ELICITING EXPECTATIONS *

SAMUEL M. HARTZMARK

ABIGAIL B. SUSSMAN

August 2024

Abstract

We document how variation in the question format used to elicit belief distributions of percentage changes, such as return expectations, influences reported beliefs. We induce distributions to serve as a ground truth and vary how we ask about these distributions. Participants effectively report *differences* in means and volatilities irrespective of elicitation method. However, elicited *levels* vary as a function of question format. Reporting full distributions yields fairly accurate responses. In contrast, directly asking for means positively biases responses. Further, volatilities from confidence intervals are not dependable. The evidence demonstrates complexities in survey design and helps researchers understand how to design and interpret surveys for meaningful results.

*Hartzmark is with Boston College Carroll School of Management and the NBER, Sussman is with the Chicago Booth School of Business. We would like to thank Niels Gormsen, Beidi Hu, Michael Ungeheuer, as well as seminar participants at Aalto, Boston College, Boston University, Consumer Financial Protection Bureau, London Business School, UMass Amherst, the European Finance Association Conference and USC Alumni Conference for helpful comments. We are grateful to Jieyi Chen and Fiona Yang for outstanding research assistance.

"What do you expect the return of the US stock market to be over the next 12 months?" (Giglio et al. (2021)), or something similar,[1] is a ubiquitous survey question in finance. This is sometimes accompanied by a confidence interval question, such as one asking respondents to report a value such that "there is a 1-in-10 chance the actual return [of the S&P 500 over the next 12 months] will be less/greater than" that value (e.g. Ben-David et al. (2013)). Responses to these questions are used to proxy for expected means and volatilities of stock returns, which are central to understanding asset prices. While there has been a dramatic increase in the use of survey-based expectations data in finance, there is minimal evidence about whether or not this is the most effective question format to elicit means and volatilities of stock returns. Open questions remain about how best to elicit such expectations and even whether these expectations can be meaningfully reported at all.

In this paper, we critically examine the benefits and drawbacks of eliciting expectations of returns by asking directly about means and confidence bounds versus an alternative method based on querying full distributions. This method, which has become increasingly popular when asking about macroeconomic variables, provides multiple bins with specific ranges of values and then asks respondents to report beliefs corresponding to complete distributions.[2] To examine the influence of question format on reported beliefs, we adapt techniques from psychology that allow us to identify a ground truth from which we can test the accuracy of responses (e.g., Goldstein and Rothschild (2014)). We ask questions about the ground truth we provide, rather than about expectations of real world future outcomes. This approach allows us to examine not only the difference in responses as a function of question type but also the accuracy of these responses in terms of levels. While some questions in finance relate to variation in expectations over time, for many key questions (e.g. the equity premium puzzle) and most models of asset prices it is the level of expectations that is of interest.

Examining nearly 7,000 respondents - including lay, high asset, and mathematically sophisticated populations - we document a number of material differences induced by the choice of question format. Our first result is that question format is not particularly important for identifying differences in means and volatilities. Across all question types we examine, the reported ordering of means and volatilities compared within question type matches that of the true distribution. This suggests that survey-based expectations data is an informative input for a number of economic applications.

[1] e.g. the Graham and Harvey CFO survey asks for the value "I expect the return to be" for the S&P 500 in 12 months. Shiller's survey asks "How much of a change in percentage terms do you expect in 1 year" for the Dow."

[2] See, for example, the Survey of Professional Forecasters when asking about inflation, GDP growth, or unemployment rates, or the Survey of Consumer Expectations when asking about inflation, earnings growth, or home prices. See Manski (2018) for an overview.

With that said, we find systematic differences in accuracy across elicitation methods in identifying the levels of these beliefs. Directly asking for mean expectations and confidence intervals does not yield particularly accurate measures of levels of beliefs. Direct elicitation of means suffers from a large positive bias. Elicited values for confidence intervals are largely independent of the width that is asked about (e.g., an 80% vs. 90% CI range). This means the question construction influences the implied volatility making it difficult to interpret meaningfully.

If participants are instead asked to build complete distributions by filling in the number of outcomes within each range, reports are fairly accurate. The reported levels of means and standard deviations tend to be close to those in the true underlying data. A shortcoming of this method is that the choice of where to center the bins influences elicited means. However, this method induces a smaller error than directly asking for means, irrespective of the center bin within the range we examine. These results suggest that improving survey methodology enables more accurate reporting of beliefs while greater attention to the survey methods used for elicitation can help researchers better interpret reported beliefs.

A sizeable literature in psychology and behavioral economics has found that question wording and context can affect responses in various settings in which individuals form and report beliefs. Thus, a core goal of the researcher designing a survey question is to determine how to ask questions as clearly and effectively as possible to maximize accuracy in responses and minimize satisficing (i.e., just good enough responding) behavior on the participant's part.[3]

Understanding the accuracy of survey responses to questions eliciting return expectations is complicated because the ground truth (i.e., a participant's *actual* beliefs about the future) is unknown. An individual can express any expectation, and it is impossible to evaluate the extent to which they accurately express their beliefs versus respond to an artifact of the question. We may learn that people report higher expectations of a given metric than materializes in reality, but this could be because of inaccuracy in either the beliefs themselves or in the ability of the survey to capture those beliefs. If different question types lead to different responses, then it is reasonable to conclude that question type matters, but one cannot evaluate which approach provides responses closer to the true expectations.

[3]In response to surveys, participants will typically respond reliably to clearly stated questions, which have responses that participants know and are willing to share, and require little cognitive effort to answer. For example, if you ask people a simple question they know the answer to (e.g., "Did you trade stocks yesterday?"), they will likely answer it correctly, irrespective of context. But, as questions become more difficult (e.g., "How many trades did you make in your retirement account over the last five years?"), people will start looking for easier solutions to the cognitively demanding and overall uncertain task (Krosnick et al., 1996). Participants will trade off accuracy with cognitive convenience and may leverage cues from the survey context to help them provide a reasonable enough response to the survey question.

To study which question types yield more meaningful responses, researchers must have a benchmark against which to judge those responses. We induce distributions to provide a ground truth by exposing participants to a series of numbers and then asking about the distribution they observed. If participants are able to respond meaningfully to a question about a known distribution, then it is reasonable to expect that the same question format would elicit similarly meaningful responses for future expectations when an appropriate benchmark is unknown to the econometrician.

The distributions we examine are stock returns or randomly generated numbers with similar distributions. These are volatile distributions of percentage changes with a small positive mean, similar to the relevant distribution in other important economic contexts (e.g., economic growth, corporate earnings growth, inflation). We find nearly identical results when the distributions are generically described as numbers, suggesting that our results are not specific to stock returns and extend to other similar distributions.

In our first treatment, we display 100 years of returns of either the S&P 500 or the Japanese Nikkei. Returns are shown one at a time in a random order. Participants are told they are seeing past stock market returns, but they are not told from which market. After viewing these returns, we ask participants about them using different techniques. A similar paradigm has been used in other settings in psychology using different distributions (Goldstein and Rothschild, 2014). In those contexts, participants were able to meaningfully digest and report back the information under certain conditions.

We focus on two main question formats. We refer to questions asking for a point estimate or range as the direct method. For means, this involves asking about average returns or expectations of returns. For uncertainty, this is based on confidence intervals, which researchers convert into estimates of variances (e.g. Ben-David et al. (2013); Boutros et al. (2020)). The alternative we explore is having participants build a distribution. We provide bins with ranges and ask how many observations fell into each bin. We refer to this approach as the distribution method. Prior research shows that eliciting full distributions can mitigate biases related to data interpretation by engaging respondents in a more comprehensive process of considering the entire distribution, rather than focusing on a specific statistic.[4]

First, we examine whether participants can meaningfully identify differences in distributions by varying whether the returns presented are from the US or Japanese stock market. In our sample, the

[4]This method has been shown to be effective at eliciting more accurate responses by encouraging participants to account for all possible values and ensuring that their responses sum to 100% (Tversky and Koehler, 1994; Haran et al., 2010). This can reduce the tendency to exaggerate certain subjective probabilities while overlooking extreme values.

US had a lower mean (7.97% vs. 10.21%) and a lower volatility (standard deviation of 18 vs. 27). Irrespective of elicitation method, participants correctly reported this pattern with a t-statistic above 5 for each method. Pooling all estimates, participants viewing the Japanese distribution report a higher mean, with a t-statistic of 9.81 and higher volatility, with a t-statistic of 12.45. Given we do not reveal what market the returns are taken from, personal knowledge of the differences between the US and Japanese markets cannot easily explain the differences we observe.

This finding has two major implications. First, it shows that participants can meaningfully express differences in the first and second moment of percentage change distributions of the type commonly of interest to financial economists.[5] Second, it shows that participants can meaningfully process the information we display in the format we display it and report this information back in a survey. This validates the extension of the experimental paradigm we use in the return space that we explore.

