help us address concerns that our estimates are confounded by measurement error, omitted variables, and/or reverse causality.

The negative conditional correlations between outcomes and B-counts are important economically. In our main specifications, a one standard deviation change in a B-count is associated with an estimated 22 to 30% reduction in objective financial condition and a 26% to 43% reduction in subjective financial condition. These magnitude of these correlations equates roughly to moving down multiple deciles in the income distribution (as implied by the conditional correlations on income decile categories); e.g., for objective financial condition, to moving someone from the 3rd to the 1st income decile, or from the 9th to the 5th decile.

The negative conditional correlations between outcomes and B-counts are also statistically robust. Across our nine main specifications the B-count always has a p-value<0.01, and this pattern also holds across different covariate specifications, including ones where we allow for measurement error in classical preferences and cognitive skills as well as in the B-count, add objective financial condition as an additional control when subjective financial condition is the outcome of interest, or drop all other covariates entirely.

We draw outcomes from other ALP surveys as well as our own, and find that the negative conditional correlations hold for different aspects of utility (per Benjamin et al.)— life satisfaction, happiness, and health status—at least when we use the Sparsity B-counts as summary statistics. The results are too imprecise to characterize for non-financial aspects when we use the Full B-count, perhaps because we chose our 17 potential sources of behavioral biases with financial decision making in particular in mind.

Decomposing the Full B-count into sub-counts sheds further light on mechanisms and welfare implications. The Full B-count's correlations with outcomes are driven more robustly by the thirteen non-math biases than by the four math biases that are more arguably reflections of classical math/cognitive skills. The results are not driven by the seven preference biases, which is notable because behavioral preferences are less clearly welfare-reducing than the ten non-preference biases (e.g., biased expectations, price perceptions, limited attention/memory).[8] Expected-direction biases (e.g., present-bias) robustly negatively correlate with outcomes, while non-expected direction biases (e.g. future-bias) do not.

Broadly, our results suggest that B-counts usefully capture heterogeneity in behavioral tendencies, and further show that such heterogeneity correlates meaningfully with consumer welfare losses. Priors can of course guide the choice of which biases to include in a behavioral summary statistic, as it did in our case with our primary (but not exclusive) focus on financial decision making and in the more narrowly-focused applications in Taubinsky and Rees-Jones (2018) and Allcott, Lockwood, and Taubinsky (forthcoming). Our research design provides a simple toolkit of elicitations, behavioral summary statistics, and approaches to dealing with measurement error that should be useful across a range of empirical applications.

In terms of implications for theory and welfare analysis, our results strongly support the foundational presumption of single-parameter behavioral models: multiple biases combine to reduce consumer welfare, in ways that vary across consumers. Sufficient statistic models with behavioral heterogeneity sometimes require sharp identification of normative choices, and we show how one can use our results and tools to check key identifying assumptions and, in cases where those assumptions are likely to hold, use our statistics to help estimate model inputs (e.g., one can use B-counts to help measure two of the three sufficient statistics required by the Taubinsky and Rees-Jones method).[9] In cases where those key identifying assumptions are unlikely to hold or sufficient statistic approaches are infeasible to implement (e.g., when experts or highly effective nudges cannot be used to identify normative choices, or when one lacks sufficiently rich identification of the relevant demand curves), we show how our approach to identifying the behavioral wedge can be a useful alternative.

The next section formalizes our approach and provides a roadmap for the rest of the paper.

[8] For more on the preference vs. non-preference bias distinction see, e.g., Baicker et al (2015); Bernheim and Taubinsky (2018).

[9] Importantly, our results provide substantial reassurance about the primary identification concerns in Allcott, Lockwood, and Taubinsky (forthcoming): confounds from omitted biases or other decision inputs.

1. Conceptual and Empirical Framework

A. Conceptual Framework

Conceptually, the B-count—our consumer-level behavioral summary statistic—is designed to help identify a "behavioral wedge" reflecting the combined effects of multiple behavioral biases. This wedge is key for measuring the welfare loss (if any) from behavioral biases, for diagnosing opportunities to improve welfare with behaviorally-targeted interventions, and for designing and evaluating such interventions.

Many behavioral summary statistic models define the behavioral wedge as the difference between the "normative utility" of an unbiased consumer and the "decision utility" of a biased one.[10] The sufficient statistic approach "does not require specifying the exact behavioral model that describes agents' choices" (Chetty 2015, p. 25), so long as the behavioral sufficient statistics capture the combined effects of multiple behavioral biases and satisfy other assumptions. Gabaix's sparsity and behavioral inattention models are more oriented toward fundamentals but also focus on a behavioral wedge, one that is generated by a single, "psychologically founded" attention cost parameter *m* that "condenses" behavioral tendencies toward simplification, inattention, and disproportionate salience (Gabaix 2014 p. 1662).[11] From that foundation "a large number of behavioral phenomena" can emerge (Gabaix 2019 p. 5), including price misperceptions, statistical fallacies, and time-inconsistent discounting.

In summary statistic models, the mean and/or variance of the behavioral wedge are often sufficient statistics for welfare analysis (Mullainathan, Schwartzstein, and Congdon 2012; Chetty 2015; Baicker, Mullainathan, and Schwartzstein 2015; Farhi and Gabaix 2018; Allcott, Lockwood, and Taubinsky forthcoming). In Taubinsky and Rees-Jones (2018), for example, the mean and variance of marginal consumers' mis-reaction to sales taxes identify the efficiency costs of small taxes, together with the price elasticity of demand. In Gabaix's model, the distribution of the *m* parameter describes consumer heterogeneity in behavioral tendencies and outcomes.

[10] We use the normative vs. decision utility nomenclature, following Allcott, Lockwood, and Taubinsky (forthcoming), because our empirical approach maps most closely into theirs. Another common nomenclature is "decision utility" vs. "experienced utility" (Kahneman, Wakker, and Sarin 1997).

[11] Gabaix uses "sparsity" to mean consumer-level whittling of the set of economic phenomena used for decisionmaking, focusing on the most relevant ones and ignoring the less relevant ones, and consequently incurring welfare costs (while saving attention costs). For example, concentrating on some prices but not others is captured by the sparsity model.

Empirically validating these models involves both documenting behavioral summary statistics' distributional properties and establishing their links to field behavior and outcomes. We do both. We first aggregate information, within-consumer, on a wide and relatively domain-general set of potential behavioral biases; that exercise yields the consumer-level Full B-count, which is our main behavioral summary statistic. The only other papers we know of that directly measure a consumer-level behavioral summary statistic, and potential heterogeneity therein, are "TR-J" (Taubinsky and Rees-Jones 2018)[12] and "ALT" (Allcott, Lockwood, and Taubinsky forthcoming). Those papers, while seminal in both theory and empirical implementation, focus on a much smaller set of potential biases and their implications for a much narrower set of decisions (purchases of small household goods).

B. Empirical Framework

After constructing the B-count and examining its distributional properties, our primary empirics estimate links between B-counts and consumer welfare, consistent with the shared view of summary statistic models that multiple behavioral biases can have reinforcing effects on field behavior and outcomes.

Much of our empirical work uses models of the form:

$$(1)\ Y_i = f(Bcount_i) + g(X_i) + h(Surv_i) + \varepsilon_i$$

Y is an outcome (e.g., saving behavior, or an index of financial condition or of happiness). We focus for the most part on broader outcomes—"aspects" of utility or marginal utility in the parlance of Benjamin and co-authors—such as self-assessed financial condition. We also examine several narrower outcomes (retirement savings; stock market participation); those are closer to the product-market-specific applications in TR-J and ALT.

Our B-counts are constructed from information on up to seventeen potential behavioral biases within-consumer. Each potential bias is measured using a stylized, non-product-specific task (in contrast to the product-specific approaches in TR-J and ALT), although we selected our biases and task frames with some focus on the financial domain. The Full B-count uses information on all 17

[12] As TR-J state, they measure the behavioral wedge indirectly (rather than from more-primitive cognitive/psychological biases), by varying tax salience within-individual: "Instead of defining [the behavioral wedge] in relation to a specific mechanisms, we define it by the behaviour that these mechanisms generate: a difference in willingness to pay depending on the presence of a tax" (p. 2466).

potential biases we elicit, while sub-counts use subsets of the 17 motivated by theory, including Gabaix's sparsity models. X is a vector of classical decision inputs (cognitive skills, life-cycle demographics, wealth when Y is subjective financial condition, etc.), and $Surv$ is a vector of measures of survey effort. i indexes consumers, and although we have a time dimension to our data and use repeated measurement to account for measurement error, we abstract from that for now to focus on our identifying variation, which comes from cross-sectional heterogeneity in the B-count. (Sections 2, 3, and 4 provide details on our approaches to sampling, measurement, and estimation.)

Equation (1) most closely parallels ALT's approach (see their Section III.D),[13] and so their primary identification concerns are instructive. One concern is that "unconfoundedness" fails to hold: that the summary statistic is conditionally correlated with X and ε. One potential confound is measurement error: if the B-count is a noisy measurement of some true summary statistic and correlated with ε, then measurement error can bias the estimated relationship between the summary statistic and outcomes. To address that, we elicit all of our behavioral biases twice, in two separate surveys nearly three years apart. Doing that allows us to use within-consumer stability in the B-count to implement an instrumental variables technique that addresses correlation between the B-count and ε (Section 4-C). Another potential confound is if there are other (non-behavioral) characteristics correlated with the B-count that we fail to measure or measure with confounding error. To address that, we measure a rich set of other covariates X, show that our results are largely invariant to X's specification (including allowing for measurement error in key components, or omitting all X variables entirely, in Section 5-B), and show that rich vector of variables X only weakly explains cross-sectional variation in the B-count (Section 6).

A second set of concerns centers on omitted behavioral biases (ALT, p. 28). We address that by eliciting an unusually rich set of potential biases. And even if one views our bias set as a mis-

[13] As ALT p. 25 states: "The process is to use surveys to elicit proxies of bias [re: nutrition knowledge and self-control with respect to sugar-sweetened beverages in their case; of up to 17 more domain-general biases in our case], estimate the relationship between bias proxies and quantity consumed [of sugar-sweetened beverages in their case; of a utility proxy or a financial product in our case], use that relationship to predict the counterfactual quantity that would be consumed if consumers instead maximized normative utility [ALT rely on the choices of experts in their sample for this; we rely on X], and finally transform the quantity difference into dollar units using the price elasticity [we lack exogenous variation in prices or wealth, which keeps us from doing quantitative welfare analysis here, but discuss some potential extensions in Section 8-B-i]."

measured estimate of some more complete set, our empirical approach accommodates that type of measurement error as well (Section 4-C). We discuss related "index weight" issues, in the context of mapping our approach into consumer welfare analysis, in Section 4-D. And finally, our results are robust to constructing the B-count from randomly selected subsets of biases (Section 7).

Equation (1) can also map into various other models that allow for heterogeneity in a behavioral summary statistic—see, e.g., Farhi and Gabaix equation 3, and Gabaix (2019) equation 54—especially if one grants, as is commonly assumed, that our *Y* variables capture important aspects of utility (Sections 4-A and 5-D). Section 8 details differences and complementarities between our approach to identification and the money-metric approach used in sufficient statistic modeling, including discussion of how to use our tools and results to help empirically assess those models' key identifying assumptions and estimate their key inputs (i.e., their sufficient statistics).

2. Research Design and Data Collection

A. Variables overview

As equation (1) illustrates, we measure four multi-dimensional sets of consumer characteristics. One set includes the behavioral biases we use to construct B-counts (detailed in Section 3-A). A second includes outcomes *Y*: objective and subjective measures of financial condition, and standard measures of other aspects of utility/well-being (Section 4-A and 5-D). A third includes classical decision inputs *X*: demographics (including life-cycle factors), classical time and risk preferences/attitudes, and cognitive and non-cognitive skills (Section 4-B). A fourth set includes survey effort: measures of time spent on our elicitations, and of item non-response (Section 4-A).

B. The American Life Panel

We administered our survey 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 surveys have been administered in the ALP, and RAND makes data publicly available after a period of initial embargo. We use data from some of those other modules to complement our data, as detailed in Section 5-D.

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. We show that our main results are mostly robust to using these weights.

C. Research design and sample

Two principles guided our research design. First, measure the richest set of individual characteristics possible, to minimize potential confounds from omitted variables and to allow exploration of relationships between B-counts and classical covariates such as demographics, cognitive skills, non-cognitive skills and classical preferences. Second, take repeated measurements at different points in time, to describe the temporal stability of behavioral summary statistics and to account for measurement error in consumer characteristics.

To those ends we administered our surveys to the same set of panelists twice, roughly three years apart. Each survey round required about one hour of survey time per panelist on average, with substantial cross-panelist heterogeneity in response time that control flexibly for, using the ALP's panelist-question-level "timings" data.

Per standard ALP practice, we paid panelists $10 per completed module. Beyond that, all but one of our elicitations are unincentivized on the margin (limited prospective memory being the exception; see Table 1 for details). We made this choice deliberately, based on research budget tradeoffs between various approaches to dealing with measurement error and identification—incentives vs. sample size vs. measuring a broad set of consumer characteristics vs. repeated elicitation over time—and scrutiny of usual motivations for paying marginal incentives. 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 and finds evidence to the contrary.[14] Researchers often hypothesize that unincentivized elicitations change inferences, but that hypothesis is not robustly supported empirically (e.g., Von Gaudecker, Van Soest, and Wengström 2011; Gneezy, Imas, and List 2015). In any case, our repeated elicitations and measurement error models should suffice to address concerns about noise. Researchers often

[14] For example, each ALP survey ends with "Could you tell us how interesting or uninteresting you found the questions in this interview?" and roughly 90% of our sample replies that our modules are "Very interesting" or "Interesting," with only 3% replying "Uninteresting" or "Very uninteresting," and 7% "Neither interesting nor uninteresting."

assume that marginal incentive mechanisms are the best way to mimic real-world stakes, but this is not generally true for behavioral consumers (Azrieli, Chambers, and Healy 2018), and tasks with hypothetical rewards like ours can offer some conceptual advantages (e.g., Montiel Olea and Strzalecki 2014).

To reduce survey fatigue, we worked with ALP staff to break each round into two separate 30-minute surveys (modules, in ALP parlance), offered about two weeks apart for most respondents. Respondents had some flexibility in choosing when to take a survey module once it was on offer.

After extensive piloting, the ALP fielded our first two Round 1 instruments (ALP modules 315 and 352) starting in November 2014. We targeted 1,500 working-age respondents, sending 2,103 initial invitations, and ultimately received 1,515 responses to Module 315, and 1,427 responses to both 315 and 352. 95% of respondents completing both modules did so by the end of February 2015. We then re-administered those same two modules (with some additional questions at the end, eliciting non-cognitive skills), seeking responses from the 1,427 panelists who completed both Round 1 modules, beginning in October 2017. Of the 1,427, 1308 remained in the ALP at Round 2 inception. Of those 1,308, we received 967 responses to the first module and 845 responses to both modules (ALP #474 and #472).[15]

3. B-Counts: Behavioral summary statistics

Here we develop our behavioral summary statistics. We define B-counts motivated by various classes of behavioral models and bias taxonomies, describe their cross-sectional distributions across panelists, and estimate their within-consumer persistence across three years.

[15] Modules 352 (Round 1 Module 2) and 472 (Round 2 Module 2) also included invitations to complete a short follow-up survey the next day. We use responses to the invitation and actual next-day behavior to measure limited memory, as detailed in Table 1.

A. Components of the B-count(s)

Each B-count aggregates information on 2-17 potential sources of behavioral biases, within-consumer. A finite research budget forces tradeoffs between the depth and breadth of bias measurements, incentives, and sample size. We prioritized biases that had been: linked to financial decisions in prior work, measured with elicitation methods that have been featured recently in top journals, are adaptable to an online environment, and could practicably fit into modules that would also measure other decision inputs and outcomes. We do not seek to measure all possible biases; rather, we start with a large set (by the standards of behavioral research), and focus on identifying whether and how that set and subsets can be empirically informative.

Among our 17 potential sources of behavioral biases, one subset relates to preferences: present-biased discounting (Read and van Leeuwen 1998; Andreoni and Sprenger 2012), loss aversion (Fehr and Goette 2007), preference for certainty (Callen et al. 2014), ambiguity aversion (Dimmock et al. 2016), and choice inconsistency (Choi et al. 2014). Other subsets capture biased beliefs, biased perceptions, and behavioral decision rules: three varieties of overconfidence (Moore and Healy 2008), narrow bracketing (Rabin and Weizsäcker 2009), exponential growth biases (Stango and Zinman 2009; Levy and Tasoff 2016), statistical fallacies (Dohmen et al. 2009; D. Benjamin, Moore, and Rabin 2017; D. Benjamin, Rabin, and Raymond 2016), and limited attention/memory (Ericson 2011).[16]

Table 1 summarizes our 17 potential sources of biases, along with our elicitation methods and their antecedents. Each bias is identified relative to the classical benchmark established in prior work (e.g., time-consistent discounting, consistency with the General Axiom of Revealed Preference, unbiased beliefs about one's own performance, unbiased perceptions of statistical properties, etc.). The Data Appendix Section 1 provides details on each of the 17, including granular data descriptions, comparisons of data quality indicators and descriptive statistics to prior work, and discussions of prior theory and evidence linking each behavioral bias to consumer decisions and outcomes.

[16] Following a common delineation in behavioral economics, we do not measure social preferences. See Dean and Ortoleva (2018) and Chapman et al. (2018a) for evidence on relationships between behavioral biases and social preferences.

B. B-counts and sub-counts: Definitions and motivations

B-counts classify behavioral biases simply: for each of the 17 potential sources of bias we measure, we classify a consumer as displaying a bias (1) or not (0). A B-count simply sums, within-consumer, a number of biases exhibited. This approach to creating summary statistics is transparent and easy to implement, here and in future work. (We consider alternate functional forms in the Results Appendix, use various subsets of the 17 potential biases throughout the paper, and discuss alternate statistical approaches to efficiently capturing information from a given set of bias measures in Section 7.)

As discussed in Section 1, conceptually one can view a behavioral summary statistic as a within-consumer aggregation of many different behavioral influences, or as a psychological underpinning for many different behavioral biases. We show below that B-counts are empirically as well as conceptually distinct from classical decision inputs, in Sections 5-B and 6.

Our different B-counts nod to these different conceptions of behavioral summary statistics; e.g., our Full B-count is a broad aggregation of all 17 potential bias sources, while our Sparsity B-counts aggregate subsets of potential biases motivated by the foundational role of the limited attention/memory parameter in Gabaix's models.[17] Our "Narrow Sparsity" B-count sums only limited attention and limited memory. Besides speaking to sparsity and other attention-based theories, the Narrow Sparsity B-count has the added benefit of being easy and quick to elicit; this is reflected in the elicitation time statistics in Table 2 Column 5 and discussed in Section 7. The "Broad Sparsity" B-count adds six more biases that can emerge from limited attention/memory in Gabaix's models: our two measures of present-biased discounting (for money and for consumption), and our four measures of price misperceptions and statistical biases: exponential growth biases, non-belief in the law of large numbers, and the gambler's fallacies.[18]

The four price misperception and statistical biases are what we call math biases. They have objectively correct answers, but they do not simply measure math mistakes or cognitive skills,

[17] Gabaix describes limited attention as a "central, unifying theme for much of behavioral economics" (2019, p. 1). He does not explicitly mention limited memory, but it is implicit: a consumer might fail to "consider" an economic variable by forgetting it, and vice versa.

[18] Narrow bracketing also seems very much in the spirit of the Gabaix models, but we do not include it in our Broad Sparsity B-count because we could not find any mention of it in Gabaix's papers. Several untabulated robustness checks suggest that including it would not change our inferences.

because they are tendencies to err in a particular direction. As an example, work on Exponential Growth Bias shows that more people underestimate the effects of compounding than overestimate it, and that people who underestimate it do so systematically across a range of financial decisions, with plausibly welfare-reducing consequences (Stango and Zinman 2009; Levy and Tasoff 2016). So, EG biases are not just mathematical mistakes: they are biases. Limited math/cognitive skills, on the other hand, generate mistakes that are non-systematic, mean-zero, and hence less likely to push people toward particular decisions (such as less saving and more borrowing) on average. We draw the distinction for two reasons. One is to confirm that our *non*-math biases are empirically relevant, and that math biases alone do not drive our observed correlations between B-counts and outcomes (Section 5-C). We also conduct a variety of other empirical tests in Sections 5-B and 6 to confirm that the math biases themselves are distinct from numeracy and other math-related cognitive skills.

Each of our 17 behavioral biases has an expected direction emphasized in prior work (Table 1 Column 3; details in Data Appendix Section 1); e.g., present-bias or underestimating exponential growth is expected while future-bias or overestimating EG is not.[19] Below we detail how expected-direction biases actually are more prevalent (Section 3-D), and more strongly correlated with outcomes (Section 5-C), than non-expected direction biases.

Our third B-sub-count couplet (besides math vs. non-math, and expected vs. non-expected) is preference vs. non-preference B-sub-counts. The latter includes biases pertaining to beliefs, price perceptions, and problem-solving approaches. The former includes our two measures of discounting biases,[20] ambiguity aversion, loss aversion/small-stakes risk aversion, our two measures of inconsistency with GARP and dominance avoidance, and preference for certainty. The mapping from preference biases to welfare implications is less clear than for non-preference biases, both theoretically and empirically, as we discuss in Section 5-C.

C. B-counts are stable within-person, over time

Table 2 Column 7 reports estimates of B-count temporal stability: round-to-round, within-person correlations over our three-year sample period. Such correlations are important because we

[19] Chapman et al. (2018a) also find that expected direction biases are relatively prevalent.

[20] Keeping in mind that discounting is more than time preference *per se*: it is a reduced-form combination of preferences, expectations, and (perceived) rates of return.

use within-person stability in B-counts to deal with measurement error when estimating correlations between outcomes (measures of utility aspects) and B-counts (Section 4-C).

The Full B-count has a within-person correlation of 0.44 across the two rounds, which is high by psychometric standards; given measurement error, even a measure that captures a stable trait will have serial correlation below one.[21] Within-sample, the Full B-count is more stable than our measure of patience, has about the same stability as our measures of risk aversion, and is less stable than our measures of cognitive skills.[22] (It is an open question whether cognitive skills are truly more stable (trait-like), and/or just measured more accurately. The latter explanation seems quite likely from a "testing" perspective, given that researchers have devoted orders of magnitude more effort to refining measures of cognitive skills than to refining the measures of behavioral biases we use here.)

The B-sub-counts have estimated round-to-round within-person correlations ranging from 0.18 to 0.49 (Table 2 Column 7). Two comparisons between B-sub-count couplets are particularly noteworthy. Expected direction biases are more than twice as stable as non-expected ones (0.41 vs. 0.18), suggesting that expected biases are more trait-like and/or easier to measure accurately. And non-preference biases are more than twice as stable as preference biases (0.49 vs. 0.23), despite us devoting more time to measuring preference biases (3 minutes per vs. 1 minute per non-preference bias).

[21] In the absence of prior work measuring the stability of behavioral summary statistics, the most relevant out-of-sample comparisons are studies of the temporal stability of single behavioral biases. Meier and Sprenger (2015) finds a one-year within-person correlation of 0.36 for a short-run money discounting parameter that is strongly present-biased on average. Chapman et al. (2018b) finds a 6-month within-person correlation of 0.21 for a measure of ambiguity aversion. Chapman et al. (2018a) elicits multiple measures ("duplicates") of 12 biases that are conceptually similar to ours, *at a single point in time*, and finds an average within-bias, within-person, across-measure correlation of about 0.6 (our calculation).

[22] More specifically, some of the relevant within-person round-to-round correlations in our sample are: 0.30 for patience, 0.58 for the Dohmen et al. (2010, 2011) measure of risk aversion, 0.32 for the Barsky et al. (1997) measure of risk aversion, 0.75 for the number series measure of fluid intelligence, and 0.70 for the first principal component of our four cognitive skills test scores. See Section 4-B and Appendix Table 1 for details on these variable definitions.

D. B-count distributional properties

Table 2 presents additional statistics for B-counts within and across our two survey rounds.[23] Besides being descriptively interesting in their own right, these statistics have implications and applications for diagnosing, modeling, and treating the influence of multiple behavioral biases, as Section 8 discusses. We start by considering prevalence, central tendencies, median elicitation time,[24] and missing data. Then we focus on cross-sectional heterogeneity at the end of this section.

Table 2 Panel A describes the Full B-count. The mean panelist exhibits about 10 biases out of a maximum 17, whether we use all round 1 data (i.e., panelists who completed both of our round 1 modules; N=1427), round 1 data only for panelists who went on to complete round 2 (N=845), or round 2 data (N=845). The median, not shown in the table, also equals 10 in each case. Nearly everyone exhibits at least one bias (Column 3 shows 100% with rounding), although no one exhibits the maximum possible 17 (Column 4). The standard deviation is roughly 2 on a mean of 10. Median survey time for eliciting the full B-count is about 34 minutes (Column 5); we focus on this more in Section 7.

Table 2 Panel B shows that our lower-dimensional Sparsity B-counts are also prevalent. The Narrow Sparsity B-count is above zero (out of a possible two biases) for 84% of our sample. Critically, the Narrow Sparsity B-count only takes about a minute of survey/task time to measure (Column 5). That, coupled with its strong conditional correlations with various outcomes (Section 5), suggests that measuring the Narrow Sparsity B-count could be a valuable and practical addition to many studies of consumer decision making. The Broad Sparsity B-count takes on a value greater than or equal to one for nearly everyone in our sample, with a mean of roughly 4.2 out of a maximum possible 8 biases and SD of 1.3.

Table 2 Panel C describes our three B-sub-count couplets. Expected-direction biases are far more prevalent than non-expected ones: the Expected-Direction B-sub-count mean is nearly as high as the Full B-count (roughly 8.5, with a SD of about 2) while the Non-expected mean is only 1.5. And while nearly everyone exhibits multiple expected-direction biases, roughly 15% of our sample exhibits zero non-expected biases (out of a possible 8). Expected-direction biases drive the

[23] Appendix Table 2 shows that B-count descriptive statistics are similar if we use the ALP sampling weights.

[24] Our elicitation time is an upper bound on the true time panelists spend, because respondents can take breaks that are imperfectly captured by the ALP’s click-to-click measures of time spent.

Full B-count, in that the two are correlated 0.87; in contrast, the Non-expected Direction B-count correlation with the Full is only 0.26. Both math and non-math biases are prevalent and heterogeneous, with cross-sectional variation in the Full B-count driven less by the Math (correlation 0.57) than the Non-math B-count (correlation 0.90). Both preference and non-preference biases are prevalent and heterogeneous, with non-preference biases driving the Full B-count more than preference biases (correlations 0.82 vs. 0.51). We consider conceptual and practical differences between behavioral preferences and other biases in Section 5-C.

Table 2 Panel D suggests that item non-response does not overly complicate interpretation of B-count variation. On average, only 1 out of maximum possible 17 biases is missing due to non-response (Column 1), with a standard deviation of about 1.5. Every panelist responds to one or more bias questions. Below we control directly for the missing B-count inputs and other measures of survey effort (Section 4-A).

Understanding the extent of cross-consumer variation in B-counts is important, given our focus on links between cross-sectional variation in B-counts and welfare measures (Sections 4 and 5), and the key role estimates of the variance of a behavioral summary statistic can play in behavioral sufficient statistics modeling (Section 8-B-ii). Table 2 suggests that B-count heterogeneity is substantial, with standard deviations of about 20-50% of the mean for our three main B-counts (Panels A and B). One might wonder if these estimates are substantially upward-biased by measurement error, but Figures 1a and 1b provide some reassurance: comparing the figures shows that dispersion in the Full B-count for our full sample is only modestly greater than for the sub-sample with identical B-counts across our two rounds.

4. Using B-counts to Model the Wedge Between Decision and Normative Utility

Having found substantial heterogeneity in our B-counts, we now detail how to use that heterogeneity to examine the fundamental assumption of behavioral summary statistic models: biases drive a wedge between normative and decision utility. Recall our empirical framework:

$$(1)\ Y_i = f(Bcount_i) + g(X_i) + h(Surv_i) + \varepsilon_i$$

Now we pay particular attention to identifying assumptions given measurement error in one of its three primary objects: utility aspects Y_i, the behavioral wedge as measured by a $Bcount_i$, and classical inputs X_i. We also consider the impact of noise from variation in survey effort $Surv_i$, and discuss the impact of measurement error and other econometric concerns on ε_i.

A. Outcomes: Measuring financial well-being and other aspects of consumer welfare

Our approach to the left hand side of (1) is to consider individual-level outcomes measuring various important "aspects" of consumer welfare, following Benjamin et al. (2014) and Benjamin et al. (2014). We focus on financial measures here, describe measures of other aspects in Section 5-D, and consider relationships between aspects and overall utility in Section 4-D. We scale all outcomes on the [0,1] interval, with higher values indicating better outcomes (Table 3).[25]

Our primary outcome is an index of *subjective financial condition*—an aspect of consumer welfare relating to household finances—that averages responses to four sets of questions about retirement savings adequacy, non-retirement savings adequacy, overall financial satisfaction, and financial stress.[26] The four index components correlate strongly and positively with each other (Appendix Table 3 Panel B): the pairwise correlations range from 0.31 to 0.53, each with p-values < 0.001.

We also measure *objective financial condition* by averaging five indicators: positive net worth, owning retirement assets, owning stocks, having saved over the past 12 months, and not having experienced any of four financial hardship indicators. These index components are strongly positively correlated with each other: the range is 0.35 to 0.56 (Appendix Table 3 Panel A). The objective index is correlated 0.57 with the subjective index (Table 3).

Our empirics allow that we measure welfare/utility with error. Putting aside issues with aggregating from single aspects to overall utility until Section 4-D, for now we allow a random error component ε_i and/or links between survey effort $g(Surv_i)$ and outcome reporting Y_i for a

[25] Re-scaling provides comparability, and we chose the [0, 1] scale because most of our outcome variables are either indicators or summary indexes. We do *not* standardize, because dividing a variable by its standard deviation can introduce additional measurement error (Gillen, Snowberg, and Yariv forthcoming).

[26] We drew the content and wording for our financial condition questions from previous American Life Panel modules and other surveys (including the National Longitudinal Surveys, the Survey of Consumer Finances, the National Survey of American Families, the Survey of Forces, and the World Values Survey). Each of our outcomes is quick and easy to measure: Appendix Table 3 and Table 3 show that each individual/component outcome takes strictly less than a minute to elicit on average, and that even our most elaborate index has a median elicitation time of only 2.67 minutes. Further details on outcome variable definitions can be found in the notes to Table 3 and Appendix Table 3.

given aspect. The vector $g(Surv_i)$ contains flexibly parameterized measures of survey response times and item non-response.[27] The survey and item response vector allows, among other things, for the possibility that non-response in other variables could be correlated with reported outcomes, and/or that rushed or very long response times on behavioral elicitations could be spuriously linked to reported outcomes.[28]

B. Classical decision inputs: Measuring skills, presumed-classical preferences, etc.

We construct a rich vector of classical decision inputs X that are presumed to drive choices in most economic models: (life-cycle) demographics such as income, gender, age, education, and family structure; presumed-classical patience and risk tolerance; and cognitive and non-cognitive skills. Altogether we measure 20 consumer characteristics with 121 variables (many of them categorical, see Appendix Table 1).[29] We measure nearly all of these inputs in both of our survey rounds, and thus can allow for the possibility that these inputs too are measured with error (Section 5-B).

We measure demographics using the ALP's standard set, collected when a panelist first registers and refreshed quarterly. We measure the other elements of X_i with widely-used elicitations administered in our modules: risk attitudes/preferences with the adaptive lifetime income gamble task developed by Barsky et al. (1997), and the financial risk-taking scale from Dohmen et al. (2010, 2011);[30] patience using the average savings rate across the 24 choices in our version of the Convex Time Budget task (Andreoni and Sprenger 2012); cognitive skills using 4 standard tests for general/fluid intelligence (McArdle, Fisher, and Kadlec 2007), numeracy (Banks and Oldfield 2007), financial literacy (crystalized intelligence for financial decision making) per

[27] Specifically, we measure respondent survey effort with three types of variables. One is the count of missing inputs to our B-count, as described in Section 3-D. The second type is indicators for item non-response, for elicitations with non-trivial item non-response rates. In our main analysis sample, these rates range from zero for many demographics, to 5% for Stroop. Our third measure of survey effort is based on the ALP's tracking of a panelist's time spent on each screen. We use decile indicators of survey time spent per survey round, either overall across both of our modules, or counting just our behavioral elicitations.

[28] In untabulated results, we exclude the top and/or bottom deciles of survey response time, or use survey response times as weights. Neither approach has a meaningful effect on the findings.

[29] Including such a rich set of classical covariates might over-control if classical covariates are correlated with behavioral tendencies, but we show below that in practice our estimated links between B-counts and outcomes are quite robust to the set of covariates (Section 5-B).

[30] These Barsky and Dohmen et al. measures are correlated 0.14 in our main analysis sample. We also elicit Dohmen et al.'s general risk taking scale, which is correlated 0.68 with the financial scale.

Lusardi and Mitchell (2014), and executive function/working memory (MacLeod 1991).[31] Pairwise correlations between these four test scores range from 0.16 to 0.42. In our second round of surveying we add elicitations of noncognitive skills to the end of our second module.[32]

In some specifications, where Y_i is our index of subjective financial condition, we add the objective financial index to X_i. This makes sense if one posits the behavioral wedge as operating conditional on resources and constraints; e.g., taking someone's budget constraint as given, do behavioral biases reduce utility/well-being? In any case, conditioning on financial resources provides an even more stringent test of the relationship between subjective financial condition and a B-count, albeit one that errs on the side of over-controlling.

C. Measuring the behavioral wedge and accounting for the overall error structure

Completing our empirical specification requires a measure of a behavioral summary statistic, for which we use a B-count, and an estimator that allows for the possibility that a B-count imperfectly measures the "true" behavioral summary statistic. Such measurement error could attenuate the estimated link between the summary statistic and outcomes, or falsely identify such a relationship where none exists (Fuller 2009; Gillen, Snowberg, and Yariv forthcoming). We address this challenge using the repeated measurements across our two survey rounds.

The standard approach to dealing with measurement error in a B-count would be to use its non-contemporaneous elicitation (Round 2 in the Round 1 model, and vice versa) as an instrument for the contemporaneous elicitation. This instrument will lead to an unbiased estimate of the relationship between the behavioral summary statistic if the measurement errors in the behavioral summary statistic are uncorrelated across rounds, and those errors are uncorrelated with the regression error (which includes, among other things, error in measuring the utility aspect Y_i). Such an approach would yield two "single IV" models for estimation, where for each consumer we have:

[31] The Data Appendix Section 2 provides details on each of these measures.

[32] Specifically, we use the validated 10-item version of the Big Five personality trait inventory (Rammstedt and John 2007). We initially decided against eliciting non-cognitive skills, given our resource constraints and the lack of prior evidence of the correlations between them and behavioral biases (see, e.g., Becker et al's (2012) review article) that would be required to confound our key inferences. But in Round 2 we decided to err on the side of caution and take some non-cognitive skills measures, after seeing Kuhnen and Melzer's (2018) evidence on correlations between financial outcomes and non-cognitive skills.

Observation	Y_i	$X_i, Surv_i$	*B-count*	*B-count IV*
1	Round 1	Round 1	Round 1	Round 2
2	Round 2	Round 2	Round 2	Round 1

We go beyond the single IV approach by implementing the "both-ways" approach of Obviously Related Instrumental Variables (Gillen, Snowberg, and Yariv forthcoming). ORIV stacks the data, using both the first elicitation to instrument for the second *and* the second elicitation to instrument for the first. As with the single IV approach, ORIV will produce an unbiased estimate of the link between the summary statistic and outcomes if the measurement errors in the behavioral summary statistic are uncorrelated across rounds.

In our setting we elicit two rounds of data for nearly *all* outcomes and covariates as well as for our variables of greatest interest (in our case, those used to construct the B-count). Thus we inflate our two observations per person to four "replicates" (per Gillen et al.):

Replicate	Y_i	$X_i, Surv_i$	*B-count*	*B-count IV*
1	Round 1	Round 1	Round 1	Round 2
2	Round 1	Round 1	Round 2	Round 1
3	Round 2	Round 2	Round 2	Round 1
4	Round 2	Round 2	Round 1	Round 2

We first estimate the model separately for replicates 1 and 2 ("round 1 ORIV") and compare those estimates to those obtained using replicates 3 and 4 ("round 2 ORIV"). We do not reject the restriction that the empirical relationships are identical for round 1 ORIV and round 2 ORIV. Thus most of our empirics in Section 5 pool the four replicates, clustering standard errors by panelist. In some specifications we also treat other covariates such as cognitive skills as measured with error, and employ ORIV for them as well (i.e., by instrumenting for round 2 covariates with round 1 covariates, and vice versa).[33]

D. Interpretation and non-classical measurement error

Here we consider some issues of interpretation and some potential sources of non-classical measurement error, none of which seems to be problematic in practice.

We intend for our outcomes to measure what Benjamin et al. call different aspects—components—of overall well-being and utility. Our objective and subjective financial indexes measure a financial aspect; in Benjamin et al. (2014) the financial aspect has a high weight in terms of overall relative marginal utility (ranking 6th out of 113 aspects on their list).

Interpreting our outcomes as aspects of utility, rather than overall utility, comes with one cost and two benefits. The cost of course is that our estimated linkages between behavioral biases and outcomes are aspect-specific (consumer-aspect level), not holistic (consumer-level). The benefits are that identification is easier and more transparent. The "easier" piece is that we avoid having to extrapolate from aspect-level to overall utility, as the Benjamin et al. papers warn against. The transparency piece is that we can identify how B-counts correlate differently with different aspects of utility (Sections 5-D and the Results Appendix).

One potential problem is misspecification of an outcome index's component weights. Our indexes weight each component equally, which could bias the coefficient on the B-count in either direction depending on the relationships between index component correlations with the B-count and index component contributions to (weights in) utility. We check this potential confound by examining whether the coefficient on the B-count differs dramatically across individual components of the indexes, and find that it does not, at least not qualitatively (Results Appendix). Hence, the results would be fairly invariant to many different combinations of weights. A similar issue could arise with the B-count, as we mentioned earlier, if our measure incorrectly weights or omits relevant biases, but ORIV addresses B-count mis-weighting.

A second potential issue is omitting an important component of aspect-level well-being from an index. This seems unlikely to be a material problem, at least for our financial condition indexes,

[33] Here we rely on the fact that our other covariates are also stable within-person over time, as detailed in footnote 22.

given the breadth of our measures. But even if there were such an omitted component, it also would have to (a) have a relatively high marginal utility weight for that aspect, and (b) have a weaker correlation with the B-count.

A third potential problem would arise if it were somehow easier for low-effort survey respondents to indicate worse outcomes than better ones, since it is presumably easier to indicate behavioral tendencies (thereby upping one's B-count) than classical ones. Controlling for $g(Surv_i)$ accounts for any systematic relationships between survey effort, self-reported outcomes, and B-counts. The survey time variables may over-control, if they reflect behavioral tendencies or other characteristics like cognitive ability, but including them is a "better safe than sorry" approach and in practice does not change our inferences (Section 5-B). Further, as the Data Appendix Section 3 details, our survey user-interfaces do not make it easier for respondents to indicate systematically better or worse outcomes.

A fourth potential problem is that measurement error in a low-dimensional B-count (e.g., Narrow Sparsity) could be due to misclassification and hence non-classical. We address this in the Results Appendix.

5. Results: Links between B-counts and outcomes

A. Primary results: B-counts and outcomes

Table 4 estimates our primary, pooled ORIV specification on the sample of 845 ALP panelists who completed all four modules across our two survey rounds.[34] The models here regress objective or subjective financial condition[35] on one of our three main B-counts (in levels)[36] and our complete set of additional covariates, with each column presenting a B-count coefficient and standard error from a single ORIV regression. Columns 3, 6, and 9 regress subjective financial condition on the same covariates and add the objective financial condition as an additional covariate.

[34] Appendix Table 4 estimates ORIV round-by-round and does not reject equality in B-count coefficients across rounds. Appendix Table 5 uses Round 2 data only, with Round 1 as instruments (to help address reverse causality), and finds similar results to our pooled specifications. See the Results Appendix for discussion.

[35] Appendix Table 6 uses each of our financial index components as separate outcomes, and finds similar qualitative results for the B-counts but with some quantitative differences. See the Results Appendix for discussion.

[36] Appendix Table 7 uses alternative functional forms of the Full B-count and finds similar results. See the Results Appendix for discussion.

The Full or Sparsity B-counts strongly and negatively conditionally correlates with outcomes (p-value <0.01) in each of Table 4's nine specifications, and the economic magnitude of β^{θ} is large in every specification. We report marginal effects in the "d(LHS)/d(1 SD B-count)" row, and the smallest one of the nine implies a 22% decline in average financial condition associated with a one standard deviation increase in a B-count (the -0.119 in Column 4, divided by the LHS mean). The d(LHS)/d(1 SD B-count) row also shows similar magnitudes across the Full and Sparsity B-counts, with 8 of the 9 estimated marginal effects within [-0.179, -0.119].[37] These marginal effects are about the same size of those on the objective financial condition index in the specifications where that index is included as an additional control variable. The Results Appendix provides comparisons to other covariates from Appendix Table 10; perhaps the most noteworthy pattern is that the B-count is unusually robustly correlated with financial outcomes.

In all, Table 4 shows that B-counts have economically and statistically strong conditional correlations with financial outcomes.

B. Results are invariant to the set of covariates/controls

Table 5 examines robustness to the set of covariates and shows that the B-count estimates are nearly invariant to the composition of our rich set of controls. These results suggest that researchers interested in cross-sectional heterogeneity can economize on measuring other covariates alongside a B-count: the overall cost of adding measurement of a behavioral summary statistic to a research design can be low (Section 7). Moreover, the stability of the B-count coefficient across specifications with vastly different controls provides reassurance that it captures a behavioral wedge and not omitted components of other covariates.

We consider 27 specifications in Table 5, each using the subjective financial index on the LHS, with a panel per each of our three main B-counts and columns permuting whether and how we include other covariates. Column 1 in each panel reproduces our main specification. Column 2 drops the demographic variables, Column 3 further drops classical preferences, Column 4 drops

[37] Appendix Table 8 compares sampling-weighted estimates to unweighted ones from Table 4 and reveals that the weighted coefficients are larger in point terms but less precise. Appendix Table 9 shows that our results are qualitatively robust to an alternative approach to dealing with measurement error, including non-classical misclassification error (Black, Berger, and Scott 2000) in the two-dimensional Narrow Sparsity B-count. See the Results Appendix for discussion.

cognitive skills, and Column 5 drops non-cognitive skills. Column 6 drops the survey time spent deciles, leaving the count of missing biases as the only covariate besides the B-count.

The B-count estimate in each of those more parsimonious specifications is very similar to those from our main specification; e.g., for the Full B-count the coefficient is -0.078 (SE 0.009) in column 6 vs. -0.084 (SE 0.018) in column 1. (This invariance to other covariates does not hold if one fails to account for measurement error in the B-count by, e.g., using OLS instead of ORIV.)[38]

Columns 7-9 further address robustness by using ORIV to allow for measurement error in not just the B-count, but also in one or both of two groups of classical inputs: presumed-classical preferences and cognitive skills.[39] These results are similar to those from the other specifications.[40]

In all, Table 5 helps solidify the inference that B-counts capture a distinctly behavioral wedge between decision and normative utility.

C. Full B-count decompositions

Table 6 decomposes the Full B-count in three ways shown/discussed earlier (Table 2 Panel C and Section 3-B), to shed light on some nuances of identification and interpretation. Each regression here takes one of our main specifications (Table 4, Columns 1 and 2) and replaces the Full B-count with a couplet: with two mutually exclusive and exhaustive B-sub-counts.

The first two regressions here decompose the Full B-count into the Math Bias and Non-Math Bias sub-counts. The non-math biases have strong negative conditional correlations with both objective and subjective financial condition, while the point estimates on math biases are close to zero (albeit imprecisely estimated).[41] These results offer further reassurance that B-counts are not

[38] E.g., Appendix Table 11 shows that the OLS estimate of the correlation between the subjective financial condition index and the Full B-count, using both rounds of data and the full set of covariates, is roughly half of that in the most parsimonious OLS specification (-0.019 vs. -0.035, each with a SE of 0.003).

[39] Ideally we would allow for measurement error in non-cognitive skills as well, but we lack multiple measures for personality traits because we did not elicit them in Round 1, as detailed in footnote 32.

[40] Appendix Table 12 estimates the same 27 specifications with the objective financial condition index as the LHS variable instead of the subjective index. It reveals a similar pattern to Table 5, with two key exceptions: dropping *all* of the other covariates makes the Full B-count and Broad Sparsity B-count correlations with objective financial condition substantially more negative (i.e., compare Column 6 to the other columns in Panels A and B).

[41] Appendix Table 13 shows similar results from dropping the cognitive skills covariates and/or decomposing the math biases into expected vs. non-expected directions.

"just math": they capture something distinct from classical conceptions/measures of cognitive skills/math ability.

The next two regressions decompose the Full B-count into expected- and non-expected directions. Recall that expected-direction biases are those held to be more common/impactful in prior work, such as present-bias, overconfidence and underestimating exponential growth. The Expected-direction B-count has strong negative conditional correlations with both financial condition indexes, with point estimates almost identical to those of the Full B-count in Table 4. The Non-expected Direction B-count (future bias, under-confidence, overestimating exponential growth, etc.) also has negative and large point estimates, but they are imprecisely estimated. Taken together these results validate behavioral economics' focus on expected direction biases, while leaving unresolved the question of whether measures of non-expected direction biases capture something substantive or merely noise.[42]

The last two columns in Table 6 compare Preference vs. Non-preference B-sub-counts. This decomposition is informative because the welfare implications of behavioral preference biases (loss aversion, ambiguity aversion, etc.) are less clear than for non-preferences (biased price perceptions and expectations, limited attention, etc.).[43] Consequently the important tests in Table 6 Columns 5 and 6 are on the Non-preference B-sub-count. One would question whether the Full B-count results are truly indicative of consumer welfare losses if it turned out that only behavioral preferences were driving the negative conditional correlation between our financial condition

[42] How one interprets the Non-expected Bias B-sub-count results is highly contingent on one's prior: our prior was agnostic, and hence our interpretation is that these noisy results tell us little about the economic importance of non-expected directional biases. But if one had a strong prior that non-expected bias measures reflect noise rather than true biases, then these results provide some support for that hypothesis. We explore this further in Section 6.

[43] A policymaker has clearer normative grounds for correcting non-preference biases. In contrast, one might consider preferences inviolate, even if they are not classically normative. A policymaker may lack grounds for trying to debias someone who is ambiguity averse, but probably has grounds for trying to debias someone who underestimates the power of the Law of Large Numbers. Hewing closer to our framework, the point is simply that if behavioral preferences are truly preferences, then the preference components of a behavioral summary statistic may not drive a wedge between decision utility and normative utility. Related, if the only material behavioral components of decision making were grounded in preferences, one might still rely on revealed preference for welfare analysis.

measures and the Full B-count. But columns 5 and 6 show that is not the case; the Non-preference Bias B-sub-count is strongly negatively correlated with both financial condition indexes.[44]

In all, Table 6 helps further solidify the inference that B-counts capture something distinctly behavioral—a behavioral wedge between decision and normative utility.

D. Other outcomes: Different (and broader) measures of consumer welfare aspects

Table 7 expands the set of outcomes to include additional aspects of utility: life satisfaction, happiness, and health status. These aspects, like the financial aspect, have high marginal utility rankings per Benjamin et al. (2014): life satisfaction ranks 11th, health ranks 3rd and happiness ranks as high as 2nd. In Column 1 we also reproduce our main results with subjective financial condition as the outcome (from Table 4), for reference.

The pairwise correlations between our measures of life satisfaction, happiness, and health status range from 0.32 to 0.65 (Appendix Table 3 Panel C; Table 3). Table 3 also shows that these measures are strongly positively correlated with our indexes of subjective financial condition (the range is 0.29 to 0.50) and objective financial condition (from 0.29 to 0.35). Except for one elicitation of life satisfaction, all of these other elicitations come from modules other than ours, in periods roughly coincident with our study period.[45] Varying response rates across these other modules produces varying sample sizes across columns in Table 7.

The Full B-count coefficients in Table 7 Panel A are imprecisely estimated zeroes, while the Sparsity B-count coefficients are more clearly negative, with all of the eight new point estimates in Panels B and C implying marginal effects <-0.04 (on bases of 0.50 to 0.70), and six of them having p-values <0.05. For the Sparsity counts, a one standard deviation increase in the B-count is associated with life satisfaction 5-15% lower on the mean. For health and happiness, the corresponding declines are roughly 10% on the mean. For comparison, the Broad Sparsity

[44] Meanwhile, the Preference Bias B-sub-count has a less robust relationship with financial condition. There are various ways to interpret these results, depending on one's priors. Our view is that the consumer welfare consequences of behavioral preferences remains an open question. We explore this further in Section 6.

[45] In deciding which measures to merge in from other modules, we define "study period" as post-our Round 1 (we could not find any relevant measure post-our Round 2), and select questions that have: a) been used in other studies; b) measure highly rated "aspects" of subjective well-being in the marginal utility sense per Benjamin et al. (2014); c) are answered at least once by at least 2/3 of our sample. See Table 3 and Appendix Table 3 for details on the construction of each variable.

coefficients are roughly the same magnitude as moving down the income distribution by one to four deciles (depending on the outcome and position in the income distribution).

What might explain the pattern here of non-results for the Full B-count coupled with stronger results on the Sparsity counts? One explanation is that we chose the full bias collection specifically with links to financial decision making/outcomes in mind (as the Data Appendix Section 1 details). The Sparsity Biases are motivated somewhat more broadly, although Gabaix too focuses to a great extent on financial choices. In any case, further examining links between different definitions/conceptions of behavioral summary statistics and different outcome domains (different aspects of utility) is a promising line of future inquiry.

E. *Additional results/robustness checks*

The Appendix discusses several additional results and robustness checks mentioned above that require some elaboration (Appendix Tables 4-10).

F. *Summary interpretation of conditional correlations*

Altogether, the results in Tables 4-7 (and accompanying Appendix Tables) indicate economically large negative conditional correlations between B-counts and various outcome measures understood to capture important aspects of consumer welfare. These results are consistent with the foundational presumption of behavioral sufficient statistic models that behavioral biases, taken together, drive a wedge between decision utility and normative utility.

6. B-counts are distinct from other decision inputs

This section ties together several sets of results showing that B-counts capture something about decision making that is distinct from measures of classical decision inputs and our other covariates.

Recapping what we have learned already: 1) B-counts are strongly correlated with outcomes (measures of consumer welfare aspects), *conditional* on our rich set of additional covariates; 2) Those correlations are robust to very different specifications of the additional covariates, suggesting that any correlations between B-counts and other covariates do not confound inferences on the link between B-counts and decisions/outcomes.

We now add: 3) Variation in the Full B-count is poorly explained by our rich set of additional covariates. In addition to rounding out our description of B-counts' statistical properties (see also Section 3), this exercise adds to the "Who is (more) behavioral" literature (e.g., D. Benjamin, Brown, and Shapiro 2013), by adding evidence on fit to the prior focus on correlations, and by adding evidence based on consumer-level metrics of behavioral tendencies to a literature that has considered behavioral biases piecemeal.

Figure 2a plots raw, consumer-level variation in the "B-proportion": the share of our 17 biases a consumer exhibits. Using the proportion instead of the level B-count accounts for missingness without overfitting. Figure 2b plots consumer-level residuals from regressing the B-proportion on the complete set of other covariates in our data (see Appendix Table 1 for the list). These residuals are rescaled to the mean of the raw B-proportion in Figure 2a for comparability. Comparing the figures illustrates how little variation in the Full B-count is explained by our complete set of other covariates; although partialing out variation explained by other covariates does produce a more normal B-count distribution, it does little to reduce dispersion (the raw vs. residualized interquartile ranges are [0.56, 0.75] vs. [0.59, 0.74]).

Figures 3a-3d provide some simple univariate comparisons further highlighting that B-counts are not simply proxies for other covariates found to correlate with behavioral biases.[46] These show distributions of our B-proportion, broken out for paired groups at the opposite ends of the income, risk aversion, education, and cognitive skills distributions. These do show the expected level differences on average; e.g., the B-count distribution is shifted somewhat rightward for those in the lowest cognitive skills quartile relative to the highest. But also noteworthy is that the B-count varies substantially within each of the sub-groups we examine. Indeed, within-group variation in the B-count dwarfs cross-group variation, even between groups that are very different by construction.

Table 8 quantifies this in a multi-variate framework, using OLS to estimate the amount of variation explained by other covariates for each of our nine B-count proportions (Table 5, Columns 7-9 offer reassurance that measurement error in the additional covariates is unlikely to affect the OLS inferences here). The estimated fits (R-squareds) range from 0.12 for the Narrow Sparsity B-count to 0.40 for the Non-preference B-count, with 0.33 for the Full B-count. The subsequent rows show estimated partial R-squareds for subsets of covariates: demographics, cognitive skills, noncognitive skills (for which we have Round 2 data only), classical preferences (risk

[46] See, e.g., Benjamin et al. (2013), Burks et al. (2009), Cesarini et al. (2012), Chapman et al. (2018a), Dean and Ortoleva (2018), Frederick (2005), Li et al. (2013). See also Dohmen et al. (2018) on the relationship between measures of classical preferences/attitudes and cognitive skills.

preferences/attitudes, and patience estimated from our money discounting task), state of residence, and the deciles capturing time spent on our behavioral elicitations. Demographics and cognitive skills tend to explain the most variation, although their fit varies widely across the different B-counts. The demographics' R-squared ranges from 0.03 for the Preference B-count to 0.24 for the Math and Non-preference B-counts. The cognitive skills' R-squared ranges from 0.01 for the Preference B-count to the 0.33 for the Non-preference B-count. The other four groups of covariates—noncognitive skills, classical preferences, state of residence, and survey time spent—never explain more than 7% of the variation in a B-count across their 36 estimates (4 groups of covariates x 9 B-counts). Particularly striking is that, having adjusted for non-response on the LHS, respondent time spent completing our behavioral elicitations explains <=1% of the variation in the B-count proportions. This is consistent with our earlier hypothesis that nearly all respondents seriously engage with our surveys (Section 2-C).

Comparing these fit estimates across the various B-counts reveals several additional noteworthy patterns. The Sparsity B-counts are relatively poorly explained by our other covariates, consistent with Sparsity constructs capturing especially distinct and behavioral influences on decision making. Math biases are much better explained by demographics (which include education) and cognitive skills than non-math biases, as one would expect. Expected-direction biases are much better fit by other covariates (0.25) than non-expected direction biases (0.13). Coupled with the latter's very imprecisely estimated correlations with outcomes (Table 6) and relatively low within-person temporal stability (Table 2), the overall picture is consistent with non-expected direction biases reflecting more noise than signal. Preference biases (Column 8) are no better explained by the other covariates than non-expected direction biases, with only 1/3 of the fit of non-preference biases (Column 9).[47]

7. Efficiently measuring, and modeling with, B-counts

For researchers interested using behavioral summary statistics—for welfare analysis, targeting, theory-testing, and/or describing cross-sectional heterogeneity—a key practical question is how to efficiently measure behavioral summary statistics. Can one do so with a narrower and cheaper set

[47] As discussed above, this is not for lack of elicitation intensity: we have roughly three minutes per potential preference bias vs. one minute per non-preference bias.

of elicitations than our full set? Several of our results thus far suggest yes, yielding three pieces of concrete advice:

- Lower-dimensional behavioral summary statistics are informative. The Narrow Sparsity B-count is a good example of how theory can guide construction of a summary statistic based on only two underlying biases (Tables 4, 5, and 7).[48]
- Other covariates may not be necessary to estimate empirically stable conditional correlations between outcomes and behavioral summary statistics (Table 5).[49]
- Having at least two sets of plausibly independent behavioral bias elicitations is crucial, as multiple elicitations yield results stable enough to obviate the need to elicit and control for other covariates (compare Table 5 to Appendix Table 11).

A further point we establish here is that even a randomly selected subset of our biases can be useful. (It is important to keep in mind that a randomly selected subset of our biases is not the same as a random sample of the universe of behavioral biases; recall that we chose our full set based on prior work linking our 17 biases to financial decision making.) Figures 4a and 4b show the distributions of coefficients and standard errors from lower-dimensional B-counts using j:[1, 17] of potential behavioral biases in our data, where for any j we randomly sample up to 2,500 bias combinations, construct a B-count from the biases in that draw, and estimate our main ORIV specification with subjective financial condition on the LHS for that draw.[50] These figures show that coefficients and standard errors from B-counts based on small j converge fairly quickly to

[48] A quantitative if not qualitative caveat re: research budgets is that measuring the Narrow Sparsity B-count, and other B-sub-counts including our limited prospective memory elicitation (Table 1), does require additional resources beyond the survey time described in Table 2: 1) An additional, very brief, survey module for the follow-up task; 2) A financial incentive to complete the follow-up task. As measuring limited prospective memory is in its infancy (at least in broad samples and for economic applications), we would not presume that our elicitation is (cost-)efficient: there may be ways to elicit a useful measure within-survey, and/or with lower incentive payments per-respondent (e.g, by using a lottery instead of a piece rate). Having said that, in our implementation the marginal cost of the limited memory elicitation ended up being modest, because only about 15% of the sample actually completes the follow up task. We ended up paying about 1427*.15*$10 in Round 1 and 845*.15*$10 in Round 2, for a total of about $3,400.

[49] A weaker recommendation, from the perspective of economizing on measurement of other covariates, is that one round of data on them may well be sufficient (Table 5, Column 7-9 suggests that one need not worry about measurement error in other covariates biasing estimates of the B-count.)