While we find that participants can successfully report ordering using both question types, participants are more accurate at responding to the distribution method than to direct prompts for both means and volatilities.[6] While the average return in the US was 7.97%, when means are directly elicited, the average responses are more than 70% higher than this at 13.48%. For Japan, the average return was 10.21%, but the directly elicited average was nearly 100% higher, at 20.06%.

When asked about distributions using bins centered at the historical mean, the responses are 8.11% for the US and 10.48% for Japan, which are not significantly different from the truth. When asked using bins that are centered at zero, the responses are lower, with values of 5.18% for the US and 7.21% for Japan. The distribution method is not a panacea, and the choice of bin construction influences responses by anchoring estimates towards the choice of scale midpoint and inducing error through the discrete approximation. With that said, the direct elicitation forecast errors are each greater than 70%, but using bins centered at zero, significantly below the true mean, leads to an error around 30%.

The distribution method also leads to more meaningful responses than the direct method does for volatility. The direct elicitation method elicits volatility using confidence intervals (e.g., Ben-David et

[5]This pattern is consistent with earlier evidence of the validity of survey measures such as positive correlations between statistical estimates and participant reports of anticipated inflation (e.g., Armantier et al. (2013); Ang et al. (2007); Hafer and Hein (1985); Bruine de Bruin and Fischhoff (2017); D'Acunto and Weber (2024); Weber et al. (2022)) and correlations between reported beliefs and trading decisions (e.g., Giglio et al. (2021); Liu et al. (2022)).

[6]This is consistent with prior research reporting participants who anchor on arbitrary values, but subsequently provide coherently ranked preferences given different levels as starting points (Ariely et al., 2003). For example, the minimum amount participants report they are willing to accept as payment to listen to an annoying sound increases as the length of time they will need to listen increases. However, the overall level of payment is higher after participants are anchored on an arbitrary higher (vs. lower) value, leaving the "objective" willingness to accept unclear.

al. (2013)). However, individuals have been found to be quite poor at responding to such prompts as their answers do not sufficiently vary in response to different confidence intervals (Teigen and JØrgensen, 2005). If confidence intervals are valid, the variance implied by asking for a 90% confidence interval should be the same as that implied by asking for a 60% or a 98% interval.

This is not what we find. Using US data, when asking for a 98% confidence interval estimate, the standard deviation of returns implied by participants' responses is 13, but asking for a 60% confidence interval estimate yields an implied standard deviation estimate of 23. These results suggest that participants do not respond meaningfully to such prompts.

When using the distribution method, participants are more accurate. For the US, the standard deviation estimates vary from 19.78 to 19.84, close to the true value of 18.35. For Japan, estimates are 24.00 and 23.02, below the true value of 27.24. Outliers in the underlying Japanese data result in the correct response (i.e., the response implied by allocating each number to the correct bin) being further from the true distribution. These values are close to what would be implied if each bin were filled in correctly.

The distribution method consistently yields more accurate results, but this could be somewhat mechanical because the bins are calibrated on the data that participants report on. Participants could infer the experimenter's intent through the bin construction, or they could anchor their responses on where the bins are set.[7] In the field, bins calibrated on historical data are unlikely to exactly match the distribution of expectations, but in our setting, they can be one and the same.[8]

To examine the influence of the specific distribution presented to participants as well as the bin construction on responses, we run a new treatment with the same bins, but a range of distributions. Each participant is randomly assigned a mean, M, from 3 to 13 and a standard deviation, S, from 13 to 23. They then view 100 draws from that distribution.

All of our stylized facts replicate in this treatment. Within question ordering is correct, and the estimates from the distribution method are notably more accurate and more precise than those from the direct method. With that said, we find that there is a muted response to differences in means using the distribution method, which may be in part due to anchoring on where the bins are centered.

Different surveys are run on different populations and the influence of question type may vary

[7] Although bin construction has been shown to influence responses (e.g., Becker et al. (2023)), participants can identify and report differences in a range of distribution types (e.g., uniform, bimodal, normal) presented to them Goldstein and Rothschild (2014) suggesting they are responsive to the information presented.

[8] Variation in the bins provided in our studies helps address this point, as participants respond to bins constructed based on historical returns of either the US or Japanese returns irrespective of the distribution they view, as well as to bins centered at 0.

with the population. The majority of our analysis examines a sample of participants from Prolific Academic (similar to MTurk). Such a population is likely comparable to surveys run on the general population.[9] Of course, some surveys focus on populations with more financial experience and mathematical sophistication.[10] We re-run our surveys on populations with significant financial assets or degrees in mathematical subjects. We do not find that having financial assets meaningfully changes our results, but the mathematically sophisticated sample reports means with lower forecast errors for both question types.

Our results suggest that for many survey populations, eliciting distributions will likely yield more meaningful results than direct elicitation methods. However, the distribution method is not without its own shortcomings, as it both anchors estimates towards the center and induces noise based on the discrete approximation required to draw meaning from participant responses. The results from our mathematically sophisticated sample suggest that statistically sophisticated professionals who are well calibrated to a relevant distribution (e.g. quant financial professionals, economic forecasters) may achieve superior results through direct elicitation (though unfortunately we do not have access to such a sample). Outside of such settings though, which represents the majority of surveys conducted, the distribution method should likely be preferred.

Researchers typically want to measure expectations of future outcomes, but this paper does not do so. Thus, further assumptions are needed for the results to be applicable to eliciting expectations. We demonstrate how participants respond to a prompt about a distribution they have in mind. A central assumption of economic theory is that people have something like a distribution in mind when making decisions. If this assumption holds, then our results should be relevant to asking about expectations.[11] If people do not have anything approximating a distribution in mind when making decisions, our results might be less relevant to this context, but more importantly this would present a significant challenge for almost every asset pricing theory.

Our paper contributes to the literature in finance on surveys by providing context to view the evidence based on prior question types. Greenwood and Shleifer (2014) study six different surveys which elicit returns and these surveys have been used in a variety of other papers, for example Vissing-Jorgensen (2003), and Malmendier and Nagel (2011) among others. See Adam and Nagel (2023) for a

[9] Examples include the Michigan Survey of Consumers, Gallup, or the Shiller survey of individual investors.

[10] Examples include the Graham and Harvey CFO survey, the Vanguard survey in Giglio et al. (2021), the professional investor Shiller survey, or various commercial surveys of professional forecasters or Fed employees.

[11] Specifically, responding to survey questions is a function of querying and integrating information in memory in response to the specific question prompt Krosnick et al. (1996); Tourangeau and Rasinski (1988).

recent review. Recent papers often craft their own questions using variants of direct elicitation (e.g. Giglio et al. (2021), Merkle and Ungeheuer (2021), Liu et al. (2022)). The Graham and Harvey CFO survey asks a confidence interval question about S&P 500 returns, which is used to measure perceived volatility (e.g. Ben-David et al. (2013)).[12]

A related literature studies expectations of macroeconomic variables, with a particular focus on inflation. For example, in a series of papers, D'Acunto et al. (2021b), D'Acunto et al. (2021a), D'acunto et al. (2023), amongst others, explore direct elicitation methods from a variety of surveys to elicit inflation expectations (see Weber et al. (2022); D'Acunto et al. (2023) for reviews). The literature has included an investigation of the validity of eliciting expectations, finding that expectations tend to be correlated with relevant outcomes. Armantier et al. (2013); Ang et al. (2007); Hafer and Hein (1985); Bruine de Bruin and Fischhoff (2017); D'Acunto and Weber (2024); Weber et al. (2022). Manski (2004) and Manski (2018) make a case for eliciting probabilistic belief distributions and provide overviews of the history of belief elicitation in this space. Research in macroeconomics such as Becker et al. (2023) and Boctor et al. (2024) have begun to identify how differences in elicitation methods can influence responses. Interestingly, while these distributions are slightly different than what we study, existing literature (primarily examining inflation expectations) has found a number of similar stylized facts such as that responses are meaningful at the level of the rank order, are often higher than considered benchmarks, and are influenced by the question type and construction of bins (see, e.g., Becker et al. (2023); Weber et al. (2022); D'Acunto and Weber (2024); Armantier et al. (2013)). The methods explored in our paper allow us to replicate many findings from these prior papers.

Importantly, we extend this literature in several ways. We are able to provide a ground truth against which to compare reported beliefs. This provides a high-quality and nuanced point of comparison unavailable to the prior literature. This allows a clean identification of the accuracy of the level of beliefs which cannot be identified using the methods from these prior papers.[13] Our method also allows us to extend our comparisons beyond means to the full distribution of beliefs. In addition, we generalize findings to stock market outcomes, which tend to have higher means and larger volatility in comparison to most macroeconomic variables. Finally, our methodology requires recall of a prior distribution, which better isolates beliefs versus preferences.[14]

[12] Alternative question formats have focused on variables related to market performance that are not returns. For example, asking for the chance the stock market will increase (e.g. Michigan Survey of Consumers), or asking about levels of future prices Glaser et al. (2019). Our paper does not directly address these formats.

[13] E.g., prior work has shown that outcomes are correlated with expectations or responses are consistent within question type, but it does not not capture deviations from the levels of actual expectations.

[14] See, e.g., Manski (2018) for a discussion of this issue.