[50] As j gets closer to 2 or 17, the number of possible combinations falls below 2,500—sampling 16 of our 17 potential sources of bias can have only 17 possible combinations, for example, as can sampling only one. Drawing 8 or 9 has 24,310 possibilities, the max, from which we draw 2,500.

what we obtain with j=17, with especially large gains in precision from measuring at least two biases.

There is, of course, much more work to be done to derive truly optimized measurement strategies and research designs for behavioral summary statistics. To take one example, while B-counts are simple by design, model selection techniques could increase power by guiding definitional, functional form, and other specification choices. (Having said that, such techniques would need to account for measurement error, and we are not aware of any that do.) To take another example, while repeated elicitations of a behavioral summary statistics are critical for dealing with measurement error, it remains to be determined which combination of timing and elicitation methods produces the most accurate and/or cost-effective measures. More broadly, efficient measurement requires consideration of tradeoffs between multiple margins of costly measurement, as we discuss in the Conclusion.

8. Using B-counts for welfare analysis and intervention design

This section provides some additional guidance on how one can use our behavioral summary statistics approach to help identify sound policy interventions for behavioral consumers and conduct welfare analysis of such interventions.

A. Policy Diagnostics

i. Might behavioral biases warrant intervention?

The threshold question for any behaviorally-motivated intervention—whether it be in health, household finance, energy, etc.—is whether behavioral biases materially reduce consumer welfare. To take a specific example from household finance, suppose a policymaker posits that behavioral biases reduce consumer welfare in either credit card or mortgage markets, or both (see, e.g., the Dodd-Frank legislation and subsequent implementing regulations). Our methods suggest several tests of this hypothesis that would help inform whether to proceed with developing behaviorally-targeted interventions:

1. Test whether B-counts are (conditionally) correlated with product-specific outcomes (i.e., replace *Y* in our equation (1) with outcomes of interest from the credit card and/or mortgage market like debt levels, severe delinquency, borrowing costs, etc.). Such outcomes are

admittedly not perfect proxies for consumer welfare, but in many empirical settings they are the best available data, and they also have the advantage of being dollar-denominated.

2. Test whether more-behavioral consumers fare worse in the requisite product market, in cases where there is plausibly exogenous variation in product usage D (where D could be market participation, debt level, etc.), using equations of the form:

   $$(2)\ Y_i = a(Bcount_i \cdot D_i) + b(Bcount_i) + c(D_i) + \ d(X_i) + e(Surv_i) + \varepsilon_i$$

   Here Y is an aspect-level welfare measure (like the ones we use in Table 7), and we are particularly interested whether the coefficient(s) on the first term function are negative and economically large.

3. Test the extent to which some "debiasing" intervention Z (disclosure, reminders, financial education, commitment, etc.) actually reduces bias, in cases with plausibly exogenous variation in Z (a pilot experiment, natural experiment, etc.), using equations of the form:

   $$(3)\ Bcount_i = f(Z_i) + g(X_i) + h(Surv_i) + \varepsilon_i$$

4. Test the extent to which Z mitigates the effects of bias, by substituting Z for D in equation (2) above and examining estimates of the *Bcount***Z* term(s):

   $$(4)\ Y_i = a(Bcount_i \cdot Z_i) + b(Bcount_i) + c(Z_i) + \ d(X_i) + e(Surv_i) + \varepsilon_i$$

The resource requirements for these sorts of diagnostic tests are modest, in the context of typical policy development and evaluation budgets. They require either a bespoke survey or adding some of our bias elicitations to routinely administered large-scale surveys.[51] There may also be opportunities to link to other sources of data on outcomes besides survey measures (e.g., supervisory data, credit reports, personal financial management apps, tax returns).

[51] There are many such surveys in the U.S. alone, in addition to the American Life Panel that we use, including the: Survey of Consumer Finances, Health and Retirement Study, Panel Survey of Income Dynamics, Consumer Expenditure Survey, National Longitudinal Surveys, Understanding America Study, ClearVoice, National Survey of American Families, National Financial Capability Study, National Financial Well-Being Survey, Medical Expenditure Panel Survey, National Health and Nutrition Examination Survey, National Health Interview Survey, Behavioral Risk Factor Surveillance System, and the Residential Energy Consumption Survey.

ii. What kinds of interventions might be optimal?

Theory shows that the empirical distribution of a behavioral summary statistic can provide qualitative guidance on designing interventions to treat the combined effects of multiple biases. For example, Baicker, Mullainathan, and Schwartzstein ("BMS", p. 1650) shows that that if there are variably biased consumers, with mean-zero bias on average, then whether a procedure's optimal copay is a subsidy or tax depends on whether it is socially beneficial on average. Another example is that if there is a mix of biased and unbiased consumers, a copay that is constant across consumers cannot deliver first-best utilization, motivating work to develop better-targeted interventions (BMS, p. 1657). Allcott, Lockwood, and Taubinsky ("ALT") provides yet another example of when it can be diagnostically useful to understand the bias mix in a population, as they find that if poor consumers are relatively more biased or more price-elastic, then inequality aversion does not necessarily push the optimal sin tax lower (p. 3, p. 12).[52]

B. Modeling diagnostics: How to do welfare analysis?

The empirical distribution of a behavioral summary statistic also affects how one should model welfare. Principally, meaningful heterogeneity in consumer-level bias should give one pause about using representative agent models (e.g., Chetty, Looney, and Kroft 2009; Gabaix 2014) and push one toward heterogenous-agent models like ALT, Taubinsky and Rees-Jones ("TR-J"), BMS, Farhi and Gabaix, and ours.

This sub-section discusses how one can use our empirical behavioral summary statistics to help choose among candidate heterogeneous-agent modeling approaches and then implement one's preferred approach. We start by describing empirical conditions that might lead one to use our empirical approach for welfare analysis—i.e. , a variant of our equation (1) or (4)—*instead* of a behavioral sufficient statistics approach. Then we discuss how our tools for measuring behavioral summary statistics can *complement* a behavioral sufficient statistics approach, by providing estimates of key sufficient statistic model inputs.

[52] Inequality aversion does of course push toward a lower tax via the redistributive motive; the key insight is that it also amplifies the internality-corrective (and hence tax-raising) motive when poor consumers are relatively more biased or more elastic.

i. Using our approach

Briefly recapping our framework and findings re: identification, our results suggest that correlations between outcome measures *Y* understood to capture utility aspects and a relatively low-dimensional *Bcount*, with the B-count measured twice per-consumer to help account for measurement error, can indeed identify the behavioral wedge needed for welfare analysis. This wedge can in turn be translated into money-metric units (consumer surplus in dollar terms), in settings where plausibly exogenous variation in income or wealth is available or a relevant aggregate demand curve is identified: in these cases *d(Y)/d(Bcount)* can be monetized by scaling it with the partial derivative of *Y* with respect to income, wealth, or prices.

As such our empirical approach to welfare analysis may, in some cases, be more technically feasible than behavioral sufficient statistic modeling. Our method requires obtaining data on the requisite outcomes and a handful of the relevant biases, together with a source of money-metric variation. Sufficient statistic methods require rich information on demand curves and normative choices that may only be obtainable with, e.g., within-subject price variation (TR-J), actionable estimates of the marginal social value of the regulated product (BMS), requisite data on experts' choices (ALT), and/or a fully debiasing intervention (Allcott and Taubinsky 2015; Chetty, Looney, and Kroft 2009).

Our approach may also be better-identified than sufficient statistic modeling; indeed one can use data like ours to check the identifying assumptions maintained by those models. One example is when a model requires a fully debiasing intervention to identify normative choice; this "pure nudge" assumption may not be valid if the intervention is not (fully) effective at debiasing consumers[53] or generates an overreaction (Bordalo, Gennaioli, and Shleifer 2019). Pure nudge assumptions can be examined using equations (2) and (3) above. A second example is using B-counts to check key assumptions required to use experts to identify normative choice. The expert approach requires not only obtaining requisite data (e.g., ALT's sample includes only 24 experts), but also that expert choices are unconfounded with unobserved heterogeneity and unbiased. Our

[53] BMS expresses skepticism: "Of course, it is implausible that a perfectly debiasing nudge exists…" (p. 1658).

results support the unconfoundedness assumption, but the unbiasedness assumption need not hold, as suggested by the findings in Linnainmaa, Melzer, and Previtero (forthcoming), where financial advisors follow their own recommendations and exhibit similarly multi-faceted and costly biases as their clients.[54] This issue further highlights the value of adding B-count elicitations to large nationally representative surveys, in this case to those surveys with occupation and/or other data that flags potential experts. Researchers could then assess whether experts are indeed unbiased or less biased before relying on experts to help identify normative choices.[55]

ii. Using B-counts to help implement behavioral sufficient statistics models

Our empirical approach also can be used to complement behavioral sufficient statistics modeling. First, as the above discussion details, one can use our approach to check and refine identifying assumptions, including the foundational assumption that multiple biases have reinforcing effects on behavior. Second, one can use our approach to help construct inputs to sufficient statistics models.

In a nutshell, if one has a way of identifying *who* is on the relevant margin(s), one can use our approach to help measure the required behavioral sufficient statistic moments of the relevant agents. To take one example, the TR-J method requires three sufficient statistics, with a B-count being helpful for estimating two of them: the average marginal bias and the variance of marginal consumers' bias. Moreover, our results offer some reassurance that one can estimate the latter directly (Section 3-D), rather than having to bound it due to concerns that measurement error will lead to an upward-biased estimate (TR-J Section 5). Another example is that one could use B-counts to help identify which marginal consumers are more biased than others, and then bound the

[54] The folk wisdom "Do as I say, not as I do" sounds another cautionary note for assuming that expert choices are unbiased, and suggests an alternative approach to identification: relying on expert recommendations (as I say) rather than assuming expert choices (as I do) reveal their preferences.

[55] Yet more examples of how B-counts can be used to examine modeling assumptions include using B-counts to check for the prevalence of biases hypothesized to be especially important (e.g., limited attention and memory as psychological foundations for Gabaix's models); to check whether bi-directional biases have the widely-hypothesized distributions, with expected directions (e.g, present-bias, under-estimating exponential growth, over-confidence, etc.) substantially more prevalent than less-expected ones (e.g., future-bias, over-estimating exponential growth, under-confidence, etc.); and to check whether average bias is indeed biased and not mean-zero, as some models require (e.g., Chetty, Looney, and Kroft 2009; Allcott and Taubinsky 2015).

size of the behavioral wedge by “comparing the demand curves of the more and less biased groups” (BMS , p. 1657).

9. Conclusion

An ardent classical economist might argue that behavioral tendencies are a collection of theoretically incoherent and/or empirically innocuous deviations from classical rationality. An ardent behavioral economist might argue that behavioral biases are important, but so multi-dimensional in how they affect decisions and welfare that a single consumer-level parameter could not hope to summarize them usefully. While strange bedfellows, these two might agree that seeking to measure and model a behavioral summary statistic is a fool’s errand.

In contrast, we find that behavioral economics can advance by capturing cross-consumer heterogeneity in overall behavioral tendencies using a single parameter. Specifically, we construct consumer-level behavioral summary statistics—B-counts—by aggregating information, within-person, across as many as 17 and as few as 2 potential sources of behavioral biases. We measure biases using streamlined, portable, and low-cost elicitations, and so our summary statistics are easy to measure, construct, and understand.

Our B-counts are strongly conditionally correlated with various outcomes, quite distinct empirically from measures of classical decision inputs and other covariates, and can be used to complement or substitute for behavioral sufficient statistic modeling approaches. One need not measure our full set of 17 potential behavioral biases to produce a valid and powerful behavioral summary statistic, and indeed it seems that measuring 2 suffices, at least when guided by theory as we are with our Narrow Sparsity B-count. Most fundamentally, our framework and results suggest that one can use directly measured summary statistics to identify consumer welfare loss associated with multiple, reinforcing behavioral biases, and that the welfare loss—the behavioral wedge between decision and normative utility—is substantial in magnitude.

We close by highlighting some opportunities for future research using our data and methods, by way of acknowledging some limitations of our work here. Our results linking B-counts to outcomes are probably better qualitative than quantitative estimates of the *total* consumer welfare loss from behavioral biases, given aggregation issues with both utility aspects and biases that remain to be sorted out. Consumer welfare loss is more clearly due to non-preference biases than preference ones, in our results as well as in theory, highlighting the need for more work on how to evaluate the welfare consequences of behavioral preferences. Consumer welfare loss may not equal social welfare loss, as behavioral biases can create opportunities for efficiency gains when there are market failures or redistributive motives (see Rees-Jones and Taubinsky forthcoming for a review), highlighting the need for more comprehensive welfare analysis. We use (partial) temporal stability in B-counts methodologically without addressing important substantive issues of how and why stability is incomplete: are behavioral summary statistics not fully trait-like, or is it more the case that summarizing behavioral tendencies is difficult to do with complete accuracy, or best done with reference to state-dependencies?

Relatedly, our initial guidance on efficient measurement highlights that more work is needed to optimize the mapping from a given set of elicitation data into a summary statistic. This may require further development of model selection techniques that account for measurement error.

More work is also needed on elicitation design. This can be achieved with experimentation. Although our elicitations are largely unincentivized on the margin—we elected to allocate more of our scarce research budget to measuring more variables for a larger sample size—one might obtain better power by trading off sample size and/or the number of biases elicited for marginal incentives on a smaller number of elicitations. These tradeoffs are worth exploring, especially given the informativeness of B-counts that are based on elicitations of only a handful—or even as few as two—potential sources of behavioral biases.

Our main takeaway for future work is that measuring a behavioral summary statistic can be a valuable and practical addition to many research designs concerned with consumer decision making and its implications.

REFERENCES

Afif, Zeina, W. Wade Islan, Oscar Calvo-Gonzalez, and Abigail Dalton. 2018. “Behavioral Science around the World: Profiles of 10 Countries.” World Bank Group: Mind, Behavior, and Development Unit.

Allcott, Hunt, Benjamin B. Lockwood, and Dmitry Taubinsky. forthcoming. “Regressive Sin Taxes, with an Application to the Optimal Soda Tax.” *Quarterly Journal of Economics*.

Allcott, Hunt, and Dmitry Taubinsky. 2015. “Evaluating Behaviorally Motivated Policy: Experimental Evidence from the Lightbulb Market.” *American Economic Review* 105 (8): 2501–38.

Andreoni, James, and Charles Sprenger. 2012. “Estimating Time Preferences from Convex Budgets.” *The American Economic Review* 102 (7): 3333–56.

Azrieli, Yaron, Christopher P. Chambers, and Paul J. Healy. 2018. “Incentives in Experiments: A Theoretical Analysis.” *Journal of Political Economy* 126 (4): 1472–1503.

Baicker, Katherine, Sendhil Mullainathan, and Joshua Schwartzstein. 2015. “Behavioral Hazard in Health Insurance.” *The Quarterly Journal of Economics* 130 (4): 1623–67.

Banks, J., and Z. Oldfield. 2007. “Understanding Pensions: Cognitive Function, Numerical Ability, and Retirement Saving.” *Fiscal Studies* 28 (2): 143–70.

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.

Benjamin, Daniel, Sebastian Brown, and Jesse Shapiro. 2013. “Who Is ‘Behavioral’? Cognitive Ability and Anomalous Preferences.” *Journal of the European Economic Association* 11 (6): 1231–55.

Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Alex Rees-Jones. 2014. “Can Marginal Rates of Substitution Be Inferred from Happiness Data? Evidence from Residency Choices.” *American Economic Review* 104 (11): 3498–3528.

Benjamin, Daniel J., Ori Heffetz, Miles S. Kimball, and Nichole Szembrot. 2014. “Beyond Happiness and Satisfaction: Toward Well-Being Indices Based on Stated Preference.” *American Economic Review* 104 (9): 2698–2735.

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.

Bernheim, B. Douglas, and Dmitry Taubinsky. 2018. “Behavioral Public Economics.” In *Handbook of Behavioral Economics: Applications and Foundations 1*, edited by B. Douglas Bernheim, Stefano DellaVigna, and David Laibson, 1:381–516. North-Holland. https://doi.org/10.1016/bs.hesbe.2018.07.002.

Black, Dan, Mark C. Berger, and Frank Scott. 2000. “Bounding Parameter Estimates With Non-Classical Measurement Error.” *Journal of The American Statistical Association* 95 (451): 739–48.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2019. “Memory, Attention, and Choice.”

Burks, S. V., J. P. Carpenter, L. Goette, and A. Rustichini. 2009. "Cognitive Skills Affect Economic Preferences, Strategic Behavior, and Job Attachment." *Proceedings of the National Academy of Sciences* 106 (19): 7745–50.

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.

Campbell, John. 2016. "Restoring Rational Choice: The Challenge of Consumer Financial Regulation." *American Economic Review* 106 (5): 1–30.

Cesarini, David, Magnus Johannesson, Patrik K. E. Magnusson, and Björn Wallace. 2012. "The Behavioral Genetics of Behavioral Anomalies." *Management Science* 58 (1): 21–34.

Chapman, Jonathan, Mark Dean, Pietro Ortoleva, Erik Snowberg, and Colin Camerer. 2018a. "Econographics."

———. 2018b. "Willingness to Pay and Willingness to Accept Are Probably Less Correlated than You Think."

Chetty, Raj. 2009. "Sufficient Statistics for Welfare Analysis: A Bridge Between Structural and Reduced-Form Methods." *Annual Review of Economics* 1 (1): 451–88.

———. 2015. "Behavioral Economics and Public Policy: A Pragmatic Perspective." *American Economic Review* 105 (5): 1–33.

Chetty, Raj, Adam Looney, and Kory Kroft. 2009. "Salience and Taxation: Theory and Evidence." *The American Economic Review* 99 (4): 1145–77.

Choi, Syngjoo, Shachar Kariv, Wieland Müller, and Dan Silverman. 2014. "Who Is (More) Rational?" *American Economic Review* 104 (6): 1518–50.

Dean, Mark, and Pietro Ortoleva. 2018. "Is It All Connected? A Testing Ground for Unified Theories of Behavioral Economics Phenomena."

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.

———. 2018. "On the Relationship between Cognitive Ability and Risk Preference." *Journal of Economic Perspectives* 32 (2): 115–34. https://doi.org/10.1257/jep.32.2.115.

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.

Ericson, Keith. 2011. "Forgetting We Forget: Overconfidence and Memory." *Journal of the European Economic Association* 9 (1): 43–60.

Farhi, Emmanuel, and Xavier Gabaix. 2018. "Optimal Taxation with Behavioral Agents."

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.

Frederick, Shane. 2005. "Cognitive Reflection and Decision Making." *Journal of Economic Perspectives* 19 (4): 25–42.

Fuller, Wayne A. 2009. *Measurement Error Models*. Vol. 305. John Wiley & Sons.

Gabaix, Xavier. 2014. “A Sparsity-Based Model of Bounded Rationality.” *The Quarterly Journal of Economics* 129 (4): 1661–1710.

———. 2019. “Behavioral Inattention.” In *Handbook of Behavioral Economics- Foundations and Applications 2, Volume 2*, edited by Douglas Bernheim, Stefano DellaVigna, and David Laibson. Elsevier.

Gillen, Ben, Erik Snowberg, and Leeat Yariv. forthcoming. “Experimenting with Measurement Error: Techniques with Applications to the Caltech Cohort Study.” *Journal of Political Economy*.

Gneezy, Uri, Alex Imas, and John List. 2015. “Estimating Individual Ambiguity Aversion: A Simple Approach.”

Güntner, Anna, Konstantin Lucks, and Julia Sperling-Magro. 2019. “Lessons from the Front Lines of Corporate Nudging.” *McKinsey Quarterly*, 2019.

Kuhnen, Camelia M., and Brian T. Melzer. 2018. “Noncognitive Abilities and Financial Delinquency: The Role of Self-Efficacy in Avoiding Financial Distress.” *The Journal of Finance* 73 (6): 2837–69.

Levy, Matthew, and Joshua Tasoff. 2016. “Exponential-Growth Bias and Lifecycle Consumption.” *Journal of the European Economic Association* 14 (3): 545–83.

Li, Ye, Martine Baldassi, Eric J. Johnson, and Elke U. Weber. 2013. “Complementary Cognitive Capabilities, Economic Decision Making, and Aging.” *Psychology and Aging* 28 (3): 595–613.

Linnainmaa, Juhani, Brian Melzer, and Alessandro Previtero. forthcoming. “The Misguided Beliefs of Financial Advisors.” *Journal of Finance*.

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.

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, Stephan, and Charles D. Sprenger. 2015. “Temporal Stability of Time Preferences.” *Review of Economics and Statistics* 97 (2): 273–86.

Montiel Olea, J. L., and T. Strzalecki. 2014. “Axiomatization and Measurement of Quasi-Hyperbolic Discounting.” *The Quarterly Journal of Economics* 129 (3): 1449–99.

Moore, Don A., and Paul J. Healy. 2008. “The Trouble with Overconfidence.” *Psychological Review* 115 (2): 502–17.

Mullainathan, Sendhil, Joshua Schwartzstein, and William J. Congdon. 2012. “A Reduced-Form Approach to Behavioral Public Finance.” *Annual Review of Economics* 4 (1): 511–40.

Poterba, James M. 2014. “Retirement Security in an Aging Population.” *American Economic Review* 104 (5): 1–30.

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.

Rees-Jones, Alex, and Dmitry Taubinsky. forthcoming. “Measuring ‘Schmeduling.’” *Review of Economic Studies*.

Stango, Victor, and Jonathan Zinman. 2009. “Exponential Growth Bias and Household Finance.” *The Journal of Finance* 64 (6): 2807–49.

Taubinsky, Dmitry, and Alex Rees-Jones. 2018. “Attention Variation and Welfare: Theory and Evidence from a Tax Salience Experiment.” *Review of Economic Studies* 85: 2462–96.

Von Gaudecker, Hans-Martin, Arthur Van Soest, and Erik Wengström. 2011. “Heterogeneity in Risky Choice Behavior in a Broad Population.” *The American Economic Review* 101 (2): 664–94.

Figure 1. Uncorrected estimate of B-count variance does seem to reflect true heterogeneity

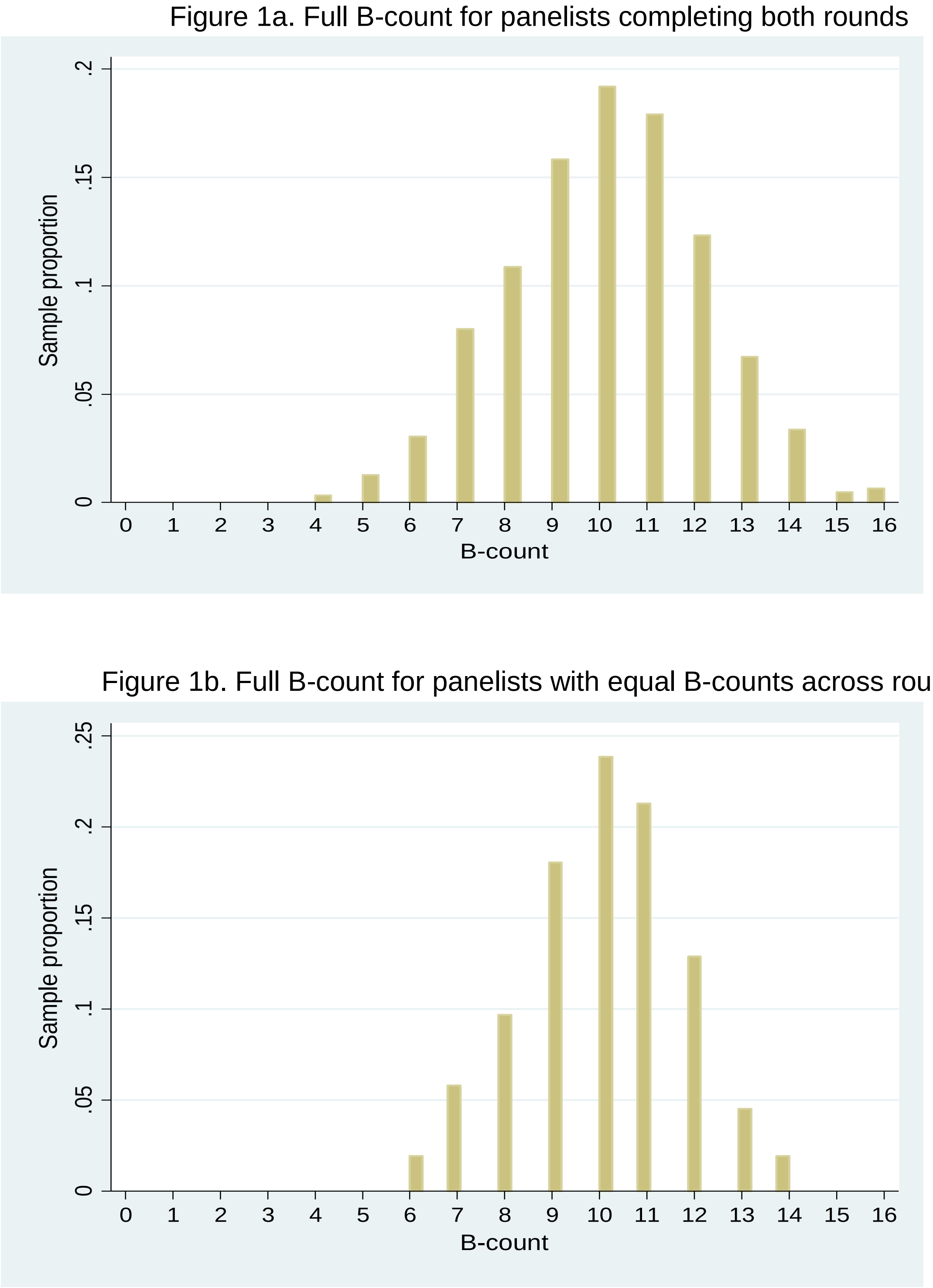


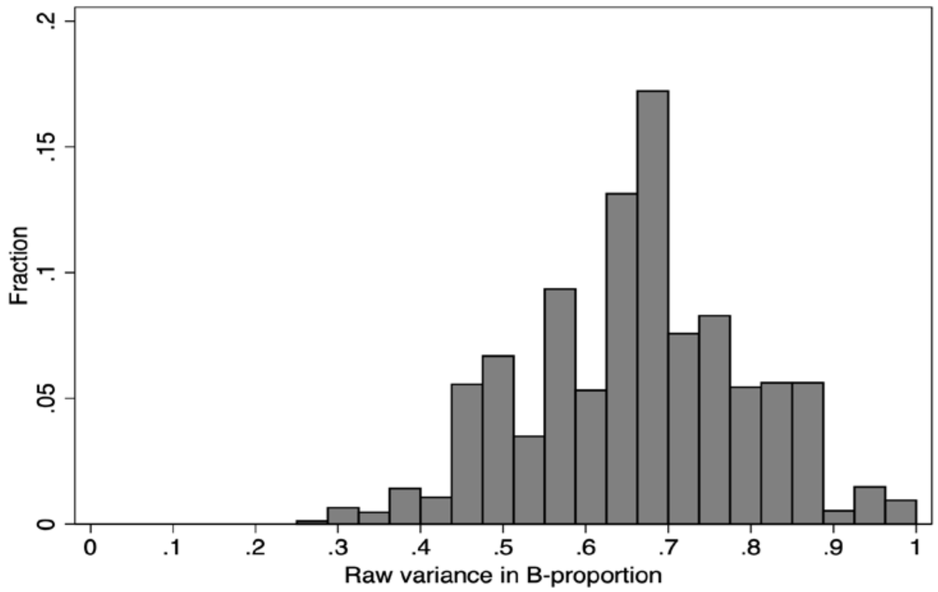
Figure 1a. Full B-count for panelists completing both rounds

Figure 1b. Full B-count for panelists with equal B-counts across rounds

We omit panelists with missing data on 2 or more of our 17 potential sources of behavioral biases, to mitigate spurious variance from variance in missingness. This leaves sample sizes of 625 individuals in the top panel and 125 in the bottom.