In addition, we contribute to literature in psychology and related fields by extending prior methodology around belief distributions to volatile distributions with a small positive mean. Research in this context has historically used direct elicitation of confidence intervals rather than eliciting more complete distributions.[15] More recent literature has begun to explore the importance of eliciting full belief distributions, often to reduce overconfidence (Haran et al. (2010); Moore and Dev (2020)).

Within the literature using this distributional approach, researchers often sidestep difficult questions such as how to create optimal bins by examining situations where the range of possible values is known and bounded (e.g. integers 1 through 10 or percent terms bounded at 0 and 100) (e.g., André et al. (2022); Goldstein and Rothschild (2014); Moore et al. (2017)). In other cases, this literature will select bins without substantial discussion of why (e.g., Hu and Simmons (2023); Prims and Moore (2017); Reinholtz et al. (2021)).[16] We contribute to this literature by providing guidance on a reasonable approach to constructing bins in situations where it is not straightforward and where responses are likely to cluster around low positive values with the potential to be either positive or negative.

While the use of survey data is still relatively nascent in finance, our results suggest it is a valid input for a number of applications. Many in finance have responded to the fact that question type influences answers as a reason that survey evidence can be discarded altogether as meaningless. By the same logic, all empirical results can be ignored because different econometric specifications lead to different conclusions or different modelling assumptions lead to different structural estimates. Clearly, all of these approaches can yield useful insights; the key is understanding and developing the methodology. Adapting the large and rich existing literature on survey design to settings tailored to economics questions is an important task, and developing research in this area will greatly enhance the ability of the field to examine questions about beliefs.

In the remainder of the paper, Section I provides an overview of the experimental design. It explains how we communicate the ground truth as a point of comparison and the two methods we use to elicit beliefs from participants. Section II reports the baseline results from a treatment using the past 100 years of historic stock returns. It illustrates that participants respond meaningfully to both elicitation methods, but that their responses tend to be higher in response to direct elictation and are

[15] See Alpert and Raiffa (1982) for an early example and Moore et al. (2015) for a review of common elicitation methods. See Soll and Klayman (2004), Speirs-Bridge et al. (2010), and Mandel et al. (2020) for comparisons of different methods of directly eliciting confidence intervals.

[16] Notably, Haran et al. (2010) varies the range and number of bins in a task estimating Washington DC temperature estimates and does not find meaningful differences in participant estimates as a function of this variation, although it does find differences in implied confidence intervals. Crosetto and de Haan (2023) test a more complex *click-and-drag* method that allows the participants to determine bins and compares this against other methods of eliciting belief distributions when mapping from one visually displayed distribution to their own.

both lower and more accurate in response to the distribution method. Section III reports on a treatment which provides participants with a variety of distributions with varying means and volatilities and finds similar results. Section IV probes populations with high assets and mathematical expertise and again finds similar patterns although those with mathematical expertise are slightly more accurate. Section V explores results across methods further, Section VI discusses key considerations for future researchers trying to elicit and interpret beliefs and Section VII concludes.

I. Experimental Design

The core experimental design entails providing people with 100 numbers, one at a time. This set of numbers serves as a distribution, which they can retrieve from memory and report back in response to questions eliciting this distribution. We adapt this method from prior research in psychology (e.g., Goldstein and Rothschild (2014)) which has used a similar approach to identify beliefs about distributions in other contexts, finding that people are quite accurate at reporting distributions based on this approach. This ability to report central characteristics of the distributions is consistent with people's overall ability to draw summary information from larger sets of information, even when given limited processing time (Chong and Treisman, 2003).

We selected this approach because it provides a way to implant a distribution which is constant across participants and accessible to the researcher. However, it is different from asking people to generate beliefs about future events. For example, when generating predictions, people may have more uncertainty and gather information from a broad range of inputs to determine their response, whereas in the current setting, participants are primarily drawing from recent memory. This raises an important question about whether eliciting one type of judgment can serve as a proxy for the other.

To gain insight into this question, we draw from seminal literature on survey methods which identifies four stages of answering questions: (1) interpreting the meaning of each question, (2) searching memories extensively for all information relevant to answering the question, (3) integrating information carefully into summary judgments, and (4) responding in ways that convey those judgments and meanings as clearly and efficiently as possible Krosnick et al. (1996); Tourangeau and Rasinski (1988). While the memory search and information integration process may be more extensive when recruiting expectations versus direct recall, these four core steps remain the same. In the current context, we provide participants with input which serves as the relevant accessible information (steps 2 and 3). The primary goal of the experiment is to examine how participants rely on the question type

in answering their questions (steps 1 and 4). While possible, we do not have reason to believe that differences in how people are searching their memories for information in these intermediate steps would interact with elicitation methods. Consequently, we would not expect variation between the current task and the process of identifying beliefs about future expectations to differentially influence participants' ability to respond to question type as a function of the recall versus expectation task.

A. Experiment

Our experiment provides participants with a distribution of returns so we can assess their ability to report on that known distribution when asked about properties of the numbers displayed. The complete text and screenshots of the experiment are included in the Internet Appendix. Upon starting the experiments, participants see the following:

> *In this survey, we are interested in understanding how people recall and summarize information presented to them.*
>
> *On the following pages you will see 100 numbers. Each number represents the* ***stock market return*** *from one of the last 100 years. The returns will be presented in random order and cover all 100 years.*

After the instructions, participants had to pass a reading check to ensure they were reading carefully and engaging with the study (Bergman et al., 2020).[17] After passing the reading check, participants viewed a page with one number which was randomly chosen from the prior 100 years, without replacement. After viewing the number for half of a second, a box appeared, and participants clicked it to see the next number. The next number was again displayed for half a second before the box appeared. This occurred 100 times until participants had seen all values from the prior 100 years. The sequence was randomized for each participant.

While we did not tell participants which stock market data we used, we presented two different series. We used the S&P 500 to proxy for the US stock market and the FTSE index to proxy for Japan. In the last 100 years, Japan had a mean return of 10.21%, higher than the 7.97% for the US. Japan's returns were also more volatile with a standard deviation of 27.24, higher than the US's of 18.35.

[17] The question asked "In this survey, you will be shown information about _____ over the past 100 years:" and provided six multiple choice options for participants, including the correct answer, "stock market returns". If participants answered incorrectly, they were given one opportunity to re-read the instructions and try again. If they responded incorrectly a second time, they were directed to the end of the survey.

We note that this is a recall task and that many patterns have been documented in how people recall sequential information. For example, they recall recent, initial, and extreme information (Deese and Kaufman, 1957; Murre and Dros, 2015; Kahneman et al., 1993). This could be a concern in our setting if participants saw the same sequence. Importantly, randomizing the order for each participant should remove any sequence effects from the sample averages. In addition, it is unclear how any recall biases would differentially influence responses as a function of the elicitation method.

The experiment used a 2 (Data source: US vs. Japan) X 2 (Response method: confidence intervals vs. distributions) X 2 (Order: direct elicitation first vs. direct elicitation second) design where each factor was randomly assigned across participants. First, the data source varied whether it displayed the 100 year history of returns from the S&P 500 (US) or FTSE (Japan).[18] Second, participants were randomly assigned questions asking about their beliefs in the form of either confidence intervals or distributions. Third, the order varied such that participants were randomized between seeing the direct elicitation question first followed by the confidence interval or distribution question, or seeing the confidence interval or distribution question first, followed by the direct elicitation question.

In addition, participants in the confidence interval (CI) conditions were further randomized into one of four conditions where they were asked about a 60, 80, 90, or 98% CI range. Participants in the distribution elicitation were further randomized into one of four sets of bins. Specifically, the construction of the bins varied in whether the bins were calibrated using US or Japanese data and whether the bins were mean-centered or zero-centered. In all cases, bin width was given by 1/2 of a standard deviation which is 9 for the US and 15 for Japan. This was true regardless of what source of data they saw (e.g. a participant who observed US data could be randomized into responding based on US or Japanese bins). Participants in the distribution treatments were also randomly assigned to enter their responses using text boxes or a graphical interface.

Every participant was asked to report a point estimate of the average value of the returns (i.e., direct elicitation). Specifically, they were asked:

> *Please consider all the stock market returns that were just displayed to you. What do you think was the average value of these returns?*

There was a box where participants could enter this number.[19]

[18] Some early treatments included only US data and bins. See Table 1 Panel D for a breakdown.

[19] To clarify how to state their beliefs, we also wrote: "To indicate the overall market is increasing (a positive return), insert a positive number. To indicate that the overall market is decreasing (a negative return), please use a negative (-) sign in front of the number." This was also included for the confidence interval questions

Participants in the confidence interval conditions were asked:

> *Please consider all the stock market returns that were just displayed to you. Imagine that someone randomly selects one value of returns from the last 100 years. Please list two values so that there is an X% chance the return will be between these two numbers.*

X was randomized across the values of 60, 80, 90, and 98. Participants were prompted to enter two values, indicating upper and lower bounds.

Different versions of the distribution treatment have been used in the literature and we wished to test whether these influence responses. Some participants in the distribution treatment saw text boxes with ranges of values next to them (e.g. 4% to 12%). Participants filled in these bins with the restriction that the sum of values entered equal 100. A total bin displayed the sum at the bottom.

More complicated interactive versions of the distribution method have been employed such as using a graphical interface that visualizes the distribution being formed (Goldstein and Rothschild, 2014; Sharpe et al., 2000).[20] To examine such a procedure, the remainder of participants saw a nearly identical task, but used a graphical display to allocate balls using the distribution builder plugin developed in Andre (2016). See the appendix for an example.

We find similar results with and without the graphical interface (see Appendix for details). Consequently, we present results for both of the methods together. Further, we transitioned later experiments to asking questions using a simpler and more basic numeric fill-in approach in some variations. Our results suggest that simple survey formats using bins can give similar responses to those using more complicated visual methods.

B. Calculation of Mean and Variance

For most of our variables of interest, further assumptions are necessary to calculate means and variances from the responses. The exception to this is the direct elicitation of mean values. Even in this case, it is unclear how best to interpret outliers, and we thus choose to winsorize the values at 1%. To calculate standard deviations from the confidence interval prompts, we follow Ben-David et al. (2013) and use the Davidson and Cooper (1976) method. Namely, we take the range of upper bound minus the lower bound and divide it by the number of standard deviations in the confidence interval.

For the distribution method, binned responses need to be converted into means and standard deviations. To do so, we impute a value for each bin equal to the midpoint of that bin (e.g. the -23%

[20] see Crosetto and de Haan (2023) for a comparison of interfaces for eliciting belief distributions.

to -15% bin represents responses of -19%). The extreme bins (more or less than X%) are open-ended. For these bins we take the values used to calibrate the bins (e.g., for the US, a mean of 8 and SD of 18), assume normality, and calculate the average of this distribution above the cutoff (e.g., responses in the less than -23% bin represent -30.77%). Using these values we take the 100 observations implied by the responses and calculate the mean and standard deviation of these values.

C. Participants

For the main sample in our experiment, we recruited nearly 6,000 participants from Prolific Academic, an online participant population similar to Amazon's Mechanical Turk. This set of participants collectively completes several treatments including the US only, US and Japan, different numbers of bins, numbers only, and random mean and standard deviation variations, with each participant taking part in only one study. For simplicity, we report participant characteristics collapsed across these variations here. See Table 1 Panel A. Results from this population are likely to be similar to most other surveys we are aware of which do not target populations with specific financial knowledge or skills.

Table 1 Panel A shows summary statistics for our sample population. The median participant was 37, with an annual income of about $65,000. About half of our sample was male and about 70% of our sample had at least a college degree.

Of course certain surveys target populations with different attributes, such as having significant financial wealth or being a professional with greater knowledge and sophistication. While all of our experiments are recall experiments that neither require nor would be expected to benefit from any knowledge of finance, it could be that people with more experience in financial markets are better calibrated to respond to questions about them. Thus we also examine a sample of about 500 participants with significant financial wealth of at least $100,000 (and a median portfolio value of $375,000). It could also be that certain market participants are better at answering these types of questions not because of their superior knowledge of financial markets, but because of their grasp of statistics. Thus we also recruited a sample of about 500 participants who are likely more advanced in such knowledge with degrees in Statistics, Engineering, Math or Physics. These specialized samples were recruited through the Luth Research platform.

Table 1 Panel D shows a participant count across all of the variations in our treatments and sample. In total our results represent 6,714 participants. Our experiments were conducted in 2023 and 2024.

II. Baseline Results

A. Detecting Differences: US vs Japan

While interest in survey evidence has increased, there is still a high level of skepticism as to whether such methods can meaningfully provide information. Thus, before turning to more nuanced questions on identifying levels, we begin by asking a more basic question: can survey questions meaningfully identify *differences* in means and standard deviations? In response, we explore the difference in the estimated means and standard deviations between the US and Japan. If respondents describe the US as having a lower mean and standard deviation, this suggests that respondents can meaningfully interpret and report information, at least related to the ordering of means and volatilities.

Further, this test serves as a validation exercise of the paradigm we use. One possibility is that the information we display is simply not digested by participants. This could occur either because participants cannot process the values in this presentation format or because they are not paying attention. If we find that participants are able to identify distributional differences, this demonstrates that participants are meaningfully processing and responding to our experimental paradigm.

We begin by studying mean returns in Table 2 Panel A. In column 1, we regress the mean estimate on a dummy variable equal to 1 if the data was from Japan and a constant. Column 1 shows that respondents viewed Japan as having returns 4.6% higher than the US. This pattern is statistically strong with a t-statistic of 9.81. Next, we include treatment fixed effects, which account for combinations of question type and its sub-components (e.g., bin width and centering, or confidence interval range). Results are materially similar. Next, we split results into the direct or the distribution method. Both are significantly positive numbers, though the difference using the direct method of 5.5% is more than twice the true level of 2.2%, while the distribution method successfully provides an estimate of 2.2%.

Table 2 Panel B repeats the exercise, but explores estimates of the standard deviation. Examining the first two columns, we again see large positive coefficients of around 5 with t-statistics above 10. Across all confidence interval questions, there is a difference of 7.3, which is closer to the true difference of 8.9 than the distribution-based estimate of 3.7.

These results provide evidence that participants are able to identify differences in means and volatilities using survey questions. It also demonstrates that our method of inducing a distribution is effective at accomplishing its basic goals. The results hold for each question type. This suggests that prior research that focuses on changes and the ordering of beliefs (e.g. becoming more or less

optimistic) likely reflect the true ordering irrespective of the question type used.

B. Level of Means

A key input to most asset pricing models is the level of expectations of means and volatilities of stock returns. Thus better understanding of such levels is key to resolving a number of important asset pricing debates. In this section, we examine our results in terms of how well participants were able to identify levels of moments from such a distribution.

We begin by examining results from participants who viewed US data in Table 3 Panel A. The true mean of the returns viewed was 7.97%. Using direct elicitation, the average reported estimates were too high, at 13.53%. This is roughly 70% higher than the correct estimate.

When using the distribution method, however, results are closer to the true answer. When the distribution is centered at the historic mean, the mean value is 8.11%, which is not statistically different from the true value. When the distribution is centered at zero, the results are lower, with a mean response of 5.18%. While lower, the 2.79% difference from the true value obtained through the distribution method is roughly half the distance from the true value compared to the 5.51% difference obtained through the direct elicitation method.

In Table 3 Panel B we explore the results for participants who observed the Japanese returns, which had a true value of 10.21%. Direct elicitation yields responses which are roughly double the true value, at 20.06%. The distribution method yields responses which are much closer to the truth. Using bins centered at zero, the estimate is slightly too low at 7.21%. Using bins centered at the mean, the estimate is 10.48%, statistically indistinguishable from the true value.

The results for the US and Japan exhibit a similar pattern. When directly eliciting means, the responses are too high and much further from the truth than when using the distribution method. This finding of overestimation is consistent with prior findings including Weber et al. (2022)'s reports of an upward bias in reports of inflation expectations elicited from surveys and Glaser et al. (2019)'s result that asking participants directly about returns yielded responses that were not only high but also higher than asking participants about prices. When eliciting means through distributions, estimates are closer to the true value. [21] Centering the distribution bins at zero, significantly below the true mean, draws the mean estimate down slightly. Centering at the true value, the results are not statistically different from the truth.

[21] This pattern is consistent with Jung et al. (2020)'s (study 10) result that asking participants directly about their beliefs of others' valuations led to overestimates relative to asking about full distributions.

Participants who built distributions also answered direct elicitation questions. Both questions were based on the exact same sequence of displayed returns and thus should reflect that participant's perception of that sequence. If our results are really capturing the influence of question type, we should find similar within person patterns.

Table 4 shows the average of the within person difference between the direct elicitation mean and the mean implied by the distribution response. We find across treatments that the same person responding about the same sequence of data reports a mean about 7.6% higher based on the direct elicitation. This is consistent with our results reflecting artifacts of the question type, not differences in opinions about what the data actually was.

An interesting question is whether the order of questioning matters. Given our randomization, some participants saw a direct elicitation question followed by a distribution question and some saw the opposite. Table 5 shows responses to a given question type when the direct elicitation question occurred first (Panel A) or second (Panel B). Results are generally similar suggesting that responses are not greatly influenced based on whether they are the first question or a followup question. Thus in the paper we report both estimates together in all the other analysis.

C. Level of Volatility

While mean estimates are clearly important to finance and economics, so are measures of risk. This section explores methods of estimating standard deviations. It finds that the distribution method allows participants to express volatility more accurately.

We begin by exploring the direct elicitation of confidence intervals. Combined with a distributional assumption, typically normality, any confidence interval can be converted into a standard deviation. The key assumption behind this calculation is that participants can meaningfully respond to a prompt about a confidence interval.

To test whether this is the case, Figure 1 shows a variety of ranges from elicited confidence intervals. For each person we take the upper value minus the lower. We average this number and graph it centered at zero. Panel A shows results from the US data and Panel B shows results from Japan. The navy line represents the elicited range while the maroon line shows the true range in the data.