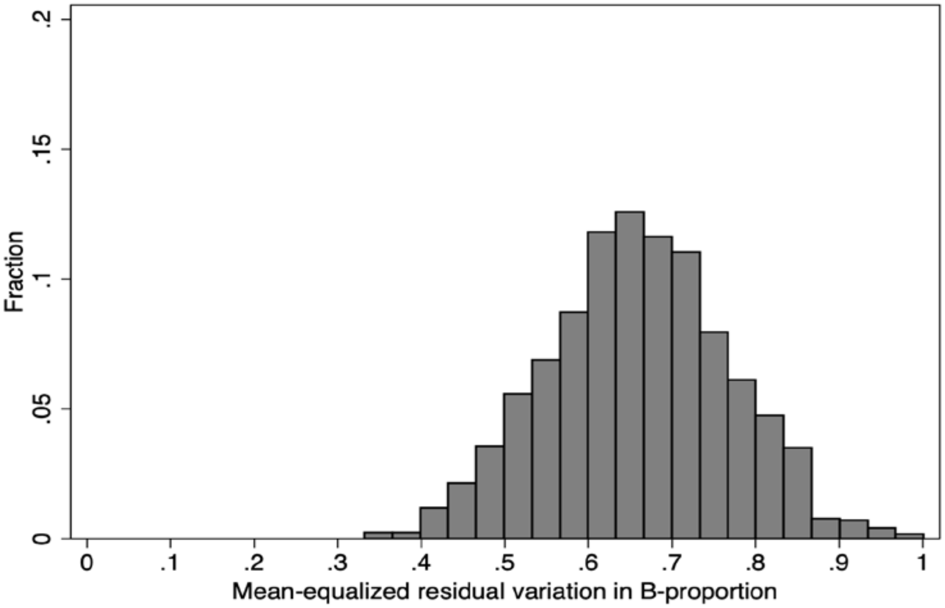
Figure 2. The Full B-count is poorly explained by other covariates

Figure 2a. B-count unconditional variation



Note: Y-axis shows sample proportion. On X-axis, B-proportion is the ratio of biases exhibited to non-missing biases. We use B-proportion instead of level B-count for comparability to Figure 1b. Interquartile range here is [0.56, 0.75] and 5th/95th percentiles are [0.44, 0.88].

Figure 2b. B-count residual variation



Note: Y-axis shows sample proportion. X-axis shows distribution of residuals from regression of B-proportion on full set of covariates (Table 8 Column 1 reports R-squareds). We use B-prop instead of level B-count to avoid overfitting. Mean of residuals is set equal to the mean of the B-proportion from Figure 1, for comparability. Interquartile range here is [0.59, 0.74] and 5th/95th percentiles are [0.47, 0.85].

Figure 3. Behavioral summary statistics are distinct from other decision inputs: B-count variation within- and across- key sub-groups

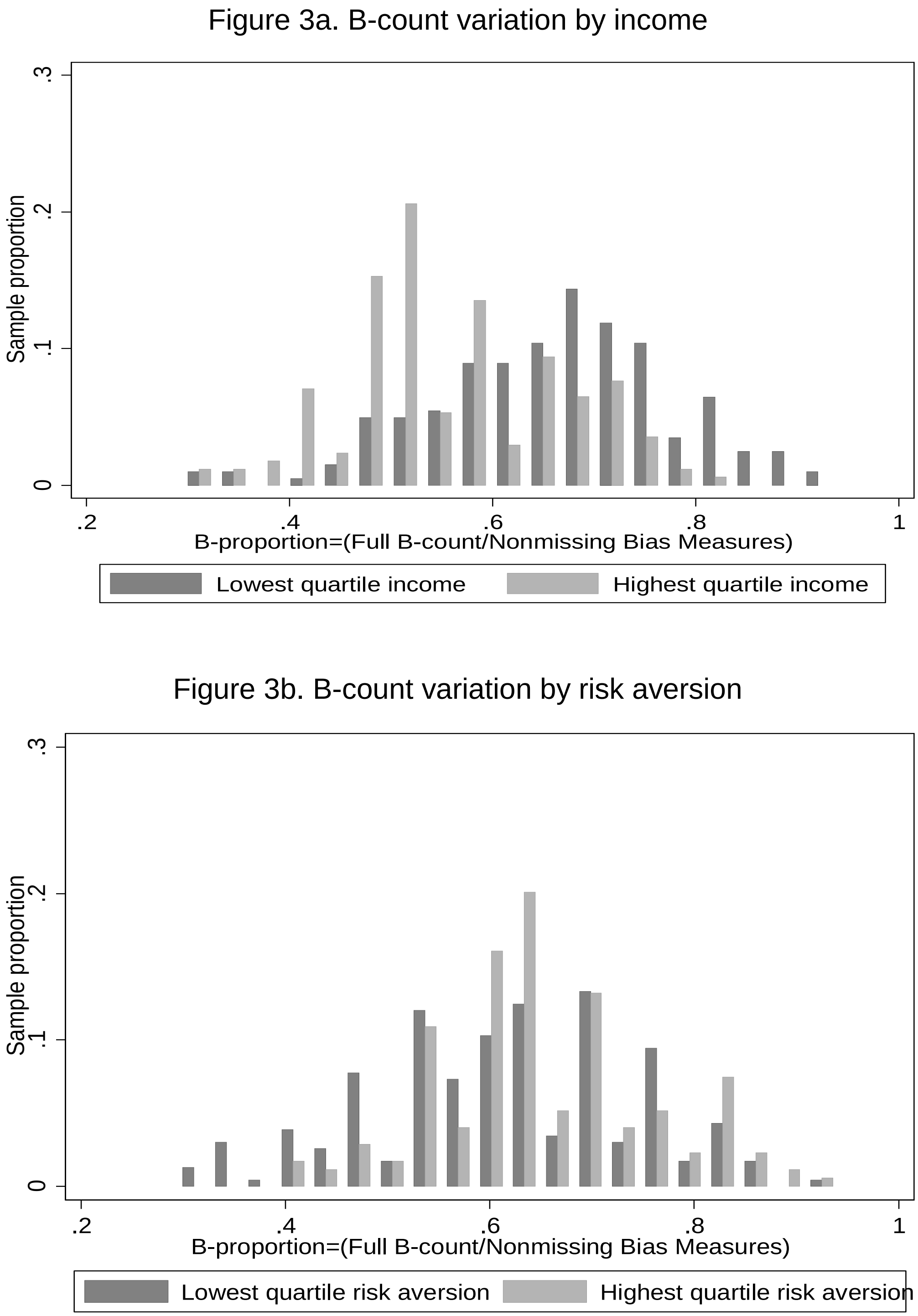


Figure 3a. B-count variation by income

Figure 3b. B-count variation by risk aversion

(Notes at bottom of next page, following Figure 3d.)

Figure 3c. B-count variation by education

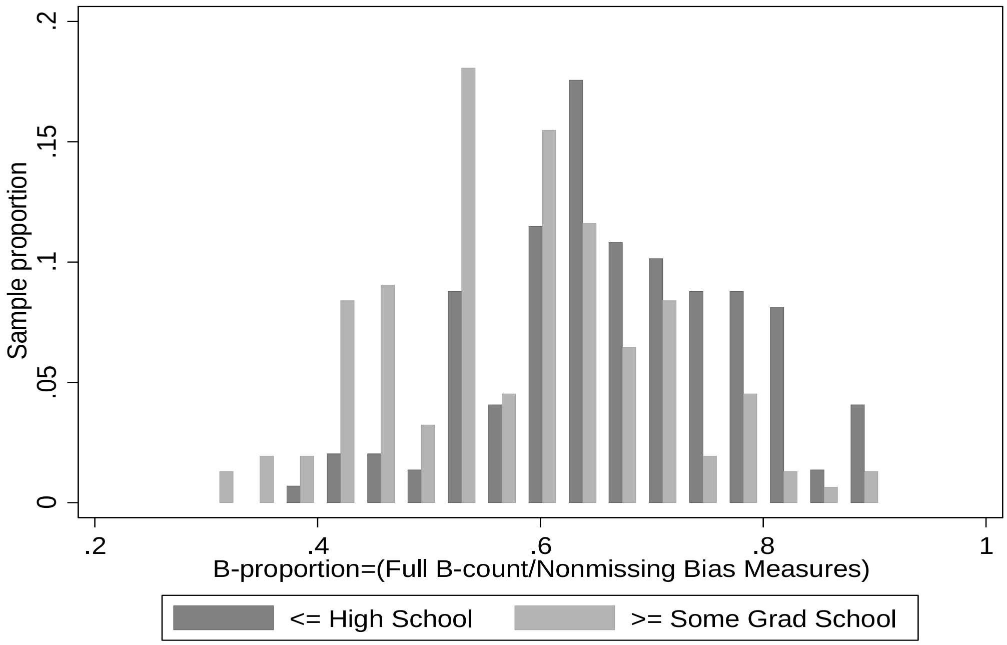
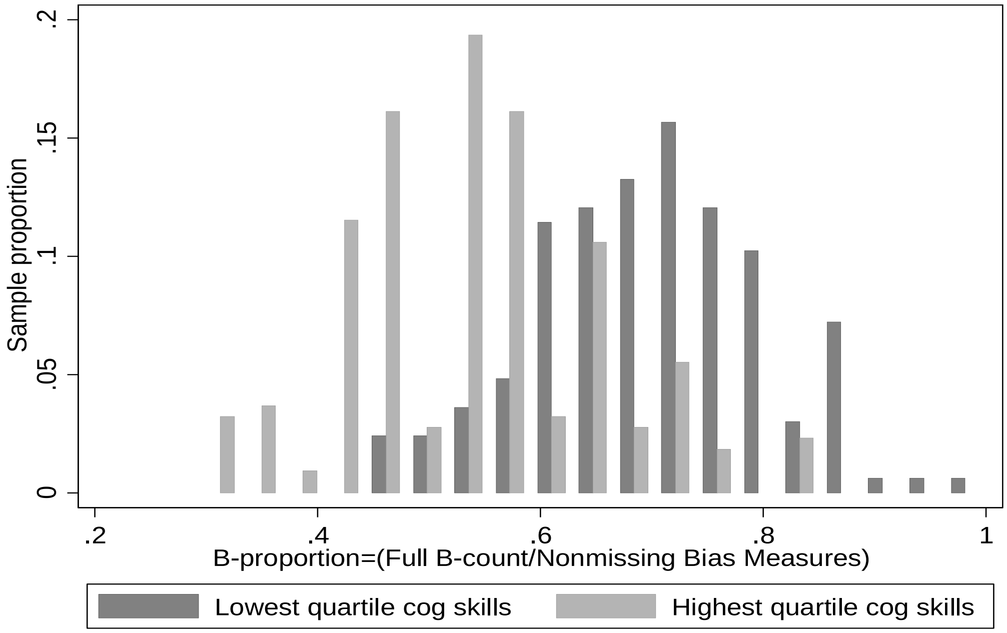


Figure 3d. B-count variation by cognitive skills



Note: Round 1 data only. On X-axis, we use B-proportion instead of level B-count to allow for item non-response to vary across sub-groups. Cognitive skills measured here with the 1st principal component of our four test scores. Risk aversion measured here with the 1st principal component of the Dohmen et al and Barsky et al measures. See Appendix Table 1 and Data Appendix Section 2 for details on individual test score and risk aversion measures.

Figure 4. B-sub-counts based on random draws of a handful of biases reproduce our main result on the Full B-count

Figure 4a. Distributions of coefficients for randomly constructed B-sub-counts

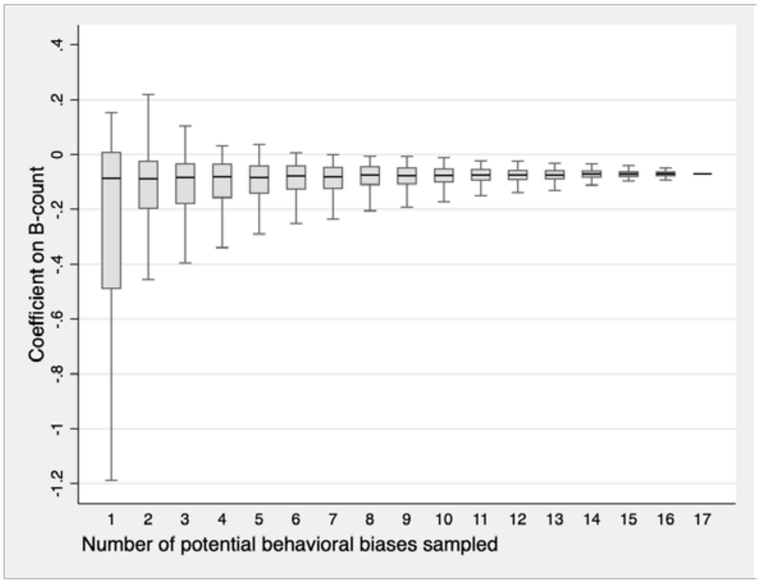
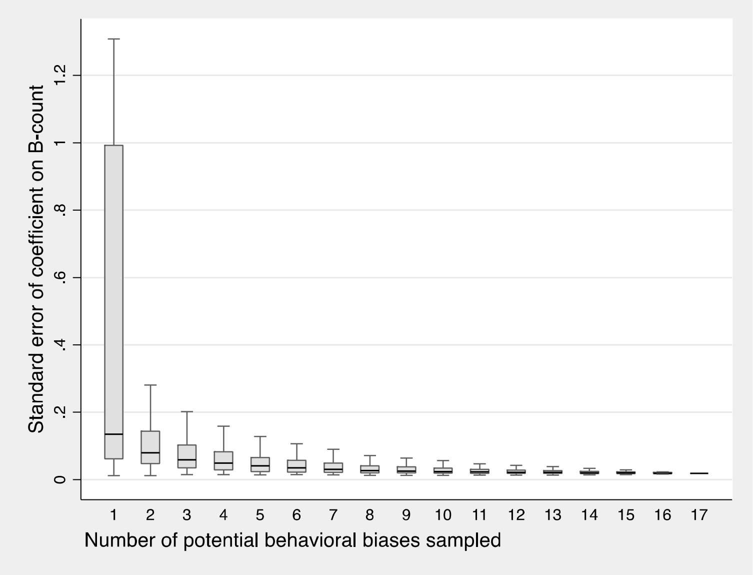


Figure 4b. Distributions of standard errors for randomly constructed B-sub-counts



Note: We randomly draw up to 2,500 bias combinations for each j (see Section 7 for details), from the full set of potential biases described in Table 1, and estimate the specification used in Table 4 Column 2 on each draw. Top and bottom whiskers show 95th/5th percentiles, top and bottom of box show IQR, and solid line within box shows median.

Table 1. Research design: Eliciting data on multiple behavioral biases, and defining bias indicators.

			Groupings for B-sub-counts		
Potential source of bias: *key antecedents*	Elicitation method description	Behavioral indicator(s), **"expected" deviation direction in bold**	**Math?**	**Preference?**	**Sparsity?**
(1)	(2)	(3)	(4)	(5)	(6)
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.	**Present-biased: discounts more when sooner date is today** Future-biased: discounts more when sooner date is 5 weeks from tdy	No	Yes	Broad
Time inconsistent discounting of money: *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.	**Present-biased: choose less healthy now, healthy 5 weeks from now** Future-biased: choose healthy now, less healthy 5 weeks from now	No	Yes	Broad
Violates GARP (with 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.	**Violates GARP: potential earnings wasted per CCEI>0** **Violates GARP and dominance avoidance: potential earnings wasted per combined-CCEI>0**	No	Yes	No
Certainty premium: *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, q ϵ[0.1, 1.0], p=0.5 on one screen and 1.0 on the other.	**Preference for cetainty: certainty premium (CP) >0** Cumulative prospect theory: certainty premium (CP)<0	No	Yes	No
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.	**Loss aversion: choosing the certain $0 payoff in one or more choices.**	No	Yes	No
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.	**Narrow-bracketing: making a choice that is dominated given implications of an earlier decision, on one or both tasks.**	No	No	No
Ambiguity aversion: *Dimmock et al. (forthcoming)*	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.	**Ambiguity Aversion: prefers bag with 45 green to bag with unknown mix.**	No	Yes	No
(Over-)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?"	**Overconfidence in perform: self-assessment > actual score** Underconfidence in perform: self-assessment < actual score	No	No	No
(Over-)confidence 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?"	**Greater diff between self-assessed and actual rank indicates more overconfidence. "Overconfident" = overconfidence above median.**	No	No	No
Overconfidence in precision: *Larrick et al (2007), Moore & Healy (2008)*	Questions about about likelihoods of different numeracy quiz scores and future income increases.	**Overconfidence in precision: responds 100% to one or both questions**	No	No	No
Non-belief in the law of large numbers (NBLLN): *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]. NBLLN = distance between response for [481, 519] and 78.	Overestimates convergence to 50-50: responds with>78% **Underestimates convergence to 50-50: responds with<78%**	Yes	No	Broad
Gambler's or hot-hand fallacy: *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?"	**Hot-hand fallacy: responds with>50%** Gambler's fallacy: responds with<50%	Yes	No	Broad
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 APR). Then elicits perceived APR implied by that payment.	**Underestimates EG: actual APR>perceived APR** Overestimates EG: actual APR<perceived APR	Yes	No	Broad
Exponential growth bias (EGB), asset-side: *Banks et al (2007)*	Elicits perceived future value of $200, earning 10% annual, after two years.	**Underestimates EG:** perceived FV<actual FV=$242 Overestimates EG: perceived FV>actual FV=$242	Yes	No	Broad
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.	**Limited attention: Indicates regret about paying too little attention given opportunity cost of attention, on one or more of the four questions**	No	No	Narrow
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."	**Limited memory: Says will complete task but does not complete.**	No	No	Narrow

The Data Appendix Section 1 provides additional details on measuring each behavioral bias."pp" = percentage points. "CCEI" = Critical Cost Efficiency Index. The "Full" B-count sums all indicators in column (3). "Expected" deviation direction, for bi-directional B-factors, is the direction typically theorized/observed in prior work. Sparsity biases are per Gabaix and discussed in Section 3-B. Both "Narrow" sparsity biases are also counted in the "Broad" sparsity B-sub-count.

Table 2. B-count descriptive statistics

	Round-by-round					Using both rounds	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	Mean	SD	Share>0	Max (possible)	Median mins survey time required	Correlation w/full B-count	Correlation (Rd1, Rd2)
Panel A: Full B-count							
Round 1 (N=1427)	10.05	2.17	1.00	16 (17)	34.43	1.00	0.44
Round 1, in Round 2 (N=845)	10.08	2.02	1.00	16 (17)	34.02		
Round 2 (N=845)	9.84	2.22	1.00	16 (17)	34.48		
Panel B: Sparsity B-counts							
Narrow: Limited attention/memory only							
Round 1	1.29	0.64	0.90	2 (2)	1.35	0.39	0.27
Round 1, in Round 2	1.24	0.65	0.88	2 (2)	1.37		
Round 2	1.15	0.68	0.84	2 (2)	1.33		
Broad: Limited attention/memory, present-biases, price misperception and statistical biases							
Round 1	4.31	1.34	1.00	8 (8)	16.50	0.69	0.38
Round 1, in Round 2	4.24	1.31	1.00	8 (8)	16.47		
Round 2	4.00	1.33	1.00	8 (8)	15.98		
Panel C: Other B-sub-counts							
Expected Direction biases							
Round 1	8.60	2.16	1.00	15 (17)	34.43	0.87	0.41
Round 1, in Round 2	8.59	2.03	1.00	15 (17)	34.02		
Round 2	8.32	2.13	1.00	14 (17)	34.48		
Non-expected Direction biases							
Round 1	1.45	1.04	0.81	5 (8)	18.20	0.26	0.18
Round 1, in Round 2	1.49	1.06	0.82	5 (8)	18.18		
Round 2	1.52	1.02	0.84	5 (8)	17.82		
Math							
Round 1	2.63	0.92	0.99	4 (4)	3.80	0.57	0.44
Round 1, in Round 2	2.61	0.88	1.00	4 (4)	3.97		
Round 2	2.44	0.91	0.99	4 (4)	4.00		
Non-math							
Round 1	7.42	1.76	1.00	12 (13)	29.62	0.90	0.32
Round 1, in Round 2	7.47	1.67	1.00	12 (13)	29.45		
Round 2	7.41	1.82	1.00	13 (13)	29.22		
Preferences							
Round 1	4.18	1.29	1.00	7 (7)	23.50	0.51	0.23
Round 1, in Round 2	4.27	1.23	1.00	7 (7)	23.47		
Round 2	4.29	1.28	1.00	7 (7)	23.08		
Non-preferences							
Round 1	5.87	1.76	1.00	10 (10)	9.77	0.82	0.49
Round 1, in Round 2	5.81	1.69	1.00	10 (10)	9.88		
Round 2	5.56	1.78	1.00	10 (10)	9.55		
Panel D. Count of missing inputs to B-counts							
Round 1	1.00	1.71	0.49	12 (17)	n/a	-0.33	0.36
Round 1, in Round 2	0.72	1.14	0.43	8 (17)			
Round 2	0.97	1.75	0.47	11 (17)			

Our data consist of two survey rounds, of two modules each, conducted 3 years apart. We include only those panelists who took both modules in Round 1 (N=1427) or all four modules across both rounds (N=845). B-count and B-sub-count definitions are summarized in Table 1 and discussed in Sections 3-A and -B. Column 5 reports median panelist time spent on questions/tasks used to measure the inputs to the B-count in that row. Round-to-round correlations for B-counts (Panels A-C) adjust for missing data by conditioning on the count of missing bias measures in each survey round.

Table 3. Measuring financial condition and subjective well-being: Our main outcome measures

	Data used					Mean	SD	Pairwise correlation					
	# of	Median mins	From	From	# panelists	(All		Financial condition		Other measures of subjective well-being			
	questions	survey time	our	other	with	rescaled to		Objective	Subjective	Life	Life satis	Happiness	Health
	per module	required	modules?	modules?	nonmissing	[0,1])		index	index	satisfaction	index	index	status
Objective financial condition index	12	2.67	yes	no	845	0.53	0.34	1.00					
Subjective financial condition index	4	0.97	yes	no	845	0.50	0.25	0.57	1.00				
Life satisfaction	1	0.19	yes	no	844	0.68	0.23	0.35	0.50	1.00			
Life satisfaction index	1	0.19	no	yes	809	0.64	0.21	0.33	0.43	0.64	1.00		
Happiness index	1	<1.0	no	yes	787	0.70	0.23	0.29	0.33	0.51	0.57	1.00	
Health status	1	<0.5	no	yes	840	0.61	0.22	0.31	0.29	0.32	0.45	0.37	1.00

Unit of observation is the individual respondent, with multiple observations per respondent averaged across survey rounds (for variables in our modules) or across other ALP modules (for variables we merge in from other ALP modules). Other ALP modules used here are all administered *between* our survey rounds; we could not find relevant data collected in modules adminstered after or during our second round. As in most of our main tables, we limit the sample frame here to panelists who completed both of our survey rounds (N=845). Correlations estimated using the two-step "polychoric" procedure in Stata.

Variable definitions: Each variable is scaled so that higher values indicate better financial condition and/or subjective well-being. Each measure here is scaled or rescaled to [0, 1] for comparability. Indexes simply take the unweighted mean of non-missing index components. See Appendix Table 3 for details on index components.

Objective financial condition index is comprised of indicators of postive net worth, positive retirement assets, holding equities, having a positive savings rate over the prior 12 months, and not having severe financial hardship during the prior 12 months.

Subjective financial condition index is comprised of measures of financial satisfaction, retirement savings adequacy, non-retirement savings adequacy, and lack of financial stress.

Life satisfaction is measured using one of three minor variants on the standard "… how satisfied are you with your life as a whole these days?" asked in many surveys worldwide. For the other-module measure, we take the within-panelist average of non-missing responses to this question across the six ALP modules in which it has appeared subsequent to our round 1 modules, as of this writing. Of the 809/845 panelists with at least one non-missing response, 640 have at least two.

Happiness is measured by taking the within-panelist average of responses to two standard questions on happiness in general and in the last 30 days. These are asked in five other ALP modules subsequent to our Round 1 modules, with 787 of our 845 panelists completing at least one of these happiness questions and 397 completing both the 30-day version and the in-general-version.

Health status is from the standard question: “Would you say your health is excellent, very good, good, fair, or poor?”. We take the within-panelist average across eight different modules in which this question has appeared subsequent to our Round 1 modules. Of the 840/845 panelists completing at least one of these, 780 complete more than one.

We lack timings data on happiness and health status questions because they do not appear in our modules, and so we estimate the time required to elicit these measures, roughly, based on questions of similar length and difficulty in our modules.

Table 4. B-counts are strongly conditionally correlated with financial outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
LHS=Financial outcome index	*Objective*	*Subjective*	*Subjective*	*Objective*	*Subjective*	*Subjective*	*Objective*	*Subjective*	*Subjective*
B-count: Full	-0.061***	-0.084***	-0.064***						
	(0.018)	(0.018)	(0.016)						
B-count: Sparsity Broad				-0.090***	-0.128***	-0.097***			
				(0.028)	(0.028)	(0.024)			
B-count: Sparsity Narrow							-0.236***	-0.328***	-0.256***
							(0.063)	(0.065)	(0.057)
Objective financial index			0.340***			0.341***			0.304***
			(0.026)			(0.026)			(0.031)
d(LHS)/d(1 SD B-count)	-0.130	-0.179	-0.135	-0.119	-0.169	-0.129	-0.157	-0.218	-0.170
d(LHS)/d(1 SD objective financial index)			0.117			0.117			0.104
mean(LHS)	0.531	0.504	0.504	0.531	0.504	0.504	0.531	0.504	0.504
N panelists	843	843	843	843	843	843	843	843	843
N with replicates	3370	3370	3370	3370	3370	3370	3370	3370	3370

* 0.10 ** 0.05 *** 0.01. Standard errors, clustered on panelist, in parentheses. Each column presents results from a single pooled Obviously Related Instrumental Variables regression (equation 3 in the text) of the LHS variable described in the column label on the variables described in the row labels + the complete set of covariates described in Appendix Table 1. Table 1 provides details on our B-count variable definitions; higher values indicate more behavioral biases. Table 3 provides details on our LHS variable definitions; higher values indicate better financial condition.

Table 5. Identifying relationships between outcomes and B-counts: Insensitivity to covariate specification

LHS=Subjective financial index	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A. Full B-Count									
Full B-count	-0.084***	-0.089***	-0.086***	-0.074***	-0.079***	-0.078***	-0.073***	-0.090***	-0.068***
	(0.018)	(0.014)	(0.017)	(0.015)	(0.018)	(0.009)	(0.019)	(0.027)	(0.017)
dY/d(1 SD B-count)	-0.179	-0.189	-0.183	-0.157	-0.168	-0.167	-0.162	-0.186	-0.163
Panel B. Broad Sparsity B-count									
Sparsity biases: attention+	-0.128***	-0.135***	-0.132***	-0.116***	-0.126***	-0.134***	-0.115***	-0.142***	-0.116***
	(0.028)	(0.023)	(0.028)	(0.024)	(0.027)	(0.016)	(0.029)	(0.039)	(0.031)
dY/d(1 SD B-count)	-0.169	-0.179	-0.175	-0.154	-0.168	-0.178	-0.152	-0.188	-0.153
Panel C. Narrow Sparsity B-count									
Sparsity biases: attention only	-0.328***	-0.274***	-0.331***	-0.326***	-0.320***	-0.292***	-0.333***	-0.325***	-0.329***
	(0.065)	(0.047)	(0.065)	(0.065)	(0.061)	(0.046)	(0.083)	(0.067)	(0.080)
dY/d(1 SD B-count)	-0.218	-0.182	-0.220	-0.217	-0.212	-0.194	-0.221	-0.216	-0.218
Missing bias count included?	yes	yes	yes	yes	yes	yes	yes	yes	yes
Demographics included?	yes	no	yes	yes	yes	no	yes	yes	yes
Classical preferences included?	yes	yes	no	yes	yes	no	yes	yes	yes
Cognitive skills included?	yes	yes	yes	no	yes	no	yes	yes	yes
Non-cognitive skills included?	yes	yes	yes	yes	no	no	yes	yes	yes
Survey time spent deciles included?	yes	yes	yes	yes	yes	no	yes	yes	yes
ORIV for B-count?	yes	yes	yes	yes	yes	yes	yes	yes	yes
ORIV for classical preferences?	no	no	no	no	no	no	yes	no	yes
ORIV for cognitive skills?	no	no	no	no	no	no	no	yes	yes
mean(LHS)	0.504	0.505	0.504	0.504	0.504	0.505	0.504	0.504	0.504
N panelists	843	843	843	843	843	843	843	843	843
N with replicates	3370	3370	3370	3370	3370	3370	3370	3370	3370

* 0.10 ** 0.05 *** 0.01. Standard errors, clustered on panelist, in parentheses. Each panel-column presents results from a single ORIV regression of our subjective financial index on the B-count described in the Panel title and row label and the other covariates described in rows at the bottom of the table. N.B. Column 1 here reproduces results from our main specifications in Table 4 (Columns 2, 5, and 8 in Table 4).

Table 6. Identifying relationships between outcomes and B-counts: Full B-count decompositions

	(1)	(2)	(3)	(4)	(5)	(6)
LHS=Financial condition index	*Objective*	*Subjective*	*Objective*	*Subjective*	*Objective*	*Subjective*
Math biases (1)	-0.011	-0.014				
	(0.044)	(0.041)				
Non-math biases (2)	-0.083***	-0.115***				
	(0.029)	(0.030)				
Expected biases (1)			-0.061***	-0.086***		
			(0.019)	(0.020)		
Non-expected biases (2)			-0.038	-0.116		
			(0.073)	(0.073)		
Preference biases (1)					-0.008	-0.069*
					(0.036)	(0.035)
Non-preference biases (2)					-0.082***	-0.090***
					(0.019)	(0.018)
pval (1)=(2)	0.240	0.091	0.715	0.624	0.050	0.553
d(LHS)/d(1 SD B-count(1))	-0.010	-0.013	-0.126	-0.178	-0.010	-0.087
d(LHS)/d(1 SD B-count(2))	-0.145	-0.201	-0.040	-0.120	-0.143	-0.157
mean(LHS)	0.531	0.504	0.531	0.504	0.531	0.504
N panelists	843	843	843	843	843	843
N	3370	3370	3370	3370	3370	3370

* 0.10 ** 0.05 *** 0.01. Same specification as Table 4, but with the Full B-count decomposed into the B-sub-count couplets described in the row labels.