Both figures show that participants are not nearly sensitive enough to variation in the specific confidence interval that is asked. Examining the maroon line, one can see that the range of intervals asked about should have a large impact on the answer given to reflect the same underlying beliefs

about uncertainty. Examining the blue line, it is clear that this is not the case. While there is some variation across ranges, it is not nearly enough to reflect the question being posed.

Table 6 shows how this pattern relates to the implied volatility. Given the lack of responsiveness to the question, the implied volatility for the 98% interval is far lower than that for the 60% interval. Using US data, participants report a standard deviation of 24 using a 60% confidence interval and of 13 using a 98% confidence interval. For Japan, these numbers are 30 and 21 respectively. This is driven not by differences of opinion about volatility, but by the distributional assumption necessary to convert the estimate to a standard deviation. Given the distributions presented are the same and only the specific confidence intervals asked about are randomized across subjects, it is worrying that we observe meaningful variation in the standard deviation estimates implied by participant responses. This indicates that participants are not responding meaningfully to the specific question being asked. Thus, it does not make sense to apply the distributional assumptions to convert the response to a standard deviation, taking as a given that participants are considering the indicated interval. At the same time, there is no clear alternative for interpreting the responses participants provide as we do not know what size interval participants have in mind when answering the question.

We next examine the implied standard deviation from the distributional method. The US estimate is 19.78 when centered at the historic mean and 19.84 when centered at 0. While these are slightly higher (and statistically significantly different) than the true value of 18.35, they are substantively fairly close to correctly estimating the actual value. When examining Japan, participants report estimates of 24.00 and 23.02 which are lower than the true value of 27. As discussed in Section V, the Japanese data is somewhat irregular, which results in a significant difference in the implied standard deviation and the sample actual even when the value entered into the distribution bins exactly match the true underlying distribution. While such an irregular distribution can occur using realized returns, this type of issue is unlikely to be relevant when examining actual beliefs. Compared to the correct value of about 23, the values are slightly higher to a similar degree as what we see for the US data.

These results suggest that standard deviations from confidence interval estimates should be avoided. Participants' responses are not sufficiently sensitive to the question prompts when stating their confidence intervals directly, which brings the validity of such estimates into question. The implied standard deviation of the distribution method more closely reflects the data.

D. Specific to stocks?

Our baseline treatment accurately describes the numbers as past stock returns. Although it is unclear how or why participants would use that information, it is possible that this influenced the results.[22] One potential influence of describing values in terms of stock market returns could be to anchor participants on a prior distribution. For example, when thinking about stock returns perhaps people have in mind a distribution with more mass in the middle, such as a normal distribution.

We repeat our treatment using the US data, where we remove any mention of the stock market and returns. The same numbers are shown, but they are labeled simply as numbers which participants are asked to recall.

If the concern was related to inducing a prior distribution, it is likely that in this version the prior distribution would be different. For example, if I state I am randomly drawing numbers, it seems just as likely that someone might happen to have a uniform distribution in mind as a normal one. This is reflected in the fact that in prior similar studies where observations are labeled generically as numbers, participants were able to identify a number of different distribution types (Goldstein and Rothschild (2014)). Regardless of the specific distribution, any prior distribution would likely be different in the treatment with a stock label compared to the treatment without.

Table 7 shows a materially similar pattern in these treatments. The direct elicitation method is too high at 13.74%, similar to the 13.36% we found when describing the numbers as stock returns. The distribution method estimate centered at zero is a bit too low at 3.98%, similar to what we found when describing it as stock returns. Finally, the distribution centered at the mean is quite accurate at 7.81%, similar to what we found describing it as stock returns. Standard deviation estimates are also similar. In the treatment describing the data as numbers, estimates are 18.85 and 19.58. This is quite close to the values of 19.35 and 19.4 from the version with stock returns.

Another possibility is that while this setting does not require financial knowledge, perhaps those with investing experience are simply more comfortable and better calibrated at responding to distributions of this type.[23] In the following section we repeat our analysis on a sample of investors with significant financial wealth, namely a portfolio of greater than $100,000 and the description of the

[22] While this could potentially influence levels, it is even harder to explain the differences between the US and Japan as the instructions did not identify the country from which the data was obtained.

[23] This experience could influence either beliefs or ability to respond to the survey prompts, similar to how estimates of expectations have been shown to vary with demographic factors in the context of inflation, such as in Bruine de Bruin et al. (2010); D'Acunto and Weber (2024); D'Acunto et al. (2021a); D'acunto et al. (2023).

numbers again presented as stock market returns.[24] The results suggest that this sample does not perform better and if anything performs slightly worse than our lay sample.

These findings suggest that the results we find reflect the survey methods and the distribution of the data we are examining. It is unlikely to reflect anything related to financial knowledge, personal experience in the stock market, or prior beliefs about the shape of stock return distributions. This suggests the results are likely to be applicable in any number of settings where the underlying distribution of beliefs is similar to what we study here.

III. Continuous Treatment

To this point, the paper has restricted itself to examining two different actual distributions of prior returns. While a sensible starting point (and how the academic literature tends to calibrate models), this method has limitations. The prior section was forced to calibrate bins based on the actual underlying data. In most settings, it is likely that the belief distribution and the calibrated bins will be somewhat different from each other. The prior method varies the match between the induced belief distribution and the bins by randomizing the underlying distribution (US vs. Japan) along with the distribution used for calibration (US vs. Japan), and by centering the bins either at the mean or at 0. However, the prior method is not flexible enough to explore the influence of bin construction systematically.

In addition, there are only a few possible distributions of past returns to examine. Such realized return distributions can have extreme realizations that are unlikely to be a part of ex-ante return expectations. Taking the Japanese data seriously as expectations suggests expectations of a 1% chance of returns of 105 or 118, but a 0% chance of returns from 106 to 117. While this makes sense for realizations, of course expectations would likely include smooth tails and generally smooth distributions.

To complement the prior study and circumvent these concerns, this section develops a new treatment based purely on simulated data.[25] The use of simulations means that we can randomize over a large number of potential distributions. While we randomize the distributions, we keep the bins the same so we can directly measure the influence of bins deviating from the underlying distribution.

Specifically, for each person we randomly assign them to view a series of numbers drawn from a normal distribution with a mean M and a standard deviation S, i.e. $N(M, S)$. We center our choices of M and S based on the US values of $M = 8$ and $S = 18$. We wish to explore a wide range around

[24] Participants were instructed: "On the following pages, you will see a series of 100 numbers, which we refer to as "stock market returns". These are randomly selected numbers chosen from a distribution that approximates (though is different from) the annual stock returns in the past 100 years of various large national stock markets."

[25] This study was preregistered on AsPredicted: #164247 (https://aspredicted.org/VHV_JKR).

these so we pick mean returns ranging from 3% to 13% and standard deviations ranging from 13 to 23. Thus for each participant, we randomly pick a mean M from the uniform distribution $U[3, 13]$ and a standard deviation S from $U[13, 23]$.[26] We then pick 100 numbers at random from $N(M, S)$ (rounded to the nearest integer) and display these 100 numbers.[27] For our analysis we examine elicited beliefs relative to the calculated mean and standard deviation of the 100 numbers subjects actually observe.

Outside of randomizing the distributions, the treatment is nearly identical to the treatment where we remove any mentions of stocks. The prior results provide evidence that results are similar with or without such a label. Thus, this approach avoids difficulties of trying to explain hypothetical stock market returns that are not related to actual markets while maintaining similar values for the numeric distributions. We do not include confidence interval questions since we expect that participants who were unresponsive to confidence interval width in the prior treatment would behave similarly here.

We use the US bins calibrated at the mean values. Thus randomizing M and S allows us to test for the influence of having bins that are not directly calibrated from the data that we are examining.

Before turning to the results, this treatment allows us to explore the inherent limitations of using coarse bins to estimate implied means and standard deviations and any mechanical bias that might entail. Specifically, for every integer combination of M and S in the ranges we explore, we took 10,000 observations from those distributions and allocated them into bins one through nine. We then calculated the implied mean and standard deviation from these answers using our standard methodology.

Figure 2 shows the results. Figure A shows that for means, while the bin method induces some noise relative to the true mean, this noise is relatively small and consistent across the range of means. The coarse bins introduce random error, but do not seem to induce a systematic bias.

Panel B shows that for standard deviations, distributions that are more volatile than that used to calibrate the bins will be biased towards understating the standard deviation. This can be seen by the green dots drifting below the orange line as the S rises above the S used to calibrate the bins. In practice this occurs because the implied value of the tail bin is based on the properties of the distribution used to calibrate the bins. If the actual distribution is much more volatile, the implied value of extreme bins is too small and the standard deviation is understated. While this induces a systematic bias, it is relatively small. The most extreme observations occur when the true standard deviation is 23, and the average implied standard deviation for the correct answers using this method is 21.6.

[26] This treatment is similar in spirit to Merkle and Ungeheuer (2021). When studying market betas they display a drop down menu centered at the true value, but add random noise to this value.

[27] To circumvent the issue of extreme outliers we throw out draws that are outside the 99% bounds and pick another number in its place.

Now we turn to participant responses. Figure 3 presents binscatter plots of the results. The fitted line is in maroon and the correct estimate is the gray dashed line. Figure A shows the results for direct elicitation and Figure B shows the results for the distribution method.

Panel A shows that the direct method estimates are far too high. The maroon line (and blue dots representing the average of each bin) lie far above the dashed gray line. Panel B shows that the distribution method responses are closer to the actual in the that maroon line and the gray line are at a similar level and somewhat overlap. It is also worth noting that the figures seem to indicate greater precision with the distribution method in that the blue dots deviate less from the maroon line.

Table 8 explores these results more formally. Examining the forecast errors in the odd columns of Panel A, the first column shows that participants were too optimistic using the direct elicitation method, stating numbers more that 7 percentage points higher than the mean value they actually observed. Columns 3 and 5 show that the distribution method was fairly accurate. Means were roughly 0.7 percentage points higher than the actual values. Standard deviations were also fairly accurate with values slightly higher than the true estimate by 1.

In addition to the point estimate, we can examine the absolute value of the forecast error which we do in the second row. We find that the overall error was much higher using the direct elicitation method. The values of over 11 are roughly 3 times higher than the distribution based estimates.

Figure 3 also illustrates what seems to be the main shortcoming of the distribution method. The slope of the maroon line is less steep than that of the true gray line. With the two lines intersecting near where the bins are centered at 8%, this does seem to indicate that responses are anchored to some degree by the choice of where the bins are centered. In contrast, the slope of the true (maroon) and estimated (gray) lines representing the direct elicitation method (Panel A) appear closer to parallel suggesting greater sensitivity, although with a positive bias.

We turn to the regression analogue of this test in Table 8 Panel B. We regress elicited values on the displayed values. All coefficients show a strong positive correlation, significant at the 1% level. The direct method is fairly accurate at reflecting different means with coefficients of roughly one.[28]

The distribution method is responsive, but less than fully. While perfect responses would yield coefficients of 1, instead coefficients are generally around 0.5. This formalizes one shortcoming of the distribution method. The choice of centering is not innocuous and will influence estimate results,

[28]Our paper winsorizes direct elicitation means at the 99% level. While this choice is fairly innocuous in most of our settings, in this instance we find different results based on winsorisation choices. For example, without winsorisation coefficients are nearly 2, and winsorising at the 95% interval they are about 0.8.

anchoring responses towards the value chosen for the middle of the distribution.

IV. Alternative Participant Populations

All results presented so far have been from a population drawn from Prolific Academic.[29] This sample is likely relevant to understanding how the general population responds to surveys, which is often where such questions are asked. With that said, there are surveys conducted on populations that are more familiar with financial settings and the type of distributions we examine. Finance professionals are also more likely to be comfortable with mathematical concepts which may affect their ability to respond to different question types. To examine this, we re-run our continuous treatment on a sample with significant financial holdings and also a more mathematically sophisticated subject pool. This population was recruited through Luth Research. Those with significant financial holdings reported assets of $100,000 or more and those with mathematical sophistication reported holding undergraduate or masters level degrees in engineering, physics, mathematics, or statistics.[30]

We start with our high financial wealth sample. Table 1 Panel C shows summary statistics. This sample has an average portfolio value of about $400,000 and an average income of about $125,000. They are about a decade older on average than our base sample, more college educated (∼90%) and weighted more male, at 76%. In addition to high financial wealth they trade fairly often, with a median number of trades of about 10 per year. Thus if our results were due to a sample that was unfamiliar with financial markets, examining this sample might yield better calibrated results. To test whether this was the case, we used the continuous treatment, but referred to the numbers as stock returns to facilitate use of specific financial knowledge.

Table 9 shows the results, which fail to provide evidence that this sample performs better at the direct elicitation task than the lay sample. Both the mean error and the absolute value of the forecast error have a higher point estimate than that found for the lay sample. The distribution method performs broadly similarly across samples. Thus, the basic patterns from the lay sample seem to hold in this sample of participants with significantly more experience in financial markets.

Of course, our experiment is a recall task and the difficulty may be more in the intuitive application of statistical knowledge rather than financial knowledge. To examine whether this is the case we next examine a sample of participants with degrees in either statistics, engineering, math, or physics.

Table 1 shows summary statistics. Compared to our baseline sample, this sample is older (aver-

[29] With the exception of the reference to the high financial wealth sample in Section D, which we detail in this section.

[30] These studies were pre-registered on AsPredicted. High asset sample: #178727; (https://aspredicted.org/DW5_P9Y); Mathematical sample: #178641 (https://aspredicted.org/6Q5_6CW)

age of 53 years), wealthier (average income of about $120,000), more male (75%) and more educated (nearly half have a graduate degree or higher). In addition to basic demographic questions, we also asked participants to "rate your understanding of statistics relative to the general US population." Participants did so on a seven point Likert scale where 1 was labeled "Much lower than the general US population", 4 was labeled "Same as the general US population" and 7 was labeled "Much higher than the general US population." About 80% of the sample rated themselves above the US population (5 or above), with about 40% rating themselves 6 or above.

Table 10 shows that this sample does appear to respond more accurately to these question prompts. Examining Panel A, the point estimate of the error is lower for both direct elicitation and also for the distribution method relative to the lay population.

Table 11 repeats the analysis for the sample rating themselves with particularly high statistical knowledge, 6 or higher. This sample performs even better. Relative to the correct answer forecast errors are 7.4 for the lay sample, 5.0 for the math sample generally and 3.7 for the high statistical knowledge sample. Thus the higher the statistical knowledge, the more precise the answer to these questions.

Interestingly, the same pattern emerges when examining answers based on the distribution method. The lay sample has an error of 0.7 (t-statistic of 3.27), the math sample has an overall forecast error of 0.4 (t-statistic of 1.46) and the high statistical knowledge sample has an error of 0.2 (t-statistic of 0.49). Thus it appears that greater sophistication leads to more precise responses using both question types while the distribution method remains more accurate than direct elicitation.

V. Further Exploration of the Results

This section delves into further details of the results for each question type to better understand why we find the pattern that we are observing.

A. Direct Elicitation

In each instance that we directly elicit a mean return, we find numbers that are too high. To further understand this pattern, we show a histogram of responses in Figure 4. Panel A shows results from the US and Panel B shows results from Japan. The true average in the data is the vertical gray dashed line and the mean of the survey responses is the red vertical line. The main stylized fact is that the red line is far to the right of the gray line.

It is worth noting that there is a significant mass (relative to surrounding answers) at the mean

response of 8% for the US and 10% for Japan. With that said, these bars are swamped by the surrounding bars which are multiples of 5. While 8% is a large mass of answers in the US data, this is less than the responses given for 10%, 15% and 20%. For Japan, 10% is of course a multiple of 5, but there are more responses given at 15% and 20%. We note that this pattern is similar to those observed in other survey settings (e.g. a similar pattern of rounding is found for inflation expectations using the Chicago Booth Expectations in D'Acunto et al. (2021b) and the Michigan Consumer Sentiment Survey in D'Acunto et al. (2023) and Vanguard investors Giglio et al. (2021)).

This pattern seems to be an important component of why direct elicitation estimates are too high. While people are often able to extract mean properties from distributions presented to them (Chong and Treisman, 2003; Whitney and Yamanashi Leib, 2018), computing a mean from a string of numbers that are presented sequentially is, arguably, a difficult task and people may not have a precise answer in their mind. Instead of responding with a precise estimate, it appears that participants are coming up with an approximate estimate and rounding to provide what they believe is a reasonable, if not perfectly accurate, response. While a useful heuristic in many settings, in practice here it leads to a large bias in level estimates.

B. Distribution Method

Next we examine the results using the distribution method in more detail. We do so not only to provide evidence that it seems to yield more meaningful results but also to highlight its limitations. The results help to guide our discussion of constructing questions to elicit beliefs in the next section.

Figure 5 graphs the true answers in the clear bars relative to the average given in red for participants who saw US data. The figures show that the answers shift as the bins change. For example, when asking questions with bins centered at the means of the distributions, the true distributions are farther to the left when compared to those centered at zero, and the responses reflect this.

More formally, the figures report the χ^2 statistic. This is calculated as the sum of the squared difference between each red bar (reported estimate) and the clear bar (the actual).[31] The p-values for US responses based on the US bins are both above traditional cutoffs (both values >0.25), failing to reject the null that the two distributions are the same. Figure 6 displays similar figures based on the Japanese data.

We also explore the underlying histograms of the continuous treatment in Figures 7 and 8. Each

[31] We report the χ^2 statistic only for figures where each bar has a non-zero mass for the actual, otherwise the statistic is undefined. The p-value is reported based on 8 degrees of freedom.

respondent sees a unique combination of M and S, so we group respondents based on quartiles of means in Figure 7 and different standard deviations in Figure 8. Visually, the responses track variation of the actual clear bins fairly well. Of the eight figures there is only one (the lowest standard deviation) which shows significant differences between the correct and reported distributions as described by p-values based on the χ^2 statistic (all other ps >0.20).