Table 7. B-count conditional correlations with measures of utility aspects

	(1)	(2)	(3)	(4)	(5)
Utility aspect measures from	Our modules		Other modules		
LHS = Utility aspect variable	*Fin index*	*Life Satisfaction*	*Life Satisfaction*	*Happiness Index*	*Self-assess Health*
Panel A. Full B-Count					
B-count: All biases	-0.084***	-0.009	-0.004	-0.010	0.002
	(0.018)	(0.013)	(0.014)	(0.016)	(0.013)
d(LHS)/d(1 SD B-count)	-0.179	-0.019	-0.008	-0.021	0.004
Panel B. Sparsity Broad B-count					
B-count: Sparsity biases attention+	-0.128***	-0.071***	-0.028	-0.049**	-0.042*
	(0.028)	(0.022)	(0.021)	(0.025)	(0.022)
d(LHS)/d(1 SD B-count)	-0.169	-0.094	-0.038	-0.066	-0.055
Panel C. Sparsity Narrow B-count					
B-count: Sparsity biases attention only	-0.328***	-0.099**	-0.093**	-0.136**	-0.100**
	(0.065)	(0.041)	(0.041)	(0.056)	(0.043)
d(LHS)/d(1 SD B-count)	-0.218	-0.066	-0.062	-0.090	-0.066
mean(LHS)	0.504	0.679	0.643	0.703	0.607
N	3370	3366	3226	3138	3350

* 0.10 ** 0.05 *** 0.01. Standard errors, clustered on panelist, in parentheses. Each panel-column presents results from a single Obviously Related Instrumental Variables regression of the LHS variable described in the column label on the variable(s) described in the row label(s) + the complete set of covariates described in Appendix Table 1. Table 1 provides details on our B-count variable definitions. Table 3 provides details on our LHS variable definitions.

Table 8. Distinctness: B-counts are not well-explained by other covariates

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	
LHS=B-count proportion numerator	*Full*	*Sparsity Narrow*	*Sparsity Broad*	*Math*	*Non-Math*	*Expected*	*Non-expected*	*Preference*	*Non-preference*	N
R-squared: All covariates in Appendix Table 1	0.33	0.12	0.25	0.34	0.21	0.25	0.13	0.13	0.40	1690
Partial R-squared: demographics	0.19	0.06	0.16	0.24	0.09	0.13	0.06	0.03	0.24	1690
Partial R-squared: cognitive skills	0.24	0.02	0.16	0.25	0.12	0.17	0.05	0.01	0.33	1690
Partial R-squared: noncognitive skills	0.02	0.02	0.03	0.01	0.02	0.02	0.00	0.01	0.02	845
Partial R-squared: classical preferences	0.04	0.00	0.03	0.02	0.04	0.03	0.02	0.06	0.03	1690
Partial R-squared: state of residence	0.05	0.05	0.06	0.07	0.04	0.05	0.07	0.04	0.06	1690
Partial R-squared: time spent on behavioral q's deciles	0.00	0.01	0.01	0.01	0.00	0.01	0.01	0.00	0.01	1690
mean(LHS)	0.62	0.62	0.54	0.68	0.60	0.53	0.21	0.65	0.60	1690

LHS variable is a proportion: a B-count scaled by the count of its potential behavioral biases with nonmissing data (i.e., by the maximum possible B-count one could observe for that B-count). Each cell presents results from a single OLS regression, using the two observations per panelist from our two rounds of surveying (except for non-cognitive skills, where we only have Round 2 data), of the LHS variable described in the column heading on the RHS variables described in the row labels. As in our other tables, we limit the sample to those who completed both of our rounds (i.e., who completed all four of our modules).

Results Appendix. Additional results and robustness checks

Appendix Table 4 confirms that the data warrant pooling. Each column presents an estimate of a B-count coefficient for either Round 1 ORIV (odd-numbered columns) or Round 2 ORIV (even-numbered columns), varying outcomes (objective or subjective financial index) and B-counts (Full B-count and the two Sparsity B-counts). The B-count coefficients are qualitatively similar across rounds, and do not reject equality at conventional p-value cutoffs.

Appendix Table 5 is one of several ways we address the possibility of spurious correlation. Here we focus on a reverse causality interpretation[55] of our main results, through which having lower financial resources produces behavioral biases. We consider this hypothesis by varying our main specification in two ways. One way is using Round 2 data only and instrumenting for the Round 2 B-count with the Round 1 B-count. That "Standard IV" approach uses only the 3-year-earlier measurements of behavioral biases to identify the correlation between our three main B-counts and subjective financial condition (Columns 1, 4, and 7). The second way is conditioning on objective financial condition when subjective financial index is the outcome (as we do in Table 4); that may err on the side of over-controlling, but allows us to address the possibility that (objectively) low financial resources produce behavioral tendencies by controlling for the former (Columns 2, 5, and 8).[56] Granting that possibility, we then instrument for Round 2 objective financial condition with Round 1 objective financial condition in Columns 3, 6, and 9. The B-count conditional correlation with subjective financial condition remains strongly negative in each of these nine specifications, suggesting that our main results are not driven by reverse causality.

Re: other spurious correlation hypotheses, we refer the reader back to Table 5 and Section 4-D. The former addresses the standard omitted-variable, unobserved heterogeneity concern by varying control variable specifications. The latter details how our survey design and controls for survey effort minimize the likelihood of spurious correlations between outcome measures and behavioral bias measures.

[55] We say "interpretation" instead of "concern" here, because if reverse causality were to drive the results, that would be important to discover in the sense that it would motivate a revamp of most behavioral models.

[56] We use "produce" instead of "exacerbate" here intentionally, to highlight another benefit of relying on discrete measures of behavioral biases: in our setup it would need to be the case the worse financial condition increases the likelihood that people indicate *any* deviation from classical benchmarks.

Appendix Table 6 decomposes the subjective and objective financial indexes into their components, and shows that links between B-counts and these outcomes are robustly negative: all 27 B-count coefficients are negative, 17 of them have p-values <0.01, and each implies an economically large marginal change in the outcome variable per one standard deviation change in the B-count. There is evidence of some quantitative heterogeneity, however, including within-index. E.g., the Full B-count coefficients on the subjective financial condition index components (each of which have p-values<0.01) range from -0.04 to -0.14 (Panel A, Columns 6-9).

Appendix Table 7 confirms robustness to other functional forms for the B-count: the natural logarithm of the B-count, the ratio of the panelist's B-count to their count of non-missing sources of potential behavioral biases, B-count quartiles (the results on which do not reject a linear relationship between outcomes and the B-count), and the "B-tile," a consumer-level measure of the *magnitude* of behavioral deviations from classical benchmarks.[57] The marginal effects in these alternative specifications hardly differ from those in Table 4 at all—note how similar are the d(Outcome)/d(1 SD B-count) levels across specifications.

Untabulated results, where we estimate the specifications in Table 4 separately for different sub-groups based on demographics, etc., do not reject equality of the B-count coefficient across sub-groups. Subject to the caveat that these tests are under-powered, these results support the assumption of a separable behavioral wedge in equation (1). They also fail to support a knife-edge interpretation of our results in which a narrow subset of panelists drives the results. And they cast doubt on the efficacy of targeting behavioral consumers based on more readily observable characteristics (see also Section 7).

[57] Some of our bias measures are continuous, permitting percentiles to take on the full range of values from 1 to 100. For discrete-response and uni-directional outcomes like loss aversion, the B–tiles take on fewer values but still measure the degree of deviation from classical benchmarks in useful ways. For example, loss aversion takes on four values: unbiased, and then three ordered responses (whether the individual respondent rejects the compound but not the single lottery, rejects the single but not the compound lottery, or rejects both) coded as 1/2/3. Any respondent accepting both lotteries receives a 0 (meets the classical benchmark), and 37% of individuals share that response. Anyone with the smallest deviation from the benchmark therefore is in the 37th percentile, and 13% of responses fall into that category. Summing, anyone in the next category is in the 50th(=37th+13th), and so on. The B-tile calculates each person's percentile ranking for each of the 17 potential sources of behavioral bias, relative to others in the sample, and sums them. If a person were to be the most biased person in the sample on all 17, that person would have a B-tile of (close to) 17.

Appendix Table 8 examines whether using the ALP's sampling weights changes our main empirical results (see Appendix Table 2 for a similar exercise re: B-count descriptive statistics). Here we compare weighted estimates to our main unweighted ones from Table 4 and reveals that the weighted coefficients are uniformly more negative (i.e., larger in an economic sense) in point terms, but less precise (e.g., while each of the six unweighted coefficients has a p-value<0.01, two of the weighted coefficients has p-value <0.01 and one has a p>0.10). Mechanically, it must be the case that panelists who are under-sampled by RAND (and therefore over-weighted) have noisier relationships between our outcomes and covariates.

Re: external validity, the glass half-empty interpretation of these results and our setup is that ALP sampling weights produce noisier inferences on behavioral summary statistics and, in any case, are based on demographics but not our variables of greatest interest; therefore, the extent to which our inferences our valid for the entire U.S. population is an open question. The glass half-full interpretation is that we have an unusually broad sample compared to most studies in the behavioral social sciences, and that our results on B-count properties and their conditional correlations with outcomes are not unduly sensitive to weighting that is designed to produce valid inferences for the U.S. population.

Appendix Table 9 shows that our OLS results are attenuated (compare Column 1 to Column 7; see also Appendix Table 11), but closer to the ORIV results when we use "well-measured" subsamples of panelists with arguably less measurement error: those with identical B-counts across rounds (Columns 3-6),[58] survey response times that are not in the tails (Columns 4 and 6), and/or those with identical financial literacy test scores across rounds (Columns 5 and 6).[59] When we use all three of those well-measured filters, the Full B-count OLS estimate is nearly identical to the ORIV (Panel A Column 6 vs. Column 7). We see a similar qualitative pattern for the Narrow Sparsity B-count in Panel B, with OLS estimates on the well-measured sub-samples indicating statistically strong and economically meaningful negative correlations that are closer in magnitude to the ORIV estimates than the full-sample OLS. However, the magnitude of the OLS estimates on the well-measured sub-samples remains substantially smaller than the ORIV, suggesting that

[58] Black et al. (2000) formalizes an approach for using sub-samples with relatively stable measures.

[59] Financial literacy is an example of a decision input that is relatively stable in a measurement sense across our rounds, and widely found to be strongly linked to financial outcomes.

misclassification is biasing ORIV estimates of the Narrow Sparsity B-count somewhat in the direction of spuriously large negative correlations.

Appendix Table 10 shows the full set of coefficients on the covariates in specifications (1)-(3) in Table 4. The table sheds light on the conditional correlations of other variables with outcomes (subject to caveats re: over-controlling). Income is positively correlated with objective financial condition, with the B-count marginal effects equating to a drop of multiple income deciles; e.g., to moving someone from the 3rd to the 1st income decile, or from the 9th to the 5th decile. Income is more weakly correlated with subjective financial well-being, consistent with research on happiness, and weakly negatively so once we control for objective financial condition. Other coefficients in the first and second columns reverse once we control for objective financial condition in column 3, showing its power as a control and highlighting the relative robustness of the correlations between the B-count and financial condition. For subjective financial condition, the most noteworthy pattern is that the B-count, and missingness thereon, have correlations that are more robust to the inclusion of objective financial condition as an additional covariate than any other variable or group of variables, with the possible exception of survey response times.

Appendix Table 1. Other covariates: Measuring classical decision inputs and survey effort

Variable	Definition/specification
Demographics:	
Gender	Indicator, "1" for female.
Age	Four categories: 18-34, 35-45, 46-54, 55+
Education	Four categories: HS or less, some college/associates, BA, graduate
Income	The ALP's 17 categories (collapsed into deciles in some specifications)
Race/ethnicity	Three categories: White, Black, or Other; separate indicator for Hispanic
Marital status	Three categories: married/co-habitating; separated/divorced/widowed; never married
Household size	Five categories for count of other members: 0, 1, 2, 3, 4+
Employment status	Five categories: working, self-employed, not working, disabled, missing
Immigrated to USA	Indicator, "1" for immigrant
State of residence	Fixed effects
Risk, patience:	
Risk aversion (financial)	100-point scale on financial risk-taking from Dohmen et al., with higher values indicating greater risk aversion
Risk aversion (income)	Adaptive lifetime income scale from Barsky et al., 1-6 with 6 indicating greatest risk aversion
Patience	Average savings rate across the 24 Convex Time Budget decisions, standardized
Cognitive and noncognitive skills	
Fluid intelligence	# correct on standard 15-question, non-adaptive number series quiz
Numeracy	# correct on Banks and Oldfield questions re: division and %
Financial literacy	# correct on Lusardi and Mitchell "Big Three" questions re: interest, inflation, and diversification
Executive attention	# correct on 2-minute Stroop test; respondents instructed to answer as many q's correctly as they can
Big Five Personality Traits	One variable per trait, from Rammstedt and John's validated 10-question test and scorecard (Round 2 only)
Survey effort and attrition	
Time spent on questions	Measured for each B-factor (and other variables), included as decile indicators relative to other respondents
Item non-response	Indicators for variables with non-trivial rates of non-response (although all are <5%): Income, employment status, risk, patience, cognitive skills, non-cognitive skills.

For more details on the cognitive skills measures, please see Data Appendix Section 2.

Appendix Table 2. Key B-count descriptive statistics, without and with population weighting

	(1)	(2)	(3)	(4)	(5)	(6)
weighted?	no	yes	no	yes	no	yes
	Mean (SD), across both rounds		Correlation with full B-count		Correlation (Round1,Round2)	
B-(sub)-count						
Full	9.96 (2.12)	9.97 (2.16)	1.00	1.00	0.44	0.44
Sparsity: Narrow	1.20 (0.66)	1.25 (0.67)	0.39	0.40	0.27	0.22
Sparsity: Broad	4.12 (1.33)	4.20 (1.32)	0.69	0.69	0.39	0.36
Expected biases	8.46 (2.09)	8.43 (2.12)	0.86	0.85	0.44	0.40
Non-expected biases	1.50 (1.04)	1.53 (1.04)	0.25	0.26	0.24	0.16
Math biases	2.52 (0.90)	2.58 (0.89)	0.57	0.56	0.44	0.38
Non-math biases	7.43 (1.74)	7.39 (1.78)	0.90	0.91	0.32	0.33
Preference biases	4.28 (1.25)	4.15 (1.29)	0.51	0.54	0.23	0.25
Non-preference biases	5.68 (1.74)	5.82 (1.71)	0.82	0.81	0.49	0.44
Missing inputs	0.84 (1.48)	0.97 (1.65)	-0.33	-0.38	0.36	0.45
N	1690	1690	1690	1690	1690	1690
N panelists	845	845	845	845	845	845

Our data consist of two survey rounds, of two modules each, conducted 3 years apart. We include only those panelists who took all four modules across both rounds (N=845). B-count and B-sub-count definitions are summarized in Table 1 and discussed in Sections 3-A and -B. Round-to-round correlations for B-counts adjust for missing data by conditioning on the count of missing bias measures in each survey round. Column 3 here reproduces Table 2 Column 6. Column 5 here reproduces Table 2 Column 7.

Appendix Table 3. Measuring financial condition and subjective well-being: Definitions, sampling, and descriptive statistics for index components

	Data used				Mean	SD	Pairwise correlation										
	# of questions per module	From our modules?	From other modules?	# panelists with nonmissing	(All rescaled to [0,1])		Net worth>0	Retirement assets>0	Owns stocks	Spent < income	No severe hardship	Financial satisfaction	Retirement saving adequacy	Non-ret saving adequacy	Lack financial stress	Happiness Last 30 days	Happiness in general
Panel A. Objective financial condition index components																	
Net worth>0	2	yes	no	821	0.50	0.50	1.00										
Retirement assets>0	2	yes	no	831	0.60	0.49	0.54	1.00									
Owns stocks	3	yes	no	835	0.54	0.50	0.56	0.96	1.00								
Spent < income in last 12 months	1	yes	no	841	0.41	0.49	0.44	0.35	0.35	1.00							
No severe hardship in last 12 months	4	yes	no	842	0.61	0.49	0.49	0.45	0.49	0.50	1.00						
Panel B. Subjective financial condition index components																	
Financial satisfaction scale	1	yes	no	842	0.59	0.26	0.35	0.32	0.33	0.43	0.49	1.00					
Retirement saving adequacy scale	1	yes	no	842	0.47	0.41	0.44	0.46	0.45	0.46	0.56	0.53	1.00				
Non-retirement saving adequacy scale	1	yes	no	843	0.49	0.37	0.34	0.17	0.18	0.35	0.37	0.31	0.49	1.00			
Lack of financial stress scale	1	yes	no	845	0.47	0.30	0.40	0.31	0.32	0.43	0.53	0.53	0.47	0.35	1.00		
Panel C. Other measures of subjective well-being: Happiness index components																	
Happiness last 30 days	1	no	yes	509	0.62	0.21	0.22	0.22	0.25	0.21	0.33	0.41	0.25	0.12	0.32	1.00	
Happiness in general	1	no	yes	675	0.75	0.26	0.27	0.27	0.25	0.26	0.29	0.39	0.24	0.09	0.34	0.65	1.00

Unit of observation is the individual respondent, with multiple observations per respondent averaged across survey rounds (for variables in our modules) or across other ALP modules (for variables we merge in from other ALP modules). Other ALP modules used here are all administered *between* our survey rounds; we could not find relevant data collected in modules adminstered after or during our second round. As in most of our main tables, we limit the sample frame here to panelists who completed both of our survey rounds (N=845). Correlations estimated using the two-step "polychoric" procedure in Stata.

Variable definitions: Each variable is scaled so that higher values indicate better financial condition and/or subjective well-being. Each measure here is scaled or rescaled to [0, 1] for comparability.

Net worth is from two summary questions drawn from the National Longitudinal Surveys: "Please think about all of your household assets (including but not limited to investments, other accounts, any house/property you own, cars, etc.) and all of your household debts (including but not limited to mortgages, car loans, student loans, what you currently owe on credit cards, etc.) Are your household assets worth more than your household debts?" and "You stated that your household's [debts/assets] are worth more than your household's [assets/debts]. By how much?"

Retirement assets is from questions asking specifically whether someone has one or more IRA accounts and one or more workplace plans, followed in each case by questions on amounts in such accounts. Questions like these are asked in the Survey of Consumer Finances, the Health and Retirement Study, and many other surveys.

Stockholding is from questions on stock mutual funds in IRAs, stock mutual funds in 401ks/other retirement accounts, and direct holdings. Questions like these are asked in the Survey of Consumer Finances, the Health and Retirement Study, and many other surveys.

Spent < income question is from the Survey of Consumer Finances: "Over the past 12 months, how did your household's spending compare to your household's income? If the total amount of debt you owe decreased, then count yourself as spending less than income. If the total amount of debt you owe increased, then count yourself as spending more than income." Response options are: "Spent more than income", "Spent same as income", and "Spent less than income".

(No) severe hardship questions are taken from the National Survey of American Families: late/missed payment for rent, mortgage, heat, or electric; moved in with other people because could not afford housing/utilities; postponed medical care due to financial difficulty; adults in household cut back on food due to lack of money. Response options for each of the four are Yes or No.

The **Financial satisfaction** question follows standard life and economic satisfaction question wording: "How satisfied are you with your household's overall economic situation?"; responses on a 100-point scale (input using slider or text box).

Retirement and non-retirement savings adequacy questions are placed one each in the two different modules, with different wording, to mitigate mechanical correlations. The questions are: "Using any number from one to five, where one equals not nearly enough, and five equals much more than enough, do you feel that your household is saving and investing enough for retirement? Please consider the income you and any other members of your household expect to receive from Social Security, 401(k) accounts, other job retirement accounts and job pensions, and any additional assets you or other members of your household have or expect to have" and "Now, apart from retirement savings, please think about how your household typically uses the money you have: how much is spent and how much is saved or invested. Now choose which statement best describes your household". These questions are variants on standard ones, but in each case our 5 response options are framed to encourage people to recognize tradeoffs between saving and consumption: any response that includes "saving more" also includes "and borrowing/spending less", and vice versa. In mapping the 5 responses into the variables used here, we code: saved-enough, more-than-enough, and much-more-than-enough as 1 (the latter two responses are rare: 3% of the sample for retirement, and 4% for non-retirement); saved < enough as 0.5; saved << enough as 0.

Financial stress question is taken from The Survey of Forces**:** "To what extent, if any, are finances a source of stress in your life?"; responses on a 100-point scale (respondents can input using slider or text box).

Life satisfaction question is measured using some one of three minor variants on the standard "… how satisfied are you with your life as a whole these days?" asked in many surveys worldwide. For the other-module measure, we take the within-panelist average of non-missing responses to this question across the six ALP modules in which it has appeared subsequent to our round 1 modules, as of this writing. Of the 809/845 panelists with at least one non-missing response, 640 have at least two.

Happiness last 30 days is measured using the standard "During the past 30 days, how much of the time have you been a happy person?" asked in many surveys worldwide. We take the within-panelist average of non-missing responses to this question across the four ALP modules in which it has appeared subsequent to our round 1 modules, as of this writing. Of the 509/845 panelists with at least one non-missing response, 474 have at least two, .

Happiness in general is measured using the standard "Taking all things together, I am generally happy" question asked in many surveys worldwide, incuding ALP module 425.

Appendix Table 4. ORIV estimates are similar across survey rounds
(Compare to Table 4)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Financial outcome index includes:	*Objective*	*Objective*	*Subjective*	*Subjective*	*Objective*	*Objective*	*Subjective*	*Subjective*	*Objective*	*Objective*	*Subjective*	*Subjective*
B-count: Full	-0.073***	-0.053**	-0.094***	-0.080***								
	(0.023)	(0.021)	(0.022)	(0.021)								
B-count: Sparsity Broad					-0.085**	-0.095***	-0.130***	-0.130***				
					(0.034)	(0.033)	(0.034)	(0.031)				
B-count: Sparsity Narrow									-0.193***	-0.300***	-0.292***	-0.397***
									(0.067)	(0.083)	(0.069)	(0.082)
d(LHS)/d(1 SD B-count)	-0.155	-0.112	-0.201	-0.169	-0.113	-0.126	-0.173	-0.172	-0.128	-0.199	-0.194	-0.264
mean(LHS)	0.522	0.539	0.493	0.515	0.522	0.539	0.493	0.515	0.522	0.539	0.493	0.515
Round included?	1 only	2 only	1 only	2 only	1 only	2 only	1 only	2 only	1 only	2 only	1 only	2 only
N panelists	841	844	841	844	841	844	841	844	841	844	841	844
N	1682	1688	1682	1688	1682	1688	1682	1688	1682	1688	1682	1688

* 0.10 ** 0.05 *** 0.01. Standard errors, clustered on panelist, in parentheses. Each column presents results from a single-round Obviously Related Instrumental Variables regression (per Section 4-C) of the LHS variable described in the column label on the variables described in the row labels + the complete set of covariates described in Appendix Table 1. Table 1 provides details on our B-count variable definitions; higher values indicate more behavioral biases. Table 3 provides details on our LHS variable definitions; higher values indicate better financial condition.

Appendix Table 5. Identifying relationships between outcomes and B-counts: Reverse causality looks unlikely

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Financial outcome index includes:	*Subjective measures*								
B-count: Full	-0.055***	-0.046***	-0.042**						
	(0.018)	(0.017)	(0.017)						
B-count: Sparsity Broad				-0.123***	-0.093***	-0.076***			
				(0.030)	(0.027)	(0.026)			
B-count: Sparsity Narrow							-0.315***	-0.231***	-0.177***
							(0.064)	(0.059)	(0.054)
Objective financial index		0.332***	0.488***		0.333***	0.513***		0.300***	0.490***
		(0.036)	(0.054)		(0.037)	(0.055)		(0.041)	(0.057)
d(LHS)/d(1 SD B-count)	-0.122	-0.102	-0.092	-0.164	-0.124	-0.102	-0.213	-0.156	-0.120
Data used	Round 2	Round 2	Round 2	Round 2	Round 2	Round 2	Round 2	Round 2	Round 2
IV for B-count with Round 1?	yes	yes	yes	yes	yes	yes	yes	yes	yes
IV for objective financial index with Round 1?	no	no	yes	no	no	yes	no	no	yes
mean(LHS)	0.515	0.515	0.515	0.515	0.515	0.515	0.515	0.515	0.515
N = N panelists	844	844	844	844	844	844	844	844	844

* 0.10 ** 0.05 *** 0.01. Standard errors in parentheses. Each column presents results from a single two-stage least square regression of the LHS variable described in the column label on the variables described in the row labels + the complete set of covariates described in Appendix Table 1. Table 1 provides details on our B-count variable definitions; higher values indicate more behavioral biases. Table 3 provides details on the subjective financial condition index construction; higher values indicate better financial condition and higher experienced utility. The difference between this table and our main specifications is that here we only use "replicate 2" and "standard IV": we use Round 2 data for all variables except for instruments.

Appendix Table 6. B-counts are strongly conditionally correlated with financial index components
(Compare to Table 4)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Index:	*Objective financial condition*					*Subjective financial condition*			
Component:	*Net worth>0*	*Retirement assets>0*	*Owns stocks*	*Spent < income*	*No severe hardship*	*Financial satisfaction*	*Retirement saving adequacy*	*Non-ret saving adequacy*	*Lack financial stress*
Panel A.									
B-count: Full	-0.077***	-0.067**	-0.047*	-0.041	-0.080***	-0.042***	-0.067***	-0.137***	-0.088***
	(0.029)	(0.026)	(0.027)	(0.029)	(0.029)	(0.016)	(0.019)	(0.030)	(0.024)
d(LHS)/d(1 SD B-count)	-0.164	-0.142	-0.099	-0.087	-0.170	-0.089	-0.142	-0.291	-0.182
mean(LHS)	0.500	0.598	0.543	0.407	0.610	0.586	0.473	0.468	0.490
N	3300	3322	3338	3360	3362	3362	3368	3362	3310
Panel B.									
B-count: Sparsity Broad	-0.103**	-0.089**	-0.050	-0.091**	-0.121***	-0.061**	-0.105***	-0.205***	-0.140***
	(0.043)	(0.041)	(0.039)	(0.043)	(0.044)	(0.024)	(0.030)	(0.046)	(0.038)
d(LHS)/d(1 SD B-count)	-0.136	-0.118	-0.066	-0.120	-0.160	-0.081	-0.139	-0.272	-0.183
mean(LHS)	0.500	0.598	0.543	0.407	0.610	0.586	0.473	0.468	0.490
N	3300	3322	3338	3360	3362	3362	3368	3362	3310
Panel C.									
B-count: Sparsity Narrow	-0.276***	-0.197**	-0.159*	-0.317***	-0.231***	-0.142***	-0.298***	-0.479***	-0.391***
	(0.095)	(0.089)	(0.087)	(0.095)	(0.089)	(0.048)	(0.067)	(0.099)	(0.086)
d(LHS)/d(1 SD B-count)	-0.183	-0.131	-0.106	-0.211	-0.153	-0.094	-0.197	-0.318	-0.259
mean(LHS)	0.500	0.598	0.543	0.407	0.610	0.586	0.473	0.468	0.490
N	3300	3322	3338	3360	3362	3362	3368	3362	3310

Each panel*column reports results from a single regression, using the same specification as Table 4 Col 1 and 2 (in Panel A here), Table 4 Col 4 and 5 (in Panel B here), or Table 4 Col 7 and 8 (in Panel C here). Sample sizes are slightly smaller here than in Table 4 because of non-response in index components. See Appendix Table 3 for index component variable definitions and statistics.

Appendix Table 7. Functional form robustness of the Full B-count's conditional correlation with financial outcomes

(Columns 1 and 6 here are same specifications as Columns 1 and 2 in Table 4)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Financial outcome index includes:	*Objective measures*					*Subjective measures*				
Full B-count	-0.061***					-0.084***				
	(0.018)					(0.018)				
ln(B-count)		-0.647***					-0.873***			
		(0.200)					(0.209)			
B-count proportion			-0.974***					-1.203***		
			(0.277)					(0.271)		
B-tile: Average percentile across all biases				-1.293***					-1.571***	
				(0.357)					(0.329)	
B-count: 2nd quartile					-0.156					-0.124
					(0.096)					(0.103)
B-count: 3rd quartile					-0.303**					-0.399***
					(0.146)					(0.144)
B-count: 4th quartile					-0.446***					-0.580***
					(0.138)					(0.141)
d(LHS)/d(1 SD B-count variable)	-0.130	-0.154	-0.130	-0.127		-0.179	-0.208	-0.160	-0.154	
dy/d(1 SD B-count quartile 2)					-0.068					-0.054
dy/d(1 SD B-count quartile 3)					-0.141					-0.186
dy/d(1 SD B-count quartile 4)					-0.208					-0.271
mean(LHS)	0.531	0.531	0.531	0.531	0.531	0.504	0.504	0.504	0.504	0.504
N	3370	3370	3370	3370	3370	3370	3370	3370	3370	3370

* 0.10 ** 0.05 *** 0.01. Standard errors, clustered on panelist, in parentheses. Each column presents results from a single pooled Obviously Related Instrumental Variables regression (equation 3 in the text) of the LHS variable described in the column label on the variable(s) described in the row labels + the complete set of covariates described in Appendix Table 1. Table 1 provides details on our Full B-count variable definition; higher values indicate more behavioral biases. Table 3 provides details on our LHS variable definitions; higher values indicate better financial condition.

Appendix Table 8. B-count conditional correlations with financial outcomes: Unweighted vs. unweighted

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Financial outcome index includes:	*Objective*		*Subjective*		*Objective*		*Subjective*		*Objective*		*Subjective*	
B-count: Full	-0.061***	-0.065**	-0.084***	-0.114***								
	(0.018)	(0.032)	(0.018)	(0.035)								
B-count: Sparsity Broad					-0.090***	-0.152*	-0.128***	-0.238***				
					(0.028)	(0.078)	(0.028)	(0.090)				
B-count: Sparsity Narrow									-0.236***	-0.500	-0.328***	-0.753*
									(0.063)	(0.334)	(0.065)	(0.451)
d(LHS)/d(1 SD B-count)	-0.130	-0.138	-0.179	-0.242	-0.119	-0.202	-0.169	-0.315	-0.157	-0.332	-0.218	-0.500
Sampling weights?	no	yes	no	yes	no	yes	no	yes	no	yes	no	yes
Same/analogous specification in Table 4	col 1	col 1	col 2	col 2	col 4	col 4	col 5	col 5	col 7	col 7	col 8	col 8
mean(LHS)	0.531	0.484	0.504	0.489	0.531	0.484	0.504	0.489	0.531	0.484	0.504	0.489
N	3370	3370	3370	3370	3370	3370	3370	3370	3370	3370	3370	3370

* 0.10 ** 0.05 *** 0.01. Odd-numbered columns are reproduced from Table 4; even-numbered columns use the same specification as the preceding column but with sampling weights.

Appendix Table 9. OLS results are attenuated, but (much) less so when limiting the sample to those with stable responses (Column 7 is reproduced from Table 4 Columns 2 and 8)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Financial outcome index includes:	*Subjective*	*Subjective*	*Subjective*	*Subjective*	*Subjective*	*Subjective*	*Subjective*
Panel A							
B-count: Full	-0.019***	-0.015***	-0.037***	-0.048***	-0.062***	-0.085***	-0.084***
	(0.003)	(0.003)	(0.011)	(0.013)	(0.016)	(0.014)	(0.016)
N panelists	844	690	154	142	114	107	844
Panel B							
B-count: Sparsity Narrow	-0.067***	-0.036***	-0.110***	-0.098***	-0.136***	-0.141***	-0.328***
	(0.009)	(0.008)	(0.017)	(0.017)	(0.019)	(0.021)	(0.065)
N panelists	844	433	411	378	289	267	844
Estimator	OLS	OLS	OLS	OLS	OLS	OLS	ORIV
B-counts equal across rounds?	All	No	Yes	Yes	Yes	Yes	All
Time spent decile in [2,9]?	No	No	No	Yes	No	Yes	No
Stable "Big 3" responses across rounds?	No	No	No	No	Yes	Yes	No

* 0.10 ** 0.05 *** 0.01. Standard errors, clustered on panelist, in parentheses. Each column presents results from a single pooled OLS regression of the subjective financial well-being index on the variables described in the row labels + the complete set of covariates described in Appendix Table 1. Table 1 provides details on our B-count variable definitions; higher values indicate more behavioral biases. Table 3 provides details on our LHS variable definitions; higher values indicate better financial condition and the index is scaled on [0,1].

Appendix Table 10. Main specifications for estimating correlation between financial condition and the Full B-count, showing results on all of the other covariates. (Same specifications as Table 4, Columns 1-3.)

	(1)	(2)	(3)
Financial outcome index includes:	*Objective measures*	*Subjective measures*	*Subjective measures*
B-count: Full	-0.061***	-0.084***	-0.064***
	(0.018)	(0.018)	(0.016)
Missing bias count	-0.044***	-0.055***	-0.040***
	(0.013)	(0.013)	(0.012)
Female	0.019	0.019	0.012
	(0.020)	(0.018)	(0.015)
Education: Some college	-0.033	-0.032	-0.020
	(0.025)	(0.023)	(0.020)
Education: B.A.	0.031	-0.016	-0.026
	(0.027)	(0.026)	(0.023)
Education: Grad school	0.050*	0.047	0.030
	(0.030)	(0.030)	(0.026)
Income: 2nd decile	0.061**	-0.011	-0.032
	(0.026)	(0.026)	(0.023)
Income: 3rd decile	0.114***	0.004	-0.034
	(0.033)	(0.032)	(0.029)
Income: 4th decile	0.182***	-0.007	-0.068**
	(0.030)	(0.030)	(0.028)
Income: 5th decile	0.241***	0.017	-0.065**
	(0.036)	(0.035)	(0.031)
Income: 6th decile	0.297***	0.047	-0.054*
	(0.034)	(0.033)	(0.030)
Income: 7th decile	0.304***	0.044	-0.059*
	(0.035)	(0.036)	(0.032)
Income: 8th decile	0.358***	0.088**	-0.034
	(0.036)	(0.036)	(0.032)
Income: 9th decile	0.360***	0.073*	-0.050
	(0.039)	(0.041)	(0.035)
Income: Top decile	0.504***	0.191***	0.020
	(0.039)	(0.048)	(0.045)
Age 35-45	0.026	-0.031	-0.040**
	(0.023)	(0.022)	(0.019)
Age 46-54	0.086***	-0.014	-0.043**
	(0.024)	(0.023)	(0.020)
Age 55+ (Max 60)	0.121***	0.028	-0.013
	(0.025)	(0.024)	(0.021)
Race: Black	-0.037	0.025	0.037
	(0.030)	(0.027)	(0.023)
Race: Other non-white	-0.064**	-0.023	-0.001
	(0.027)	(0.029)	(0.027)
Latino	-0.041	0.013	0.027
	(0.027)	(0.025)	(0.023)
Immigrant	0.050*	0.026	0.009
	(0.027)	(0.027)	(0.025)
Peviously married	0.010	0.019	0.016
	(0.021)	(0.019)	(0.017)
Never married	0.026	-0.015	-0.024
	(0.022)	(0.021)	(0.018)
Other household members: 1	-0.012	-0.033*	-0.029*
	(0.020)	(0.018)	(0.016)
Other household members: 2	-0.015	-0.020	-0.015
	(0.021)	(0.021)	(0.019)
Other household members: 3	-0.035	-0.040	-0.028
	(0.026)	(0.026)	(0.022)
Other household members: 4	-0.037	-0.029	-0.016
	(0.033)	(0.030)	(0.026)
Work status: Self-employed	-0.022	-0.042	-0.035
	(0.032)	(0.030)	(0.026)
Work status: Not working	-0.030	0.007	0.017
	(0.028)	(0.025)	(0.023)
Work status: Disabled	-0.155***	-0.087***	-0.034
	(0.032)	(0.031)	(0.027)
Work status: Unknown	-0.135**	0.015	0.061
	(0.063)	(0.078)	(0.068)
Patience in CTB task on 0 to 1 scale	0.019	0.022	0.015
	(0.029)	(0.027)	(0.023)
Patience missing	0.029	0.028	0.018
	(0.035)	(0.034)	(0.030)
Risk aversion: Financial on -1 to 0 scale	-0.058*	-0.035	-0.016

	(0.033)	(0.030)	(0.026)
Risk aversion: financial missing	-0.023	-0.001	0.007
	(0.080)	(0.096)	(0.099)
Risk aversion: lifetime income	0.012**	0.009*	0.005
	(0.006)	(0.005)	(0.005)
Risk aversion: income missing	0.036	0.097	0.084
	(0.107)	(0.081)	(0.077)
Fluid intelligence score	-0.006	-0.009*	-0.007*
	(0.005)	(0.005)	(0.004)
Fluid intelligence missing	0.022	-0.046	-0.053
	(0.091)	(0.088)	(0.087)
Numeracy score	0.004	-0.011	-0.013
	(0.015)	(0.012)	(0.011)
Numeracy missing	-0.027	-0.084*	-0.075*
	(0.055)	(0.046)	(0.045)
Financial literacy score	0.026**	-0.013	-0.022**
	(0.011)	(0.011)	(0.009)
Financial literacy missing	0.019	-0.070	-0.076
	(0.091)	(0.121)	(0.131)
Stroop score/100	0.013	-0.011	-0.015
	(0.032)	(0.028)	(0.025)
Stroop missing	0.001	-0.039	-0.039
	(0.037)	(0.039)	(0.035)
Survey effort: 2nd decile	0.033	-0.039	-0.051**
	(0.030)	(0.026)	(0.023)
Survey effort: 3rd decile	0.028	-0.032	-0.041*
	(0.029)	(0.028)	(0.025)
Survey effort: 4th decile	0.008	-0.068**	-0.071***
	(0.031)	(0.027)	(0.024)
Survey effort: 5th decile	0.016	-0.061**	-0.067***
	(0.032)	(0.029)	(0.025)
Survey effort: 6th decile	0.022	-0.070**	-0.078***
	(0.031)	(0.029)	(0.026)
Survey effort: 7th decile	0.038	-0.033	-0.046*
	(0.031)	(0.029)	(0.026)
Survey effort: 8th decile	0.037	-0.057**	-0.070***
	(0.031)	(0.029)	(0.025)
Survey effort: 9th decile	0.007	-0.076***	-0.078***
	(0.031)	(0.029)	(0.026)
Survey effort: 10th decile	0.013	-0.026	-0.031
	(0.030)	(0.028)	(0.026)
Extraversion score	0.002	0.006	0.006
	(0.004)	(0.004)	(0.003)
Agreeableness score	0.002	0.008*	0.008*
	(0.005)	(0.004)	(0.004)
Conscientiousness score	0.015***	0.010**	0.005
	(0.005)	(0.005)	(0.004)
Neuroticism score	-0.005	-0.009**	-0.007**
	(0.004)	(0.004)	(0.004)
Openness score	-0.014***	-0.008*	-0.003
	(0.005)	(0.004)	(0.004)
Personality variables missing	-0.026	-0.013	-0.004
	(0.046)	(0.045)	(0.037)
Objective financial index			0.340***
			(0.026)
State of residence fixed effects		Individual states not shown	
pval demographics=0	0.000	0.000	0.004
pval cognitive skills=0	0.304	0.121	0.023
pval noncognitive skills=0	0.001	0.000	0.005
pval classical preferences=0	0.109	0.223	0.542
pval survey effort=0	0.898	0.104	0.028
pval state FE=0	0.000	0.000	0.000
mean(LHS)	0.531	0.504	0.504
N	3370	3370	3370

* 0.10 ** 0.05 *** 0.01. Standard errors, clustered on panelist, in parentheses. Each column presents results from a single pooled Obviously Related Instrumental Variables regression (equation 4 in the text) of the LHS variable described in the column label on the variables described in the row labels. Appendix Table 1 provides details on the other covariate definitions. Table 1 provides details on our B-count variable definitions; higher values indicate more behavioral biases. Table 3 provides details on our LHS variable definitions; higher values indicate better financial condition.

Appendix Table 11. OLS coefficients are attenuated and sensitive to dropping other covariates

(Compare to Table 5)

LHS=Subjective financial index	(1)	(2)	(3)	(4)	(5)	(6)
B-count: Full	-0.019***	-0.035***				
	(0.003)	(0.003)				
B-count: Sparsity Broad			-0.026***	-0.051***		
			(0.005)	(0.005)		
B-count: Sparsity Narrow					-0.067***	-0.083***
					(0.009)	(0.010)
Covariates in Appendix Table 2 included?	All	B-miss only	All	B-miss only	All	B-miss only
Comparable ORIV spec in Table 5	Pan A Col 1	Pan A Col 6	Pan B Col 1	Pan B Col 6	Pan C Col 1	Pan C Col 6
d(LHS)/d(1 SD B-count)	-0.040	-0.075	-0.034	-0.068	-0.044	-0.055
mean(LHS)	0.504	0.505	0.504	0.505	0.504	0.505
N	1685	1690	1685	1690	1685	1690

* 0.10 ** 0.05 *** 0.01. OLS, with standard errors clustered on panelist. Each column presents results from a single OLS regression, using both rounds of data (two obs per panelist), of the subjective financial index on the variables described in the row labels. "B-miss" refers to the count of missing behavioral biases.

Appendix Table 12. Identifying relationships between outcomes and B-counts: Sensitivity to covariate specifications
(Same as Table 5, but with objective financial index as dependent variable instead of subjective financial index)

	LHS=Objective financial index								
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Panel A. Full B-Count									
Full B-count	-0.061***	-0.082***	-0.063***	-0.059***	-0.059***	-0.098***	-0.058***	-0.060***	-0.053***
	(0.018)	(0.016)	(0.018)	(0.015)	(0.018)	(0.012)	(0.018)	(0.019)	(0.017)
dY/d(1 SD B-count)	-0.130	-0.174	-0.134	-0.124	-0.125	-0.208	-0.122	-0.127	-0.112
Panel B. Sparsity Broad B-count									
Sparsity biases: attention+	-0.090***	-0.108***	-0.094***	-0.093***	-0.091***	-0.165***	-0.087***	-0.089***	-0.086***
	(0.028)	(0.026)	(0.029)	(0.025)	(0.028)	(0.022)	(0.028)	(0.029)	(0.028)
dY/d(1 SD B-count)	-0.119	-0.144	-0.124	-0.123	-0.120	-0.219	-0.115	-0.117	-0.114
Panel C. Sparsity Narrow B-count									
Sparsity biases: attention only	-0.236***	-0.169***	-0.240***	-0.237***	-0.241***	-0.233***	-0.235***	-0.236***	-0.232***
	(0.063)	(0.052)	(0.063)	(0.063)	(0.061)	(0.055)	(0.065)	(0.063)	(0.064)
dY/d(1 SD B-count)	-0.157	-0.112	-0.159	-0.157	-0.160	-0.155	-0.156	-0.157	-0.154
Missing bias count included?	yes	yes	yes	yes	yes	yes	yes	yes	yes
Demographics included?	yes	no	yes	yes	yes	no	yes	yes	yes
Classical preferences included?	yes	yes	no	yes	yes	no	yes	yes	yes
Cognitive skills included?	yes	yes	yes	no	yes	no	yes	yes	yes
Non-cognitive skills included?	yes	yes	yes	yes	no	no	yes	yes	yes
IV for B-count?	yes	yes	yes	yes	yes	yes	yes	yes	yes
IV for classical preferences	no	no	no	no	no	no	yes	no	yes
IV for cognitive skills	no	no	no	no	no	no	no	yes	yes
mean(LHS)	0.531	0.532	0.531	0.531	0.531	0.532	0.531	0.531	0.531
N	3370	3380	3370	3370	3370	3380	3370	3370	3370

* 0.10 ** 0.05 *** 0.01. Standard errors, clustered on panelist, in parentheses. Each panel-column presents results from a single ORIV regression of our objective financial index on the B-count described in the Panel title and row label and the other covariates described in rows at the bottom of the table. I.e., this table presents results for specifications identical to those in Table 5 except for the LHS variable.

Appendix Table 13. Identifying relationships between outcomes and B-counts: Unpacking the Math B-count

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Financial condition index includes:	*Objective*	*Objective*	*Subjective*	*Subjective*	*Objective*	*Objective*	*Subjective*	*Subjective*
Math biases (1)	-0.011	-0.018	-0.014	0.003				
	(0.044)	(0.040)	(0.041)	(0.038)				
Non-math biases (2)	-0.083***	-0.078***	-0.115***	-0.110***	-0.083***	-0.078***	-0.114***	-0.111***
	(0.029)	(0.027)	(0.030)	(0.028)	(0.029)	(0.027)	(0.032)	(0.031)
Exepcted Direction Math Biases (3)					-0.016	-0.025	-0.040	-0.029
					(0.044)	(0.042)	(0.047)	(0.045)
Non-expected Direction Math Biases (4)					0.012	0.010	0.108	0.128
					(0.121)	(0.107)	(0.138)	(0.128)
Fluid intelligence score	-0.005		-0.008*		-0.005		-0.006	
	(0.005)		(0.005)		(0.006)		(0.006)	
Fluid intelligence missing	0.037		-0.024		0.034		-0.038	
	(0.094)		(0.087)		(0.093)		(0.090)	
Numeracy score	0.004		-0.010		0.006		-0.003	
	(0.015)		(0.013)		(0.017)		(0.016)	
Numeracy missing	-0.022		-0.076		-0.015		-0.042	
	(0.057)		(0.047)		(0.064)		(0.057)	
Financial literacy score	0.031***		-0.006		0.031***		-0.007	
	(0.012)		(0.011)		(0.012)		(0.012)	
Financial literacy missing	0.033		-0.050		0.033		-0.047	
	(0.094)		(0.119)		(0.095)		(0.119)	
Stroop score	0.000		0.000		0.000		0.000	
	(0.000)		(0.000)		(0.000)		(0.000)	
Stroop missing	0.006		-0.031		0.006		-0.031	
	(0.038)		(0.041)		(0.038)		(0.043)	
pval (1)=(2)	0.240	0.304	0.091	0.051				
pval (2)=(3)					0.274	0.373	0.245	0.196
pval (2)=(4)					0.462	0.454	0.143	0.099
reproduced from Table 7?	col 1		col 2					
mean(LHS)	0.531	0.531	0.504	0.504	0.531	0.531	0.504	0.504
N panelists	843	843	843	843	843	843	843	843
N	3370	3370	3370	3370	3370	3370	3370	3370

* 0.10 ** 0.05 *** 0.01. One ORIV regression per column of the LHS variable described in the column label on the RHS variables described in the row label plus all of the additional covariates described in Appendix Table 1, *except* that even-numbered columns here do not include the cognitive skills covriates. Standard errors clustered on panelist.

Data Appendix

1. Measuring Behavioral Biases

This section details, for each of the 17 potential sources of behavioral bias we measure:

i) The motive for eliciting that potential source of bias (B-factor) and the mechanism through which that factor might affect financial condition;

ii) our elicitation method and its key antecedents;

iii) data quality indicators, including item non-response;

iv) sample size (as it compares to that for other B-factors);

v) definitions and prevalence estimates of behavioral *indicators*, with background on the distinctions between expected direction (standard) vs. less-expected (non-standard) direction biases where applicable;

vi) descriptions of the *magnitude* and *heterogeneity* of behavioral deviations, including descriptions of the distribution and—where the data permit—estimates of key parameters used in behavioral models;

Since our empirical work here is purely descriptive, we focus on our Round 1 data (ALP modules 315 and 352) to get the largest possible sample of panelists. We provide comparisons to prior work wherever possible.

A. Present- or future-biased discounting (money)

Time-inconsistent discounting has been linked, both theoretically and empirically, to low levels of saving and high levels of borrowing (e.g., Laibson 1997; Meier and Sprenger 2010; Toubia et al. 2013).

We measure discounting biases with respect to money using the Convex Time Budgets (CTB) method created by Andreoni and Sprenger (2012). In our version, fielded in ALP module 315 (the first of our two surveys), subjects make 24 decisions, allocating 100 hypothetical tokens each between (weakly) smaller-sooner and larger-later amounts. See Data Appendix Figure 1 for an example. The 24 decisions are spread across 4 different screens with 6 decisions each. Each screen varies start date (today or 5 weeks from today) x delay length (5 weeks or 9 weeks); each decision within a screen offers a different yield on saving. Among the 1,515 individuals who take our first module in Round 1, 1,502 subjects make at least one CTB choice, and the 1,422 who complete at least the first and last decisions on each of the 4 screens comprise our CTB sample.

The CTB already has been implemented successfully in field contexts in the U.S. (Barcellos and Carvalho 2014; Carvalho, Meier, and Wang 2016) and elsewhere (Giné et al. 2018). In exploring data quality and prevalence below we focus on comparisons to Andreoni and Sprenger (2012), and Barcellos and Carvalho (2014).[1] AS draw their sample from university students. BC's sample is drawn from the ALP, like ours (module 212 in their case), but they use a different adaptation of the CTB.

Indicators of response quality are encouraging for the most part. Interior allocations are more common in our sample than in AS, and comparable to BC. More of our subjects exhibit some variance in their allocations than AS or BC. Our subjects are internally consistent overall—e.g., exhibiting strong correlations in choices across different screens and delay dates—but 41% do exhibit some upward-sloping demand among 20 pairs of decisions, a figure that is within the range commonly found in discount rate elicitations but high compared to the 8% in AS.[2]

We calculate biased discounting, for each individual, by subtracting the consumption rate when the sooner payment date is five weeks from today from the consumption rate when the sooner payment date is today, for each of the two delay lengths. We then average the two differences to get a continuous measure of biased discounting. In keeping with AS, BC and several other recent papers (including Carvalho, Meier, and Wang (2016) and Goda et al. (2017)), we find little if any present-bias on average, with a median discount bias of zero, and a 1pp mean tilt toward future bias.[3]

Indicators of behavioral deviations here are bi-directional: we label someone as present-biased (future-biased) if the average difference is >0 (<0). We deem present-bias the "standard"

[1] Carvalho, Meier, and Wang use the American Life Panel like we and Barcello and Carvalho, but on a lower-income sample (ALP module 126).

[2] High rates of non-monotonic demand are not uncommon in discount rate elicitation: Andreoni and Sprenger (2012) report rates ranging from 10 to 50 percent in their literature review. In Barcellos and Carvalho 26% of subjects exhibit some upward-sloping demand, among only 4 pairs of decisions. In our sample non-monotonic demand is strongly correlated within-subject across the four screens, and decreases slightly by the final screen, suggesting that responses are picking up something systematic.

[3] Bradford et al. (2017) do find present-bias on average in their Qualtrics sample, classifying >50% as present-biased and 26% as future-biased.

direction, since future-bias is relatively poorly understood.[4] Counting any deviation from time-consistent discounting as biased, 26% of our sample is present-biased and 36% is future-biased. These prevalence estimates fall substantially if we set a higher threshold for classifying someone as behavioral; e.g., if we count only deviations > |20|pp, then only 3% of the sample is present-biased and 5% future-biased. Compared to prior prevalence estimates, our zero-threshold ones are in the middle of the range (Data Appendix Table 1). E.g., BC's CTB elicitation in the ALP shows 29% with any present-bias, and 37% with any future-bias. Goda et al. use a different elicitation method—a "time-staircase" multiple price list (Falk et al. 2016)—and classify 55% of their nationally representative sample (from the ALP and another online panel) as present-biased. In the AS sample 14% exhibit any present-bias and 12% any future-bias.

Previous studies estimate relationships between directly elicited discounting biases and outcomes in broad samples (Bradford et al. 2017; Eisenhauer and Ventura 2006; Goda et al. 2017).[5] We use CTBs rather than Multiple Price Lists, test more flexible functional forms, and control for a much richer set of (behavioral) factors that could be correlated with both discounting and outcomes.[6]

B. Present- or future-biased discounting (food)

In light of evidence that discounting can differ within-subject across domains (e.g., Augenblick, Niederle, and Sprenger 2015), we also obtain a coarse measure of discounting biases for consumption per se, by asking two questions that follow Read and van Leeuwen (1998) : "*Now imagine that you are given the choice of receiving one of two snacks for free, [right now/five weeks from now]. One snack is more delicious but less healthy, while the other is healthier but less delicious. Which would you rather have [right now/five weeks from now]: a delicious snack that is not good for your health, or a snack that is less delicious but good for your health?* We fielded these questions in our second Round1 module.

Of the 1427 persons taking our second survey, 1423 answer one of the two snack questions, and 1404 respond to both. 61% choose the healthy snack for today, while 68% choose it for five

[4] Although see Koszegi and Szeidl (2013) for a theory of future-biased discounting.

[5] Other papers have explored links between discounting biases and field behavior using direct elicitations on narrower samples, with narrower sets of covariates; see e.g., Chabris et al. (2008), Meier and Sprenger (2010), Burks et al. (2012), and Li et al. (2015).

[6] Other key differences include Bradford et al. (2017) lacking controls for cognitive skills, and Eisenhauer and Ventura (2006) only controlling for income.

weeks in the future, with 15% exhibiting present bias (consume treat today, plan to eat healthy in the future) and 7% future bias (consume healthy today, plan to eat treat in the future).[7] Barcellos and Carvalho's ALP subjects answered similar questions in their baseline survey, albeit with only a one-week instead of a five-week delay, with 6% exhibiting present-bias and 9% future-bias. Read and van Leeuwen (1998) offer actual snacks to a convenience sample of employees in Amsterdam but do not calculate individual-level measures of bias. They do find substantial present-bias on average. We do not know of any prior work estimating correlations between measures of consumption discounting biases and field outcomes.

C. Inconsistency with General Axiom of Revealed Preference (and dominance avoidance)

Our third and fourth behavioral factors follow Choi et al. (2014), which measures choice inconsistency with standard economic rationality. Choice inconsistency could indicate a tendency to make poor (costly) decisions in field contexts; indeed, Choi et al. (2014) find that more choice inconsistency is conditionally correlated with less wealth in a representative sample of Dutch households.

We use the same task and user interface as in Choi et al. (2014) but abbreviate it from 25 decisions to 11.[8] Each decision confronts respondents with a linear budget constraint 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. 1,270 of the 1,427 individuals taking our second Round 1 module make all 11 decisions, and comprise our sample for measuring choice inconsistency.[9] See Data Appendix Figure 2 for an example.

Following Choi et al., we average across these 11 decisions, within-consumer, to benchmark choices against two different standards of rationality. One benchmark is a complete and transitive preference ordering adhering to the General Axiom of Revealed Preference (GARP),

[7] If we limit the sample to those who did not receive the informational/debiasing treatment about self-control in ALP module 212 (Barcellos and Carvalho), we find 15% with present bias and 8% with future bias (N=748).

[8] We were quite constrained on survey time and hence conducted a pilot in which we tested the feasibility of capturing roughly equivalent information with fewer rounds. 58 pilot-testers completed 25 rounds, and we estimated the correlation between measures of choice inconsistency calculated using the full 25 rounds, and just the first 11 rounds. These correlations are 0.62 and 0.88 for the two key measures.

[9] 1424 individuals view at least one of the instruction screens, 1,311 are recorded as completing at least one round of the task, and 1,270 are recorded as completing each of the 11 rounds.

as captured by the Afriat (1972) Critical Cost Efficiency Index. 1-CCEI can be interpreted as the subject's degree of choice inconsistency: the percentage points of potential earnings "wasted" per the GARP standard. But as Choi et al. discuss, consistency with GARP is not necessarily the most appealing measure of decision quality because it allows for violations of monotonicity with respect to first-order stochastic dominance (FOSD).[10] Hence, again following Choi et al., our second measure captures inconsistency with both GARP and FOSD.[11] Note that these measures of inconsistency are unidirectional: there is no such thing as being *overly* consistent.

Our distribution of individual-level CCEI estimates is nearly identical to Choi et al.'s— if we use only the first 11 rounds of choices from Choi et al. to maximize comparability to our setup. Our median (1-CCEI) is 0.002, suggesting nearly complete consistency with GARP. The mean is 0.05. The median (1-combined-CCEI), capturing FOSD violations as well, is 0.10, with a mean of 0.16. Choice inconsistency is substantially higher when using the full 25 rounds in both our pilot data and Choi et al. (e.g., mean CCEI of 0.12 in both samples), and we have verified that this is a mechanical effect (more rounds means more opportunities to exhibit inconsistency) rather than deterioration in consistency as rounds increase, by finding that CCEIs measured over small blocks of consecutive rounds remain constant as the average round number of those blocks increases.

Data Appendix Table 1 shows that our prevalence estimates are also nearly identical to those from the Choi et al (2014) data. In our data, 53% of subjects exhibit any inconsistency with GARP, and 96% exhibit any inconsistency with GARP or FOSD. If we set a 20pp threshold for classifying someone as inconsistent, only 7% are inconsistent with GARP, and 31% are inconsistent with GARP or FOSD. Looking more directly at heterogeneity, we see standard deviations of 0.08 and 0.18, and 10th-90th percentile ranges of 0.16 and 0.41.

Choi et al. find that choice inconsistency with GARP is conditionally correlated with lower net worth, but that choice inconsistency with GARP+dominance avoidance is not.

[10] E.g., someone who always allocates all tokens to account X is consistent with GARP if they are maximizing the utility function U(X, Y)=X. Someone with a more normatively appealing utility function—that generates utility over tokens or consumption per se—would be better off with the decision rule of always allocating all tokens to the cheaper account.