While the figures do a fairly good job and are responsive to the different bins, things are in general muted at the extremes, just as in Goldstein and Rothschild (2014). Namely, the bar that should have the highest value based on the distribution provided tends to have slightly less mass than it should while the tails tend to have slightly more. This pattern is consistent with participant inferences about survey design such as responding based on gricean norms and beliefs that they should be putting some mass in each of the bins (e.g., Bergman et al. (2020); Altman (1983); Hu and Simmons (2023)).

Further consistent with the importance of sensible bin construction, note that when the bin range extends well beyond the data, for example when Japanese bins are used for US data, there is positive mass in these bins. Such noise is an inherent aspect of any survey.

Different bin constructions appear to influence participant accuracy. For example, participants who see US data report a mean that is about 2% higher with the bins centered at the US mean of 8% relative to when the bins used are centered at the Japanese mean of 10%. Also, participants report a SD of 18 using US bins and 23 using Japanese.

Table 12 explores why this is the case. The first row of each table shows the estimate of a given parameter that occurs using each bin construction. The second row shows the "correct" answer, that is what the implied mean or standard deviation would be if someone were to respond with the exact number of observations in the data for each bin.

Examining the results, the pattern appears to be largely driven by variation due to bin size and how we translate responses to implied moments. For example, the estimate of 8.83 using the US mean bins is quite close to the correct answer (i.e., the mean implied by completely accurate responding) of 8.12. The estimate using the Japan bins of 6.49 is quite close to the correct answer of 6.53. Thus, some of the variation is driven by variation in the correct answer. Participants remain quite accurate using most bin constructions.

We note that this specific issue is likely less relevant when asking about future expectations. It would be an odd set of expectations that put a significant mass on a return of 118%, but no other mass in that neighborhood. For future expectations, the more relevant concern is if the data used to

calibrate the bins is not appropriately matched to the distribution of expectations. Consistent with this, in the continuous treatment, there are minimal differences between a correct estimate and a true standard deviation see Figure 2. For this reason, in the next section we suggest winsorising such extreme observations before calibrating bins.

This speaks to the importance of bin construction when eliciting the beliefs. While participants seem to be able to respond fairly accurately to most bin constructions, bin construction does influence relative accuracy of responses relative to the true values. In general, the bin calibrations corresponding to the country where the data are from (i.e., the true underlying distribution) yield correct answers closer to the truth. For the US, the sample average is 7.97 and the two correct US numbers are very close at 8.18 and 8.01. The Japanese bins yield noisier correct values further away from the truth at 6.56 and 8.7. As the the center of the bins moves further from the true mean, responses become less accurate, as evidenced by responses to the selection of bins with mean zero. In this case, the selection of mean has no connection to the underlying distribution and may be confusing for participants. Note that even in this case, the distribution method outperforms direct elicitation in terms of accuracy.

The outliers in the Japanese data yields bins that are quite wide, which directly influence on the distance from "correct" answers to the "true" answers. As discussed below, in settings asking about future expectations, it likely makes sense to account for such outliers before constructing the bins.

B.1 Number of Bins

Thus far, we have limited our attention to nine bins as it seemed to offer enough variation to allow participants to express their beliefs, while still being relatively parsimonious. Ultimately whether this achieves its goal is an empirical question which we can test. This section finds that nine bins seems at least as good as, if not slightly better, than the other options which we explore.

In this section we repeat the continuous treatment, but we vary the number of bins to explore five bins, nine bins and 13 bins together. We keep the bins based on the US and attempt to cover the same range in each treatment, meaning the open-ended bins are meant to be as close as possible to our baseline nine bin assumption.[32] Other than changing the bins, everything else is identical.

Table 13 explores forecast errors and finds that using five bins leads to the largest errors, while using nine bins appears at least as accurate as thirteen. Panel A explores the error for elicited means. The five bin method yields estimates of the mean roughly 1% too high. The 13 bin method and the

[32]The nine bin and five bin treatments have the same open-ended bins of greater than 39% and less than -23%, but the 13 bin treatment uses greater than 33% and less than -19% in order to keep the bins in integer values.

nine bin method are both more accurate and the nine bin estimate is closest to the true value. Turning to the standard deviation in Panel B, the five bin method again yields the highest error. The nine bin and the 13 bin are both closer to the true value, with the 13 bin yielding results closest to the correct estimate. Panel C and D explore the absolute value of the forecast error which again displays that the five bin seems to perform worse along this metric, while the nine and 13 bin perform similarly.

Table 13 Panel C explores how these estimates vary with different values of M and S. The results suggest that the five bin version is slightly more responsive in means, with a coefficient of 0.58, compared to 0.47 for the 9 bin and 0.41 for the 13 bin. The results for the standard deviation exhibit less of a difference, with the five bin and nine bin being roughly the same.

Overall we think the results suggest that our baseline choice of nine bins is appropriate. While there may be a small benefit in fewer bins in responding to different means, this benefit seems swamped by the cost of the positive bias in level of the elicited mean. This is similar to the results that we found with direct elicitation. It is consistent with the benefits of the distribution method forcing individuals to contemplate the problem more deeply (e.g., Goldstein and Rothschild (2014)), and that five fairly wide bins are not as effective at accomplishing this task.

VI. Constructing Questions to Elicit Beliefs

Based on these results, we have a number of suggestions for conducting these types of surveys regarding use of the distribution method, confidence intervals, and direct elicitation. Given the overall outperformance of the distribution method in eliciting responses that accurately map on to the mean and variance of the distribution being considered, we propose using this approach as the default. In contrast, we recommend avoiding confidence interval questions when aiming to estimate levels of variance. Based on our results and prior papers, participants struggle to meaningfully respond to the range of a given confidence interval. If participants are not responding based on the range given, it is unclear how one would effectively transform the response to an implied volatility estimate.

Evidence suggests that direct elicitation can have specific advantages, but it should be used with caution. Direct elicitation of means appears to yield meaningful ranking and comparisons, but it carries with it a positive bias. Consequently, we recommend focusing the use of direct elicitation on cases when the experimenter's aim is to make direct comparisons across time or experimental treatments.

An additional benefit of direct elicitation is that it yields responses which are unaffected by in-

formation communicated through the choice of bins and does not require assumptions to translate participant responses to means. If direct elicitation can be responded to effectively, it would be preferred to the distribution method (based on the shortcomings of using bins discussed above). With that said, none of the populations that we examined were able to respond more accurately based on direct elicitation. Thus it is unclear precisely what such a population where direct elicitation is preferred should be. However, evidence from our mathematically sophisticated population suggests that direct elicitation could be useful if the sample population is sufficiently well calibrated to a given economic setting and comfortable enough with statistical concepts.

While by no means perfect, the evidence suggests that in most settings the distribution method is likely to be preferred. However, our results also show that choices of bin construction matter and that appropriately constructing such bins will likely lead to more meaningful results.

In most settings, when asking for future expectations, calibrating on past data is likely appropriate. The exact choices should reflect the context of the expectations being measured as to what is reasonable. In addition to determining the relevant time period, there is an interesting question as to how best to handle extreme observations. For example, the Japan based bins we use are quite wide compared to the US, 9 versus 15. While some of this is driven by stable differences, much simply comes from including the outlier observations around World War II. When asking about future returns, it seems likely that an investor in Japan places a probability very close to 0 of achieving returns of 118%, and thus such number probably shouldn't be particularly important to bin construction.

This speaks to picking the relevant sample from which to construct the bins. Looking before 1980, the Japanese sample is volatile with a SD of 30 and high returns of 13.7. In recent samples, returns have been stable with a SD of 20 and low returns of 5.5. Thus, if one were to ask investors about their current expectations, it likely makes more sense to calibrate on more recent data for bin construction.

VII. Conclusion

This paper develops a method to identify the accuracy of responses to questions about volatile distributions of percent changes centered at a small positive number. We find that people can accurately identify means and volatilities in this setting. Irrespective of question format, participants are able to identify whether means are higher or lower and whether distributions are more or less volatile. When it comes to identifying levels though, we find that direct elicitation yields less accurate results. Instead, procedures asking participants to report full distributions yield more meaningful results.

While the paper documents limitations of direct elicitation, this approach maintains certain advantages over the distributional method in theory. In particular, direct elicitation does not provide participants with benchmarks (e.g., bin ranges) that participants might view as informative when entering their responses. Further, since participants state the distribution's mean directly, researchers do not add noise to the response as is the case when interpreting binned responses. Theoretically, these benefits could allow participants to state their true beliefs more accurately.

Importantly, the studies in our paper are able to test participant's ability to report on known belief distributions. This allows us to determine, empirically, that participants are not capable of responding to questions directly asking about their beliefs in the absence of prompts requiring participants to consider the full distribution. Thus, in practice (vs. theory), researchers can more effectively recover the true underlying statistical properties using the distribution method. This superior performance holds in cases where the bins provided vary substantially, suggesting a limited role of the bin selection influencing participant responses, and in spite of limitations regarding interpretation.