[11] The second measure calculates 1-CCEI across the subject's 11 actual decisions and "the mirror image of these data obtained by reversing the prices and the associated allocation for each observation" (Choi et al. p. 1528), for 22 data points per respondent in total.

D. *Risk attitude re: certainty (certainty premium)*

Behavioral researchers have long noted a seemingly disproportionate preference for certainty (PFC) among some consumers and posited various theories to explain it: Cumulative Prospect Theory (Daniel Kahneman and Tversky 1979; Amos Tversky and Kahneman 1992), Disappointment Aversion (Bell 1985; Loomes and Sugden 1986; Gul 1991), and u-v preferences (Neilson 1992; Schmidt 1998; Diecidue, Schmidt, and Wakker 2004). PFC may help to explain seemingly extreme risk averse behavior, which could in turn lead to lower wealth in the cross-section.

We use Callen et al.'s (2014) two-task method[12] for measuring a subject's *certainty premium* (CP).[13] Similar to Holt and Laury tasks, in one of the Callen et al. tasks subjects make 10 choices 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). Both Callen et al. and we fix Y and X at 450 and 150 (hypothetical dollars in our case, hypothetical Afghanis in theirs), fix p at 0.5, and have q range from 0.1 to 1.0 in increments of 0.1. In the other task, p = 1, so the subject chooses between a lottery and a certain option. Our two tasks are identical to Callen et al.'s except for the currency units. But our settings, implementation, and use of the elicited data are different. Callen et al. administer the tasks in-person, using trained surveyors, at polling centers and homes in Afghanistan. They use the data to examine the effects of violence on risk preferences.

1,463 of 1,505 (97%) of our subjects who started the tasks completed all 20 choices (compared to 977/1127 = 87% in Callen et al.). As is typical with Holt-Laury tasks, we exclude some subjects whose choices indicate miscomprehension of or inattention to the task. 11% of our subjects multiple-switch on our two-lottery task (compared to 10% in Callen et al.), and 9% of our subjects multiple-switch on the lottery vs. certain option tasks (compared to 13% in Callen et al.). 14% of our subjects switch too soon for monotonic utility in the two-lottery—in rows [2, 4] in the two-lottery task—compared to 13% in Callen et al. All told, 19% of our subjects exhibit a puzzling switch (17% in Callen et al.), leaving us with 1,188 usable observations. Of these

[12] Callen et al. describes its task as "a field-ready, two-question modification of the uncertainty equivalent presented in Andreoni and Sprenger (2016)."

[13] The Callen et al. tasks also elicit non-parametric measures of classical risk aversion: a higher switch point indicates greater risk aversion. We discuss these measures in Section 1-D of the paper.

subjects, 1,049 switch on both tasks, as is required to estimate CP. Of these 1,049, only 30% switch at the same point on both tasks, in contrast to 63% in Callen et al.

We estimate CP for each respondent i by imputing the likelihoods q* at which i expresses indifference as the midpoint of the q interval at which i switches, and then using the two likelihoods to estimate the indirect utility components of the CP formula. As Callen et al. detail, the CP "is defined in probability units of the high outcome, Y, such that one can refer to certainty of X being worth a specific percent chance of Y relative to its uncertain value." We estimate a mean CP of 0.16 in our sample (SD=0.24, median =0.15), compared to 0.37 (SD=0.15) in Callen et al. Their findings suggest that much of the difference could be explained by greater exposure to violence in their sample.

As Callen et al. detail, the sign of CP also carries broader information about preferences. CP = 0 indicates an expected utility maximizer. CP>0 indicates a preference for certainty (PFC), as in models of disappointment aversion or u-v preferences. We classify 77% of our sample as PFC type based on an any-deviation threshold. This falls to 73%, 60%, or 42% if we count only larger deviations >0 (5pp, 10pp, or 20pp) as behavioral. In Callen et al. 99.63% of the sample exhibits PFC. CP<0 indicates a cumulative prospect theory (CPT) type, and we classify 23%, 20%, 13% or 7% as CPT under the different deviation thresholds. We denote PFC as the standard bias, simply because CP>0 is far more common than CP<0 in both our data and Callen et al.'s.

Callen et al. find significant correlations between the CP and financial outcomes, in particular with avoiding late loan repayments,[14] but their data lack controls for cognitive skills and other B-factors.

E. Loss aversion/small-stakes risk aversion

Loss aversion refers to placing higher weight on losses than gains, in utility terms. It is one of the most influential concepts in the behavioral social sciences, with seminal papers—e.g., Tversky and Kahneman (1992) and Benartzi and Thaler (1995)—producing thousands of citations. Loss aversion has been implicated in various portfolio choices (Barberis 2013) and consumption dynamics (Kőszegi and Rabin 2009) that can lead to lower wealth.

[14] The theoretical mapping from late loan repayments to our indices of financial condition is unclear under limited liability, and the average relationship (not conditioning on borrowing) more ambiguous, since borrowing could lead to (weakly) greater or lesser wealth if consumers are behavioral (Zinman 2014).

We measure loss aversion using the two choices developed by Fehr and Goette (2007) in their study of the labor supply of bike messengers (see Abeler et al. (2011) for a similar elicitation method). Choice 1 is between a lottery with a 50% chance of winning $80 and a 50% chance of losing $50, and zero dollars. Choice two is between playing the lottery in Choice 1 six times, and zero dollars. As Fehr and Goette (FG) show, if subjects have reference-dependent preferences, then subjects who reject lottery 1 have a higher level of loss aversion than subjects who accept lottery 1, and subjects who reject both lotteries have a higher level of loss aversion than subjects who reject only lottery 1. In addition, if subjects' loss aversion is consistent across the two lotteries, then any individual who rejects lottery 2 should also reject lottery 1 because a rejection of lottery 2 implies a higher level of loss aversion than a rejection of only lottery 1. Other researchers have noted that, even in the absence of loss aversion, choosing Option B is compatible with small-stakes risk aversion.[15] We acknowledge this but use “loss aversion” instead of “loss aversion and/or small-stakes risk aversion” as shorthand. Small-stakes risk aversion is also often classified as behavioral because it is incompatible with expected utility theory (Rabin 2000).

Response rates suggest a high level of comfort with these questions; only two of our 1,515 subjects skip, and only two more who answer the first question do not answer the second. 37% of our 1,511 respondents reject both lotteries, consistent with relatively extreme loss aversion, compared to 45% of FG’s 42 subjects. Another 36% of our subjects accept both lotteries, consistent with classical behavior, compared to 33% in FG. The remaining 27% of our subjects (and 21% of FG’s) exhibit moderate loss aversion, playing one lottery but not the other, with our main difference from FG being that 14% of our subjects (vs. only 2% of theirs) exhibit the puzzling behavior of playing lottery 1 but not lottery 2. Although one wonders whether these 14% misunderstood the questions, we find only a bit of evidence in support of that interpretation: those playing the single but not compound lottery have slightly lower cognitive skills than other loss averters, conditional on our rich set of covariates, but actually have higher cognitive skills than the most-classical group. And playing the single but not the compound lottery is uncorrelated with our measure of ambiguity aversion, pushing against the interpretation that the

[15] A related point is that there is no known “model-free” method of eliciting loss aversion (Dean and Ortoleva 2018).

compound lottery is sufficiently complicated as to appear effectively ambiguous (Dean and Ortoleva 2018).

All told 64% of our subjects indicate some loss aversion, defined as rejecting one or both small-stakes lotteries, as do 67% in FG. In Abeler et al.'s (2011) student sample, 87% reject one or more of the four small-stakes lotteries with positive expected value. The Abeler et al. questions were also fielded in an ALP module from early 2013 used by Hwang (2016); 70% of that sample exhibits some loss aversion. In von Gaudecker et al.'s nationally representative Dutch sample, 86% exhibit some loss aversion, as inferred from structural estimation based on data from multiple price lists. We also order sets of deviations to indicate greater degrees of loss aversion, based on whether the individual respondent rejects the compound but not the single lottery, rejects the single but not the compound lottery, or rejects both.

Despite the massive amount of work on loss aversion, research exploring links between directly elicited measures of loss aversion and field behavior is only beginning. von Gaudecker et al. (2011) do not explore links between loss aversion and field behavior. Dimmock and Kouwenberg (2010) do, like von Gaudecker et al. using CentERdata, but lack many important covariates. Fehr and Goette (2007) find that loss aversion moderates the effect of a wage increase, but their sample includes only bike messengers and lacks measures of many other potentially moderating factors. Abeler et al. (2011) find that loss aversion is strongly correlated with effort choices in the lab among their student sample, but again they lack data on many covariates of interest. Hwang (2016) uses the Abeler et al. measures to infer a strong correlation between loss aversion and insurance holdings in an earlier ALP module, but lacks many important covariates and the only other behavioral factor considered is an interaction between loss aversion and a measure of the Gambler's Fallacy (labeled "Heuristics" in the Hwang paper).[16]

F. Narrow bracketing and dominated choice

Narrow bracketing refers to the tendency to make decisions in (relative) isolation, without full consideration of other choices and constraints. Rabin and Weizsacker (2009) show that

[16] Hwang (2016) also discusses the potential (mediating) role of narrow framing/bracketing but lacks a directly elicited measure of such.

narrow bracketing can lead to dominated choices—and hence expensive and wealth-reducing ones—given non-CARA preferences.

We measure narrow bracketing and dominated choice (NBDC) using two of the tasks in Rabin and Weizacker (2009). Each task instructs the subject to make two decisions. 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. RW administer their tasks with students and, like us, in a nationally representative online panel (Knowledge Networks in their case). Like us, payoffs are hypothetical for their online panel.

Our first task follows RW's Example 2, with Decision 1 between winning \$100 vs. a 50-50 chance of losing \$300 or winning \$700, and Decision 2 between losing \$400 vs. a 50-50 chance of losing \$900 or winning \$100.[17] As RW show, someone who is loss averse and risk-seeking in losses will, in isolation (narrow bracketing) tend to choose A over B, and D over C. But the combination AD is dominated with an expected loss of \$50 relative to BC. Hence a broad-bracketer will never choose AD. 29% of our subjects choose AD, compared to 53% in the most similar presentation in RW.

Our second task reproduces RW's Example 4, with Decision 1 between winning \$850 vs. a 50-50 chance of winning \$100 or winning \$1,600, and Decision 2 between losing \$650 vs. a 50-50 chance of losing \$1,550 or winning \$100. As in task one, a decision maker who rejects the risk in the first decision but accepts it in the second decision (A and D) violates dominance, here with an expected loss of \$75 relative to BC. 23% of our subjects choose AD, compared to 36% in the most similar presentation in RW. As RW discuss, a new feature of task two is that AD sacrifices expected value in the second decision, not in the first. This implies that for all broad-bracketing risk averters AC is optimal: it generates the highest available expected value at no variance. 50% of our subjects choose AC, compared to only 33% in the most similar presentation in RW. I.e., 50% of our subjects do NOT broad-bracket in this task, compared to 67% in RW.

Reassuringly, responses across our two tasks are correlated; this is especially reassuring given that the two tasks appear non-consecutively in the survey, hopefully dampening any

[17] Given the puzzling result that RW's Example 2 was relatively impervious to a broad-bracketing treatment, we changed our version slightly to avoid zero-amount payoffs. Thanks to Georg Weizsacker for this suggestion.

tendency for a mechanical correlation. E.g., the unconditional correlation between choosing AD across the two tasks is 0.34.

1,486 subjects complete both tasks (out of the 1,515 who respond to at least one of our questions in module 315). Putting the two tasks together to create summary indicators of narrow bracketing, we find 59% of our subjects exhibiting some narrow bracketing in the sense of not broad-bracketing on both tasks, while 13% narrow-bracket on both tasks. These are uni-directional indicators: we either classify someone as narrow-bracketing, or not. RW do not create summary indicators across tasks, but, as noted above, their subjects exhibit substantially more narrow bracketing at the task level than our subjects do.

Research linking directly-elicited measures of NBDC to field outcomes is just beginning. The only paper we know of in this vein, Gottlieb and Mitchell (2015), uses a different method for measuring narrow bracketing—one that does not allow for dominated choice—the Tversky and Kahneman (1981) "sensitivity to framing" questions regarding the policy response to an epidemic. 30% of subjects in the Health and Retirement Study choose different policy options under the two different frames, an indicator of framing sensitivity, and this indicator is negatively correlated with the holding of long-term care insurance, conditional on a rich set of covariates include a measure loss aversion.

G. Ambiguity aversion

Ambiguity aversion refers to a preference for known uncertainty over unknown uncertainty—preferring, for example, a less-than-50/50 gamble to one with unknown probabilities. It has been widely theorized that ambiguity aversion can explain various sub-optimal portfolio choices, and Dimmock et al. (2016) find that it is indeed conditionally correlated with lower stockholdings and worse diversification in their ALP sample (see also Dimmock, Kouwenberg, and Wakker (2016)).

We elicit a coarse measure of ambiguity aversion using just one or two questions about a game that pays $500 if you select a green ball. The first question offers the choice between a Bag One with 45 green and 55 yellow balls vs. a Bag Two of unknown composition. 1,397 subjects respond to this question (out of 1,427 who answer at least one of our questions on ALP module 352). 73% choose the 45-55 bag, and we label them ambiguity averse. The survey then asks these subjects how many green balls would need to be in Bag One to induce them to switch. We subtract this amount from 50, dropping the 99 subjects whose response to the second question is >45 (and the 10 subjects who do not respond), to obtain a continuous measure of ambiguity aversion that ranges from 0 (not averse in the first question) to 50 (most averse=== the three subjects who respond "zero" to the second question). The continuous measure (N=1,288) has a mean of 14 (median=10), and a SD of 13. If we impose a large-deviation threshold of 10 (20% of the max) for labeling someone as ambiguity averse, 50% of our sample exceeds this threshold and another 16% are at the threshold. Our elicitation does not distinguish between ambiguity-neutral and ambiguity-seeking choices (for more comprehensive but still tractable methods see, e.g., Dimmock, Kouwenberg et al. (2016), Dimmock, Kouwenberg, and Wakker (2016), Gneezy et al. (2015)), and so our measure of deviation from ambiguity-neutrality is one-sided.

Despite the coarseness of our elicitation, comparisons to other work suggest that it produces reliable data. Our ambiguity aversion indicator correlates with one constructed from Dimmock et al.'s elicitation in the ALP (0.14, p-value 0.0001, N=789), despite the elicitations taking place roughly 3 years apart. Prevalence at our 10pp large-deviation cutoff nearly matches that from Dimmock, Kouwenberg et al.'s (2016) ALP sample and Butler et al.'s (2014) Unicredit Clients' Survey sample from Italy, and our prevalence of any ambiguity aversion, 0.73 is similar to Dimmock, Kouwenberg, and Wakker's (2016) 0.68 from the Dutch version of the ALP .

Our examination of links to field behaviors builds on the papers by Dimmock and co-authors cited above, which estimate conditional correlations between ambiguity aversion and financial market behavior. We broaden the set of both outcomes and control variables (especially other B-factors).[18]

H. Overconfidence: Three varieties

Overconfidence has been implicated in excessive trading (Daniel and Hirshleifer 2015), over-borrowing on credit cards (Ausubel 1991), paying a premium for private equity (Moskowitz and Vissing-Jorgensen 2002; although see Kartashova 2014), and poor contract choice (Grubb 2015), any of which can reduce wealth and financial security.

[18] The other paper we know of examining correlations between ambiguity attitudes and field behavior is Sutter et al.'s (2013) study of adolescents in Austria.

We elicit three distinct measures of overconfidence, following e.g., Moore and Healy (2008).

The first measures it in level/absolute terms, by following the three Banks and Oldfield numeracy questions, in our second Round 1 module, with the question: *"How many of the last 3 questions (the ones on the disease, the lottery and the savings account) do you think you got correct?"* We then subtract the respondent's assessment from her actual score. 39% of 1,366 subjects are overconfident ("overestimation" per Moore and Healy) by this measure (with 32% overestimating by one question), while only 11% are underconfident (with 10% underestimating by one question). Larrick et al. (2007), Moore and Healy, and other studies use this method for measuring overestimation, but we are not aware of any that report individual-level prevalence estimates (they instead focus on task-level data, sample-level summary statistics, and/or correlates of cross-sectional heterogeneity in estimation patterns).

The second measures overconfidence in precision, as indicated by responding "100%" on two sets of questions about the likelihoods (of different possible Banks and Oldfield quiz scores or of future income increases). This is a coarse adaptation of the usual approaches of eliciting several confidence intervals or subjective probability distributions (Moore and Healy). In our data 34% of 1,345 responding to both sets respond 100% on >=1 set, and 10% on both.

The third measures confidence in placement (relative performance), using a self-ranking elicited before taking our number series test: *"We would like to know 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?"* We find a better-than-average effect in the sample as a whole (70% report a percentile>median) that disappears when we ask the same question immediately post-test, still not having revealed any scores (50% report a percentile>median). We also construct an individual-level measure of confidence in placement by subtracting the subject's actual ranking from his pre-test self-ranking (N=1,395). This measure is useful for capturing individual-level heterogeneity ordinally, but not for measuring prevalence because the actual ranking is based on a 15-question test and hence its percentiles are much coarser than the self-ranking.

We are not aware of any prior work exploring conditional correlations between the sorts of overconfidence measures described above and field outcomes.

I. Non-belief in the Law of Large Numbers

Under-weighting the importance of the Law of Large Numbers (LLN) can affect how individuals treat risk (as in the stock market), or how much data they demand before making decisions. In this sense non-belief in LLN (a.k.a. NBLLN) can act as an "enabling bias" for other biases like loss aversion (Benjamin, Rabin, and Raymond 2016).

Following Benjamin, Moore, and Rabin (see also D Kahneman and Tversky 1972; Benjamin, Rabin, and Raymond 2016), we measure non-belief in law of large numbers (NBLLN) using responses to the following question:

> *… say the computer flips the coin 1000 times, and counts the total number of heads. Please tell us what you think are the chances, in percentage terms, that the total number of heads will lie within the following ranges. Your answers should sum to 100.*

The ranges provided are [0, 480], [481, 519], and [520, 1000], and so the correct answers are 11, 78, 11.

1,375 subjects respond (out of the 1,427 who answer at least one of our questions in Module 352),[19] with mean (SD) responses of 27 (18), 42 (24), and 31 (20). We measure NBLLN using the distance between the subject's answer for the [481, 519] range and 78. Only one subject gets it exactly right. 87% underestimate; coupled with prior work, this result leads us to designate underestimation as the "standard" directional bias. The modal underestimator responds with 50 (18% of the sample). The other most-frequent responses are 25 (10%), 30 (9%), 33 (8%), and 40 (7%). Few underestimators—only 4% of the sample—are within 10pp of 78, and their mean distance is 43, with an SD of 17. 9% of the sample underestimates by 20pp or less. 13% overestimate relative to 78, with 5% of the sample quite close to correct at 80, and another 5% at 100. Benjamin, Moore, and Rabin (2017) do not calculate individual-level measures of underestimation or overestimation in their convenience sample, but do report that the sample means are 35%, 36%, and 29% for the three bins. The comparable figures in our data are 27%, 42%, and 31%.

[19] Only 26 subjects provide responses that do not sum to 100 after a prompt, and each response for an individual range is [0, 100], so we do not exclude any subjects from the analysis here.

We are not aware of any prior work exploring conditional correlations between directly-elicited NBLLN and field outcomes.

J. Gambler's Fallacies

The Gambler's Fallacies involve falsely attributing statistical dependence to statistically independent events, in either expecting one outcome to be less likely because it has happened recently (recent reds on roulette make black more likely in the future) or the reverse, a "hot hand" view that recent events are likely to be repeated. Gambler's fallacies can lead to overvaluation of financial expertise (or attending to misguided financial advice), and related portfolio choices like the active-fund puzzle, that can erode wealth (Rabin and Vayanos 2010). Because the hot hand fallacy is more closely linked to harmful financial behaviors such as "return-chasing" or over-valuing the talent of stock-pickers (Rabin and Vayanos 2010), for analyses linking the fallacies to field behavior we denote hot-hand as the "standard" bias and cold-hand as "non-standard."

We take a slice of Benjamin, Moore, and Rabin's (2017) elicitation for the fallacies:

> *"Imagine that we had a computer "flip" a fair coin… 10 times. The first 9 are all heads. What are the chances, in percentage terms, that the 10th flip will be a head?"*

1,392 subjects respond, out of the 1,427 respondents to module 352. The cold-hand fallacy implies a response < 50%, while the hot-hand fallacy implies a response > 50%. Our mean response is 45% (SD=25), which is consistent with the cold-hand but substantially above the 32% in Benjamin, Moore, and Rabin. Another indication that we find less evidence of the cold-hand fallacy is that, while they infer that "at the individual level, the gambler's fallacy [cold-hand] appears to be the predominant pattern of belief" (2013, p. 16), we find only 26% answering < "50." 14% of our sample responds with >"50" (over half of these responses are at "90" or "100"). So 60% of our sample answers correctly. Nearly everyone who responds with something other than "50" errs by a substantial amount—e.g., only 2 % of the sample is [30, 50) or (50, 70]. Sixteen percent of our sample answers "10,"[20] which Benjamin, Moore, and Rabin speculates is an indicator of miscomprehension; we find that while subjects with this indicator do

[20] 34% of the sample in Benjamin, Moore, and Raymond respond "10%" on one or more of their ten questions.

have significantly lower cognitive skills than the unbiased group, they actually have higher cognitive skills than the rest of subjects exhibiting a gambler's fallacy.

Dohmen et al. (2009) measure the fallacies using a similar elicitation that confronts a representative sample of 1,012 Germans, taking an in-person household survey, with:

Imagine you are tossing a fair coin. After eight tosses you observe the following result: tails-tails-tails-heads-tails-heads-heads-heads. What is the probability, in percent, that the next toss is "tails"?

986 of Dohmen et al.'s respondents provide some answer to this question, 95 of whom say "Don't know." Among the remaining 891, 23% exhibit cold-hand (compared to 26% in our sample), and 10% exhibit hot-hand (compared to 14% in our sample). Conditional on exhibiting cold-hand, on average subjects err by 29pp (40 pp in our sample). Conditional on exhibiting hot-hand, the mean subject error is 27pp (39pp in our sample).

Dohmen et al. also explore correlations between unemployment or bank overdrafts and their directly-elicited fallacy measures, conditional on age, gender, education, income, and wealth. They find evidence of positive correlations between hot-hand and unemployment and between cold-hand and overdrafting.

K. Exponential growth bias: Two varieties

Exponential growth bias (EGB) produces a tendency to underestimate the effects of compounding on costs of debt and benefits of saving. It has been linked to a broad set of financial outcomes (Levy and Tasoff 2016; Stango and Zinman 2009).

We measure EGB, following previous papers, by asking respondents to solve questions regarding an asset's future value or a loan's implied annual percentage rate. Our first measure of EGB follows in the spirit of Stango and Zinman (2009, 2011) by first eliciting the monthly payment the respondent would expect to pay on a $10,000, 48 month car loan. The survey then asks "… What percent rate of interest does that imply in annual percentage rate ("APR") terms?" 1,445 panelists answer both questions, out of the 1,515 respondents to Module 315. Most responses appear sensible given market rates; e.g., there are mass points at 5%, 10%, 3%, 6% and 4%.

We calculate an individual-level measure of “debt-side EGB” by comparing the difference between the APR *implied* by the monthly payment supplied by that individual, and the *perceived* APR as supplied directly by the same individual. We start by binning individuals into under-estimators (the standard bias), over-estimators, unbiased, and unknown (37% of the sample).[21] The median level difference between the correct and stated value is 500bp, with a mean of 1,042bp and SD of 1,879bp. Among those with known bias, we count as biased 70%, 64%, 47%, and 34% under error tolerance of zero, 100bp, 500bp, and 1000bp. Under these tolerances we count 3%, 13%, 41%, and 61% as unbiased, and 27%, 22%, 10%, and 3% as negatively biased. This is less EGB than Stango and Zinman (2009, 2011) see from questions in the 1983 Survey of Consumer Finances, where 98% of the sample underestimates, and the mean bias is 1,800bp or 3,800bp depending on the benchmark. The time frames of the questions differ, which may account for the difference (and is why we do not estimate an EGB structural model parameter to compare with our prior work or that of Levy and Tasoff).

Stango and Zinman (2009; 2011) find that more debt-side EGB is strongly conditionally correlated with debt composition, worse loan terms, and less savings and wealth. But those papers lack direct controls for cognitive skills and other behavioral factors.

Our second measure of EGB comes from a question popularized by Banks and Oldfield (2007) as part of a series designed to measure basic numeracy: “Let's say you have $200 in a savings account. The account earns 10 percent interest per year. You don’t withdraw any money for two years. How much would you have in the account at the end of two years?” 1,389 subjects answer this question (out of the 1,427 respondents to Module 352), and we infer an individual-level measure of “asset-side EGB” by comparing the difference between the correct future value ($242), and the future value supplied by the same individual.[22] We again bin individuals into underestimators (the standard bias), overestimators, unbiased, and unknown (14% of the

[21] Non-response is relatively small, as only 4% of the sample does not respond to both questions. Most of those we label as unknown-bias give responses that imply or state a 0% APR. 7% state payment amounts that imply a negative APR, even after being prompted to reconsider their answer. We also classify the 4% of respondents with implied APRs >=100% as having unknown bias.

[22] Responses to this question are correlated with responses to two other questions, drawn from Levy and Tasoff (2016), that can also be used to measure asset-side EGB, but our sample sizes are smaller for those two other questions and hence we do not use them here.

sample).[23] Among those with known bias (N=1,222), the median bias is $0, with a mean of $2 and SD of $14.[24] 44% of our sample provides the correct FV. 47% of our sample underestimates by some amount, with most underestimators (29% of the sample) providing the linearized (uncompounded) answer of $240. Nearly all other underestimates provide an answer that fails to account for even simple interest; the most common reply in this range is “$220.” Only 9% of our sample overestimates the FV, with small mass points at 244, 250, 400, and 440.

Other papers have used the Banks and Oldfield question, always—to our knowledge—measuring accuracy as opposed to directional bias and then using a 1/0 measure of correctness as an input to a financial literacy or numeracy score (e.g., James Banks, O’Dea, and Oldfield 2010; Gustman, Steinmeier, and Tabatabai 2012). Our tabs from the 2014 Health and Retirement Study suggest, using only the youngest HRS respondents and our oldest respondents to maximize comparability (ages 50-60 in both samples), that there is substantially more underestimation in the HRS (74%, vs. 48% in our sample). 14% overestimate in the HRS among those aged 50-60, vs. 9% in our sample.

Goda et al. (2017) and Levy and Tasoff (2016) measure asset-side EGB using more difficult questions in their representative samples. They find that 9% and 11% overestimate FVs, while 69% and 85% underestimate. We do not construct an EGB parameter to compare to theirs, because our questions lack their richness and yield heavy mass points at unbiased and linear-biased responses.

The only prior paper we know looking directly at links between a measure of asset-side EGB and field outcomes is Goda et al., who use data on fewer behavioral factors. They find significant negative correlations between asset-side EGB and retirement savings.

[23] We label as unknown the 8% of the sample answering with future value < present value, the 3% of the sample answering with a future value > 2x the correct future value, and the 3% of the sample who skip this question.

[24] For calculating the mean and SD we truncate bias at -42 for the 4% sample answering with future values 284<FV<485, to create symmetric extrema in the bias distribution since our definition caps bias at 42.

L. Limited attention and limited memory

Prior empirical work has found that limited attention affects a range of financial decisions (e.g., Barber and Odean 2008; DellaVigna and Pollet 2009; Karlan et al. 2016; Stango and Zinman 2014). Behavioral inattention is a very active line of theory inquiry as well (e.g., Bordalo, Gennaioli, and Shleifer 2017; Kőszegi and Szeidl 2013; Schwartzstein 2014).

In the absence of widely used methods for measuring limited attention and/or memory, we create our own, using five simple questions and tasks.

The first three ask, "Do you believe that your household's [horizon] finances… would improve if your household paid more attention to them?" for three different horizons: "day-to-day (dealing with routine expenses, checking credit card accounts, bill payments, etc.)" "medium-run (dealing with periodic expenses like car repair, kids' activities, vacations, etc.)" and "long-run (dealing with kids' college, retirement planning, allocation of savings/investments, etc.)" Response options are the same for each of these three questions: "Yes, and I/we often regret not paying greater attention" (26%, 23%, and 35%), "Yes, but paying more attention would require too much time/effort" (8%, 11%, and 12%), "No, my household finances are set up so that they don't require much attention" (15%, 16%, and 13%), and "No, my household is already very attentive to these matters" (52%, 51%, and 41%). We designed the question wording and response options to distinguish behavioral limited inattention ("Yes… I/we often…")—which also includes a measure of awareness thereof in "regret"—from full attention ("… already very attentive"), rational inattention, and/or a sophisticated response to behavioral inattention ("Yes, but… too much time/effort"; "… set up so that they don't require much attention").

Responses are strongly but not perfectly correlated (ranging 0.56 to 0.69 among pairwise expressions of regret). A fourth measure of limited attention is also strongly correlated with the others, based on the question: "Do you believe that you could improve the prices/terms your household typically receives on financial products/services by shopping more?"[25] 18% respond "Yes, and I/we often regret not shopping more," and the likelihood of this response is correlated 0.25 with each of the regret measures above. 1,483 subjects answer all four questions, out of the

[25] This question is motivated by evidence that shopping behavior strongly predicts borrowing costs (Stango and Zinman 2016).

1,515 respondents to Module 315. Summing the four indicators of attentional regret, we find that 49% of subjects have one or more (earning a classification of behavioral inattention), 29% have two or more, 19% three or more, and only 6% have all four.

We also seek to measure limited prospective memory, following previous work suggesting that limited memory entails real costs like forgetting to redeem rebates (e.g., Ericson 2011). We offer an incentivized task to subjects taking module 352: “The ALP will offer you the opportunity to earn an extra $10 for one minute of your time. This special survey has just a few simple questions but will only be open for 24 hours, starting 24 hours from now. During this specified time window, you can access the special survey from your ALP account. So we can get a sense of what our response rate might be, please tell us now whether you expect to do this special survey.” 97% say they intend to complete the short survey, leaving us with a sample of 1,358. Only 14% actually complete the short survey.

Our indicator of behavioral limited memory— (not completing the follow-up task conditional on intending to complete)—is a bit coarse. We suspect that some noise is introduced because our elicitation makes it costless to express an intention to complete (in future research we plan to explore charging a small “sign up” fee), thereby including in the indicator’s sample frame some subjects who rationally do not complete the task. Relatedly, although we set the payoff for task completion to be sufficiently high to dominate any attention/memory/time costs in *marginal* terms for most subjects (the effective hourly wage is in the hundreds of dollars), it may well be the case that the *fixed* cost exceeds $10 for some respondents.

Ours is the first work we know of estimating conditional correlations between field outcomes and directly elicited measures of limited attention/memory in a broad sample.

2. Measuring Cognitive Skills

We measure fluid intelligence using a 15-question, non-adaptive number series (McArdle, Fisher, and Kadlec 2007). Number series scores correlate strongly with those from other fluid intelligence tests like IQ and Raven's.

We measure numeracy using: "If 5 people split lottery winnings of two million dollars ($2,000,000) into 5 equal shares, how much will each of them get?" and "If the chance of getting a disease is 10 percent, how many people out of 1,000 would be expected to get the disease?" (Banks and Oldfield 2007). Response options are open-ended. These questions have been used in economics as numeracy and/or financial literacy measures since their deployment in the 2002 English Longitudinal Study of Ageing, with subsequent deployment in the Health and Retirement Study and other national surveys.

We measure financial literacy using Lusardi and Mitchell's (2014) "Big Three": "Suppose you had $100 in a savings account and the interest rate was 2% per year. After 5 years, how much do you think you would have in the account if you left the money to grow?"; "Imagine that the interest rate on your savings account was 1% per year and inflation was 2% per year. After 1 year, how much would you be able to buy with the money in this account?"; and "Please tell me whether this statement is true or false: "Buying a single company's stock usually provides a safer return than a stock mutual fund." Response options are categorical.

We measure executive function using a two-minute Stroop task (MacLeod 1991). Our version displays the name of a color on the screen (red, blue, green, or yellow) and asks the subject to click on the button corresponding to the color the word is printed in (red, blue, green, or yellow; not necessarily corresponding to the color name). Answering correctly tends to require using conscious effort to override the tendency (automatic response) to select the name rather than the color. The Stroop task is sufficiently classic that the generic failure to overcome automated behavior (in the game "Simon Says," when an American crosses the street in England, etc.) is sometimes referred to as a "Stroop Mistake" (Camerer 2007). Before starting the task, the computer shows demonstrations of two choices (movie-style)—one with a correct response, and one with an incorrect response—and then gives the subject the opportunity to practice two choices on her own. After practice ends, the task lasts for two minutes.

3. Survey Formatting and Non-classical Measurement Error

Data Appendix Table 3 provides reassurance that, *a priori*, there is little reason to think that low survey effort *per se* could contribute to a mechanical correlation between worse financial condition and more behavioral biases. A necessary condition for that confound is that it is somehow easier, from a survey effort perspective, to indicate worse than better financial condition. The table shows that this is unlikely to be the case, given how questions are scripted and response options are arrayed.

Data Appendix Table 4 provides some additional descriptive reassurance with data, showing a lack of systematic relationship between survey time spent (across all questions for both Round 1 modules) and financial condition responses, with the possible exception of the lowest time spent quintile.

As the main text details, we deal with this potential confound formally, by controlling flexibly for survey effort in both survey rounds with flexible controls for non-response and for survey time spent, and by dropping those in the lowest decile of time spent as a robustness check.

References for Data Appendix

Abeler, Johannes, Armin Falk, Lorenz Goette, and David Huffman. 2011. “Reference Points and Effort Provision.” *American Economic Review* 101 (2): 470–92.

Afriat, S. N. 1972. “Efficiency Estimation of Production Functions.” *International Economic Review* 13 (3): 568.

Andreoni, James, and Charles Sprenger. 2012. “Estimating Time Preferences from Convex Budgets.” *The American Economic Review* 102 (7): 3333–56.

———. 2016. “Prospect Theory Revisited: Unconfounded Experimental Tests of Probability Weighting.”

Augenblick, Ned, Muriel Niederle, and Charles Sprenger. 2015. “Working over Time: Dynamic Inconsistency in Real Effort Tasks.” *The Quarterly Journal of Economics* 130 (3): 1067–1115.

Banks, J., and Z. Oldfield. 2007. “Understanding Pensions: Cognitive Function, Numerical Ability, and Retirement Saving.” *Fiscal Studies* 28 (2): 143–70.

Banks, James, Cormac O’Dea, and Zoë Oldfield. 2010. “Cognitive Function, Numeracy and Retirement Saving Trajectories.” *The Economic Journal* 120 (548): F381–410.

Barber, Brad, and Terrence Odean. 2008. “All That Glitters: The Effect of Attention on the Buying Behavior of Individual and Institutional Investors.” *Review of Financial Studies* 21 (2): 785–818.

Barberis, Nicholas C. 2013. “Thirty Years of Prospect Theory in Economics: A Review and Assessment.” *Journal of Economic Perspectives* 27 (1): 173–96.

Barcellos, Silvia, and Leandro Carvalho. 2014. “Information about Self-Control and Intertemporal Choices.”

Bell, David E. 1985. “Disappointment in Decision Making under Uncertainty.” *Operations Research* 33 (1): 1–27.

Benartzi, S., and R. H. Thaler. 1995. “Myopic Loss Aversion and the Equity Premium Puzzle.” *The Quarterly Journal of Economics* 110 (1): 73–92.

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.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2017. “Memory, Attention, and Choice.” *NBER Working Paper #23256*.

Bradford, David, Charles Courtemanche, Garth Heutel, Patrick McAlvanah, and Christopher Ruhm. 2017. “Time Preferences and Consumer Behavior.” *Journal of Risk and Uncertainty* 55 (2–3): 119–45.

Burks, Stephen, Jeffrey Carpenter, Lorenz Götte, and Aldo Rustichini. 2012. “Which Measures of Time Preference Best Predict Outcomes: Evidence from a Large-Scale Field Experiment.” *Journal of Economic Behavior & Organization* 84 (1): 308–20.

Butler, Jeffrey V., Luigi Guiso, and Tullio Jappelli. 2014. “The Role of Intuition and Reasoning in Driving Aversion to Risk and Ambiguity.” *Theory and Decision* 77 (4): 455–84.

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.

Chabris, Christopher F., David Laibson, Carrie L. Morris, Jonathon P. Schuldt, and Dmitry Taubinsky. 2008. “Individual Laboratory-Measured Discount Rates Predict Field Behavior.” *Journal of Risk and Uncertainty* 37 (2–3): 237–69.

Choi, Syngjoo, Shachar Kariv, Wieland Müller, and Dan Silverman. 2014. “Who Is (More) Rational?” *American Economic Review* 104 (6): 1518–50.

Daniel, Kent, and David Hirshleifer. 2015. “Overconfident Investors, Predictable Returns, and Excessive Trading.” *Journal of Economic Perspectives* 29 (4): 61–88.

Dean, Mark, and Pietro Ortoleva. 2018. “Is It All Connected? A Testing Ground for Unified Theories of Behavioral Economics Phenomena.”

DellaVigna, Stefano, and Joshua M Pollet. 2009. “Investor Inattention and Friday Earnings Announcements.” *The Journal of Finance* 64 (2): 709–49.

Diecidue, Enrico, Ulrich Schmidt, and Peter P Wakker. 2004. “The Utility of Gambling Reconsidered.” *Journal of Risk and Uncertainty* 29 (3): 241–59.

Dimmock, Stephen G., and Roy Kouwenberg. 2010. “Loss-Aversion and Household Portfolio Choice.” *Journal of Empirical Finance* 17 (3): 441–59.

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.

Dimmock, Stephen, Roy Kouwenberg, and Peter P Wakker. 2016. “Ambiguity Attitudes in a Large Representative Sample.” *Management Science* 62 (5): 1363–80.

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.

Eisenhauer, Joseph G, and Luigi Ventura. 2006. “The Prevalence of Hyperbolic Discounting: Some European Evidence.” *Applied Economics* 38 (11): 1223–34.

Ericson, Keith. 2011. “Forgetting We Forget: Overconfidence and Memory.” *Journal of the European Economic Association* 9 (1): 43–60.

Falk, Armin, Anke Becker, Thomas Dohmen, David Huffman, and Uwe Sunde. 2016. “The Preference Survey Module: A Validated Instrument for Measuring Risk, Time, and Social Preferences.”

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.

Giné, Xavier, Jessica Goldberg, Dan Silverman, and Dean Yang. 2018. “Revising Commitments: Field Evidence on the Adjustment of Prior Choices.” *The Economic Journal* 128 (608): 159–88.

Gneezy, Uri, Alex Imas, and John List. 2015. “Estimating Individual Ambiguity Aversion: A Simple Approach.”

Goda, Gopi Shah, Matthew R Levy, Colleen Flaherty Manchester, Aaron Sojourner, and Joshua Tasoff. 2017. “Predicting Retirement Savings Using Survey Measures of Exponential-Growth Bias and Present Bias.”

Gottlieb, Daniel, and Olivia S. Mitchell. 2015. “Narrow Framing and Long-Term Care Insurance.”

Graham, Liam, and Dennis J Snower. 2013. “Hyperbolic Discounting and Positive Optimal Inflation.” *Macroeconomic Dynamics* 17 (03): 591–620.

Grubb, Michael D. 2015. “Overconfident Consumers in the Marketplace.” *Journal of Economic Perspectives* 29 (4): 9–36.

Gul, Faruk. 1991. “A Theory of Disappointment Aversion.” *Econometrica: Journal of the Econometric Society* 59 (3): 667–86.

Gustman, Alan L, Thomas L Steinmeier, and Nahid Tabatabai. 2012. “Financial Knowledge and Financial Literacy at the Household Level.” *American Economic Review* 102 (3): 309–13.

Harris, Christopher, and David Laibson. 2013. “Instantaneous Gratification.” *The Quarterly Journal of Economics* 128 (1): 205–48.

Hwang, In Do. 2016. “Prospect Theory and Insurance Demand.”

İmrohoroğlu, Ayşe, Selahattin İmrohoroğlu, and Douglas H. Joines. 2003. “Time-Inconsistent Preferences and Social Security.” *The Quarterly Journal of Economics* 118 (2): 745–84.

Kahneman, D, and A Tversky. 1972. “Subjective Probability: A Judgement of Representativeness.” *Cognitive Psychology* 3: 430–54.

Kahneman, Daniel, and Amos Tversky. 1979. “Prospect Theory: An Analysis of Decision Under Risk.” *Econometrica* 47 (2): 263–91.

Karlan, Dean, Margaret McConnell, Sendhil Mullainathan, and Jonathan Zinman. 2016. “Getting to the Top of Mind: How Reminders Increase Saving.” *Management Science* 62 (12): 3393–3411.

Kartashova, Katya. 2014. “Private Equity Premium Puzzle Revisited.” *American Economic Review* 104 (10): 3297–3334.

Kőszegi, Botond, and Matthew Rabin. 2009. “Reference-Dependent Consumption Plans.” *American Economic Review* 99 (3): 909–36.

Kőszegi, Botond, and Adam Szeidl. 2013. “A Model of Focusing in Economic Choice.” *The Quarterly Journal of Economics* 128 (1): 53–104.

Laibson, David. 1997. “Golden Eggs and Hyperbolic Discounting.” *Quarterly Journal of Economics* 112 (2): 443–77.

Larrick, Richard P, Katherine A Burson, and Jack B Soll. 2007. “Social Comparison and Confidence: When Thinking You’re Better than Average Predicts Overconfidence (and When It Does Not).” *Organizational Behavior and Human Decision Processes* 102 (1): 76–94.

Levy, Matthew, and Joshua Tasoff. 2016. “Exponential-Growth Bias and Lifecycle Consumption.” *Journal of the European Economic Association* 14 (3): 545–83.

Li, Ye, Jie Gao, A. Zeynep Enkavi, Lisa Zaval, Elke U. Weber, and Eric J. Johnson. 2015. “Sound Credit Scores and Financial Decisions despite Cognitive Aging.” *Proceedings of the National Academy of Sciences* 112 (1): 65–69.

Loomes, Graham, and Robert Sugden. 1986. “Disappointment and Dynamic Consistency in Choice under Uncertainty.” *The Review of Economic Studies* 53 (2): 271–82.

Meier, Stephan, and Charles Sprenger. 2010. “Present-Biased Preferences and Credit Card Borrowing.” *American Economic Journal: Applied Economics* 2 (1): 193–210.

Moore, Don A., and Paul J. Healy. 2008. “The Trouble with Overconfidence.” *Psychological Review* 115 (2): 502–17.

Moskowitz, T.J., and A. Vissing-Jorgensen. 2002. “The Returns to Entrepreneurial Investment: A Private Equity Premium Puzzle.” *American Economic Review* 92 (4): 745–78.

Neilson, William S. 1992. “Some Mixed Results on Boundary Effects.” *Economics Letters* 39 (3): 275–78.

Pérez Kakabadse, Alonso, and Ignacio Palacios Huerta. 2013. “Consumption and Portfolio Rules with Stochastic Hyperbolic Discounting.”

Rabin, Matthew. 2000. “Risk Aversion and Expected-Utility Theory: A Calibration Theorem.” *Econometrica* 68 (5): 1281–92.

Rabin, Matthew, and Dimitri Vayanos. 2010. “The Gambler’s and Hot-Hand Fallacies: Theory and Applications.” *Review of Economic Studies* 77 (2): 730–78.

Rabin, Matthew, and Georg Weizsäcker. 2009. “Narrow Bracketing and Dominated Choices.” *American Economic Review* 99 (4): 1508–43.

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.

Schmidt, Ulrich. 1998. “A Measurement of the Certainty Effect.” *Journal of Mathematical Psychology* 42 (1): 32–47.

Schwartzstein, Joshua. 2014. “Selective Attention and Learning.” *Journal of the European Economic Association* 12 (6): 1423–52.

Stango, Victor, and Jonathan Zinman. 2009. “Exponential Growth Bias and Household Finance.” *The Journal of Finance* 64 (6): 2807–49.

———. 2011. “Fuzzy Math, Disclosure Regulation, and Credit Market Outcomes: Evidence from Truth-in-Lending Reform.” *Review of Financial Studies* 24 (2): 506–34.

———. 2014. “Limited and Varying Consumer Attention: Evidence from Shocks to the Salience of Bank Overdraft Fees.” *Review of Financial Studies* 27 (4): 990–1030.

———. 2016. “Borrowing High vs. Borrowing Higher: Price Dispersion and Shopping Behavior in the U.S. Credit Card Market.” *Review of Financial Studies* 29 (4): 979–1006.

Sutter, Matthias, Martin G Kocher, Daniela Glätzle-Rützler, and Stefan T Trautmann. 2013. “Impatience and Uncertainty: Experimental Decisions Predict Adolescents’ Field Behavior.” *American Economic Review* 103 (1): 510–31.

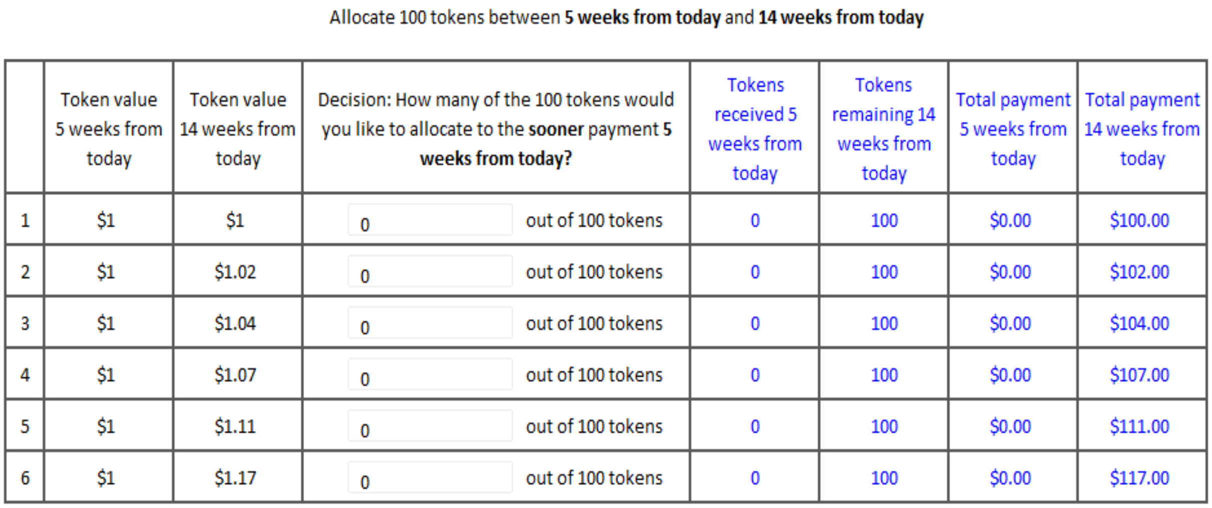
Toubia, Olivier, Eric Johnson, Theodoros Evgeniou, and Philippe Delquié. 2013. “Dynamic Experiments for Estimating Preferences: An Adaptive Method of Eliciting Time and Risk Parameters.” *Management Science* 59 (3): 613–40.

Tversky, A., and D. Kahneman. 1981. “The Framing of Decisions and the Psychology of Choice.” *Science* 211 (4481): 453–58.

Tversky, Amos, and Daniel Kahneman. 1992. “Advances in Prospect Theory: Cumulative Representation of Uncertainty.” *Journal of Risk and Uncertainty* 5 (4): 297–323.

Von Gaudecker, Hans-Martin, Arthur Van Soest, and Erik Wengström. 2011. “Heterogeneity in Risky Choice Behavior in a Broad Population.” *The American Economic Review* 101 (2): 664–94.

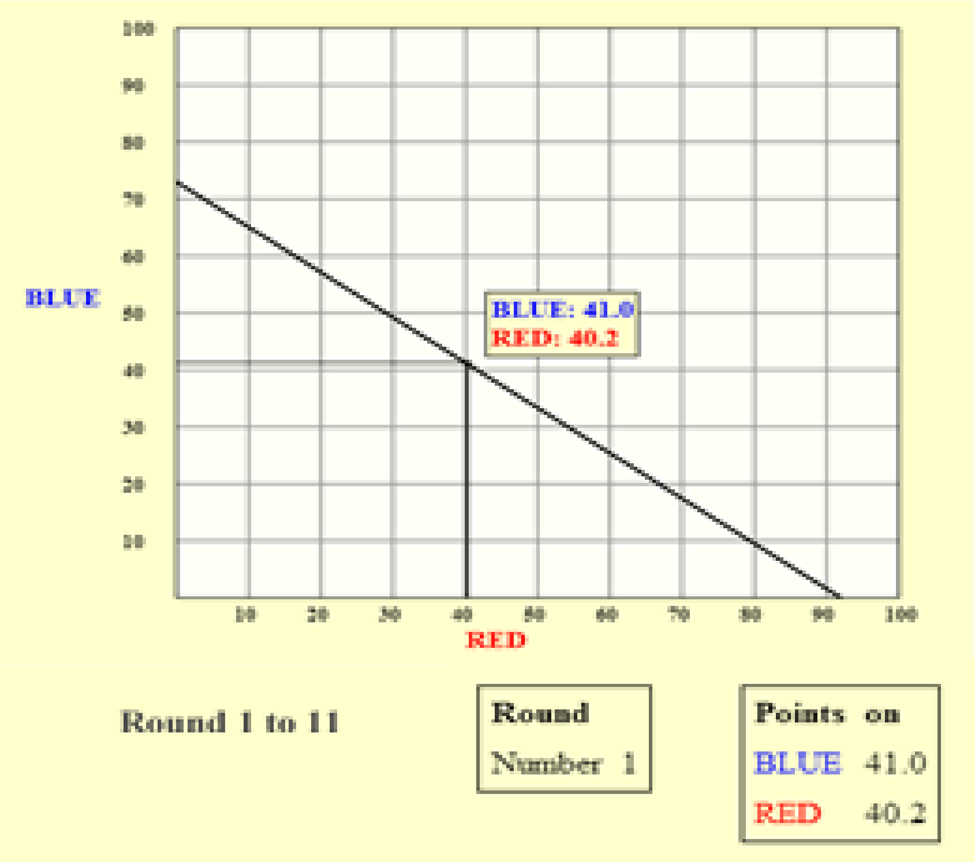
Zinman, Jonathan. 2014. “Consumer Credit: Too Much or Too Little (or Just Right)?” *Journal of Legal Studies* 43 (S2 Special Issue on Benefit-Cost Analysis of Financial Regulation): S209–37.



Allocate 100 tokens between **5 weeks from today** and **14 weeks from today**

	Token value 5 weeks from today	Token value 14 weeks from today	Decision: How many of the 100 tokens would you like to allocate to the **sooner** payment **5 weeks from today?**	Tokens received 5 weeks from today	Tokens remaining 14 weeks from today	Total payment 5 weeks from today	Total payment 14 weeks from today
1	$1	$1	0 out of 100 tokens	0	100	$0.00	$100.00
2	$1	$1.02	0 out of 100 tokens	0	100	$0.00	$102.00
3	$1	$1.04	0 out of 100 tokens	0	100	$0.00	$104.00
4	$1	$1.07	0 out of 100 tokens	0	100	$0.00	$107.00
5	$1	$1.11	0 out of 100 tokens	0	100	$0.00	$111.00
6	$1	$1.17	0 out of 100 tokens	0	100	$0.00	$117.00

Data Appendix Figure 1. Discounting choices, screenshot
(1 of 4 screens, 6 choices per screen)



Round 1 to 11

Round Number 1

Points on
BLUE 41.0
RED 40.2

Data Appendix Figure 2. Consistency with GARP choices, screenshot
(1 of 11 rounds, 1 choice per round).

Data Appendix Table 1. Behavioral bias prevalence: Comparisons to prior work using representative samples

	(U.S. samples in bold)		
	Our sample	Prior work	
		Comp 1	Comp 2
Time-inconsistent money discounting: Present-biased	**0.26**	**0.29**[1]	**0.55**[2]
Time-inconsistent money discounting: Future-biased	**0.36**	**0.37**	
Time-inconsistent snack discounting: Present-biased	**0.15**	**0.06**[1]	
Time-inconsistent snack discounting: Future-biased	**0.07**	**0.09**	
Violates GARP	**0.53**	0.51[3]	
Violates GARP plus dominance avoidance	**0.96**	0.96	
Loss-averse	**0.64**	**0.70**[4]	0.86[5]
Narrow-brackets	**0.59**		**0.30**[7]
	Task 2: 0.29	**Task 2: 0.53**[6]	
	Task 4: 0.50	**Task 4: 0.67**	
Ambiguity-averse	**0.73**	**0.52**[8]	0.68[9]
Gambler's Fallacy: Hot hand	**0.14**	0.10	
Gambler's fallacy: Cold hand	**0.26**	0.23[10]	
Exponential growth bias, loan-side: Underestimates APR	**0.7**	**0.98**[11]	
Exponential growth bias, loan-side: Overestimates APR	**0.27**	**0.00**	
Exponential growth bias, asset-side: Underestimates FV	**0.47**	**0.69**[2]	**0.85**[12]
Exponential growth bias, asset-side: Overestimates FV	**0.09**	**0.09**	**0.11**

Notes: The B-factors not listed here but included in other tables are those for which we could not find a prevalence estimate from a representative sample. See Data Appendix for details on elicitations, prevalence and distributions. In some cases we take comparisons directly from prior work, and in others we use data from other papers to perform our own calculations. "GARP" = General Axiom of Revealed Preference. "APR" = Annual Percentage Rate. "FV" = Future Value.

Footnotes:

[1] - Barcellos and Carvahlo (2014), source data are from ALP.

[2] - Goda et al. (2017), sources are ALP and Understanding America Survey.

[3] - Choi et al. (2011), source is CentER panel (Netherlands).

[4] - Hwang (2016), source is ALP. We define loss aversion as rejecting one or more of the four small-stakes lotteries with positive expected value.

[5] - von Gaudeker et al. (2011), source is CentER panel (Netherlands).

[6] - Rabin and Weizacker (2009), source is KnowledgeNetworks

[7] - Gottleib and Mitchell (2015), source is Health and Retirement Study (older Americans).

[8] - Dimmock et al. (2016), source is ALP.

[9] - Dimmock, Kouwenberg and Wakker (forthcoming), source is CentER panel (Netherlands).

[10] - Dohmen et al. (2009), source is German SocioEconomic Panel.

[11] - Stango and Zinman (2009, 2011), source is Survey of Consumer Finances.

[12] - Levy and Tasoff (2016), source is KnowledgeNetworks

Data Appendix Table 2. Survey formatting should not bias toward worse financial condition reporting

		response options			
Variable	# of questions used	# response options per q.	orientation	placement of choice(s) indicating worse condition	ordering details
net worth>0	1	3	vertical	middle	Assets compared to debts? [Yes/no/about the same]
retirement assets>0	2	2	vertical	n/a*	"Enter total amount: $[fill].00"
owns stocks	3	2	vertical	n/a*	"About what percent of your household's [IRA/KEOGH; 401(k)/other retirement accounts] are invested in stocks or mutual funds (not including money market mutual funds)?"
				n/a*	Aside from anything you have already told us about, do you or another member of your household have any shares of stock or stock mutual funds? If you sold all those and paid off anything you owed on them, about how much would your household have?
spent < income last 12 months	1	3	vertical	top	Spent [more than/same as/less than] income
financial satisfaction	1	slider	horizontal	left side of scale	0 to 100 point scale, lower numbers indicate lower satisfaction
retirement saving adequate	1	5	vertical	top	Ordered 1/5 from "not nearly enough" to "much more than enough"
non-retirement saving adequate	1	5	vertical	bottom	Ordered 1/5 from "wish my household saved a lot less" to "wish my household saved a lot more"
severe distress last 12 mos	4	2	vertical	top	Yes/no for each question, with yes on top.
financial stress	1	slider	horizontal	right side of scale	0 to 100 point scale, higher numbers indicate higher stress

Variables here are the components of our objective and subjective financial condition indices; see Appendix Table 3 for more details.

* - these responses provided check-boxes indicating "zero" as answers, below the section for the continuous response.

Data Appendix Table 3. Survey response time and financial condition components

	Financial condition component outcomes: Share with indicator of better condition									
	"Hard" outcomes: Balance sheet positions, flows, and events					"Soft" outcomes: Subjective perceptions				
Survey time decile	net worth>0	retirement assets>0	owns stocks	no severe distress last 12 months	spent < income last 12 months	financial satisfaction > median	retirement saving adequate	non-ret saving adequate	fin stress < median	Overall
1	0.33	0.36	0.35	0.60	0.30	0.40	0.27	0.25	0.46	0.37
2	0.37	0.50	0.46	0.60	0.36	0.50	0.26	0.31	0.52	0.43
3	0.47	0.53	0.52	0.58	0.35	0.45	0.24	0.26	0.55	0.44
4	0.52	0.60	0.54	0.60	0.37	0.40	0.28	0.24	0.48	0.45
5	0.47	0.61	0.55	0.56	0.43	0.47	0.26	0.20	0.52	0.45
6	0.49	0.59	0.57	0.59	0.42	0.51	0.25	0.25	0.51	0.46
7	0.50	0.54	0.50	0.47	0.29	0.49	0.21	0.29	0.49	0.42
8	0.46	0.58	0.48	0.51	0.36	0.44	0.30	0.24	0.56	0.44
9	0.41	0.48	0.46	0.58	0.33	0.44	0.24	0.25	0.48	0.41
10	0.39	0.52	0.49	0.50	0.36	0.50	0.35	0.22	0.50	0.42

Notes: Survey time decile is for total survey completion time in minutes. Financial condition components are described in greater detail in Table 4.