We do not explore survey variants in similar settings which are asked in a manner that avoids percentage change distributions. For example, some surveys do not ask for beliefs about a percentage change, but instead ask for the odds the variable of interest will increase,[33] or what level the variable will be in the future (e.g. Glaser et al. (2019)). This question type inherently provides less information than either the direct elicitation or distribution methods that we examine, but the response may be sufficient to answer certain economic questions. In these cases, it is not yet clear which question type yields more accurate responses. A comparison of a broader set of methods for eliciting beliefs is worthy of future study.

While our results point to the general validity of survey evidence, they also point to its limitations. It is clearly a mistake to take all survey evidence as equal or to ignore the noise inherent in such methods. We hope that by better understanding such issues through a systematic study of survey design in such settings, we can effectively and appropriately employ survey data.

Our results show that participants are able to effectively report properties of a known distribution. This suggests that as long as participants have a distribution in mind, they should be able to respond to survey prompts accurately. It is an open question as to how and whether participants actually think about such parameters though. For example, when making choices about the stock market, how do participants typically conceive of future potential outcomes? Do these beliefs take the form of

[33] e.g. The Michigan Survey of Consumers asks for the probability of a stock market increase.

distributions and if not, what form do they take?

These are important questions that lie beyond the scope of the current investigation. We note that if investors do not meaningfully form expectations on something approximating a return distribution, this is an important aspect of expectation formation that needs to be understood and should be incorporated as a fundamental component of future models. We leave the study of this important question for future papers.

References

Adam, Klaus and Stefan Nagel, "Expectations data in asset pricing," in "Handbook of Economic Expectations," Elsevier, 2023, pp. 477–506.

Alpert, Marc and Howard Raiffa, "21. A progress report on the training of probability assessors," 1982.

Altman, Edward I, "A complete guide to predicting, avoiding, and dealing with bankruptcy," *Corporate Financial Distress, New York*, 1983.

Andre, Quentin, "distBuilder v 1.1.0," November 2016.

André, Quentin, Nicholas Reinholtz, and Bart De Langhe, "Can consumers learn price dispersion? evidence for dispersion spillover across categories," *Journal of Consumer Research*, 2022, *48* (5), 756–774.

Ang, Andrew, Geert Bekaert, and Min Wei, "Do macro variables, asset markets, or surveys forecast inflation better?," *Journal of monetary Economics*, 2007, *54* (4), 1163–1212.

Ariely, Dan, George Loewenstein, and Drazen Prelec, ""Coherent arbitrariness": Stable demand curves without stable preferences," *The Quarterly journal of economics*, 2003, *118* (1), 73–106.

Armantier, Olivier, Wändi Bruine de Bruin, Simon Potter, Giorgio Topa, Wilbert Van Der Klaauw, and Basit Zafar, "Measuring inflation expectations," *Annu. Rev. Econ.*, 2013, *5* (1), 273–301.

Becker, Christoph K, Peter Duersch, and Thomas Eife, "Measuring inflation expectations: How the response scale shapes density forecasts," *Available at SSRN 4323706*, 2023.

Ben-David, Itzhak, John R Graham, and Campbell R Harvey, "Managerial miscalibration," *The Quarterly journal of economics*, 2013, *128* (4), 1547–1584.

Bergman, Abigail, Alex Chinco, Samuel M Hartzmark, and Abigail B Sussman, "Survey curious? start-up guide and best practices for running surveys and experiments online," *Start-Up Guide and Best Practices For Running Surveys and Experiments Online (October 05, 2020)*, 2020.

Boctor, Valerie R, Olivier Coibion, Yuriy Gorodnichenko, and Michael Weber, "On Eliciting Subjective Probability Distributions of Expectations," Technical Report, National Bureau of Economic Research 2024.

Boutros, Michael, Itzhak Ben-David, John R Graham, Campbell R Harvey, and John W Payne, "The persistence of miscalibration," Technical Report, National Bureau of Economic Research 2020.

Chong, Sang Chul and Anne Treisman, "Representation of statistical properties," *Vision research*, 2003, *43* (4), 393–404.

Crosetto, Paolo and Thomas de Haan, "Comparing input interfaces to elicit belief distributions," *Judgment and Decision Making*, 2023, *18*, e27.

D'Acunto, Francesco, Ulrike Malmendier, and Michael Weber, "What do the data tell us about inflation expectations?," in "Handbook of economic expectations," Elsevier, 2023, pp. 133–161.

Davidson, LB and DO Cooper, "A simple way of developing a probability distribution of present value," *Journal of Petroleum Technology*, 1976, *28* (09), 1069–1078.

de Bruin, Wändi Bruine and Baruch Fischhoff, "Eliciting probabilistic expectations: Collaborations between psychologists and economists," *Proceedings of the National Academy of Sciences*, 2017, *114* (13), 3297–3304.

de Bruin, Wandi Bruine, Wilbert Van der Klaauw, Julie S Downs, Baruch Fischhoff, Giorgio Topa, and Olivier Armantier, "Expectations of inflation: The role of financial literacy and demographic variables," *Journal of Consumer Affairs*, 2010, *44* (2), 381–402.

Deese, James and Roger A Kaufman, "Serial effects in recall of unorganized and sequentially organized verbal material.," *Journal of experimental psychology*, 1957, *54* (3), 180.

D'Acunto, Francesco and Michael Weber, "Why survey-based subjective expectations are meaningful and important," *Annual Review of Economics*, 2024, *16*.

D'acunto, Francesco, Daniel Hoang, Maritta Paloviita, and Michael Weber, "IQ, expectations, and choice," *The Review of Economic Studies*, 2023, *90* (5), 2292–2325.

D'Acunto, Francesco, Ulrike Malmendier, and Michael Weber, "Gender roles produce divergent economic expectations," *Proceedings of the National Academy of Sciences*, 2021, *118* (21), e2008534118.

_ , _ , Juan Ospina, and Michael Weber, "Exposure to grocery prices and inflation expectations," *Journal of Political Economy*, 2021, *129* (5), 1615–1639.

Giglio, Stefano, Matteo Maggiori, Johannes Stroebel, and Stephen Utkus, "Five facts about beliefs and portfolios," *American Economic Review*, 2021, *111* (5), 1481–1522.

Glaser, Markus, Zwetelina Iliewa, and Martin Weber, "Thinking about prices versus thinking about returns in financial markets," *The Journal of Finance*, 2019, *74* (6), 2997–3039.

Goldstein, Daniel G and David Rothschild, "Lay understanding of probability distributions," *Judgment and Decision making*, 2014, *9* (1), 1–14.

Greenwood, Robin and Andrei Shleifer, "Expectations of returns and expected returns," *The Review of Financial Studies*, 2014, *27* (3), 714–746.

Hafer, Rik W and Scott E Hein, "On the accuracy of time-series, interest rate, and survey forecasts of inflation," *Journal of Business*, 1985, pp. 377–398.

Haran, Uriel, Don A Moore, and Carey K Morewedge, "A simple remedy for overprecision in judgment," *Judgment and Decision Making*, 2010, *5* (7), 467–476.

Hu, Beidi and Joseph P Simmons, "Does constructing a belief distribution truly reduce overconfidence?," *Journal of Experimental Psychology: General*, 2023, *152* (2), 571.

Jung, Minah H, Alice Moon, and Leif D Nelson, "Overestimating the valuations and preferences of others.," *Journal of Experimental Psychology: General*, 2020, *149* (6), 1193.

Kahneman, Daniel, Barbara L Fredrickson, Charles A Schreiber, and Donald A Redelmeier, "When more pain is preferred to less: Adding a better end," *Psychological science*, 1993, *4* (6), 401–405.

Krosnick, Jon A, Sowmya Narayan, and Wendy R Smith, "Satisficing in surveys: Initial evidence," *New directions for evaluation*, 1996, *1996* (70), 29–44.

Liu, Hongqi, Cameron Peng, Wei A Xiong, and Wei Xiong, "Taming the bias zoo," *Journal of Financial Economics*, 2022, *143* (2), 716–741.

Malmendier, Ulrike and Stefan Nagel, "Depression babies: Do macroeconomic experiences affect risk taking?," *The quarterly journal of economics*, 2011, *126* (1), 373–416.

Mandel, David R., Robert N. Collins, Evan F. Risko, and Jonathan A. Fugelsang, "Effect of confidence interval construction on judgment accuracy," *Judgment and Decision Making*, 2020, *15* (5), 783–797.

Manski, Charles F, "Measuring expectations," *Econometrica*, 2004, *72* (5), 1329–1376.

_ , "Survey measurement of probabilistic macroeconomic expectations: progress and promise," *NBER Macroeconomics Annual*, 2018, *32* (1), 411–471.

Merkle, Christoph and Michael Ungeheuer, "Beliefs about Beta: Upside Participation and Downside Protection," *Available at SSRN 3794224*, 2021.

Moore, Don A and Amelia S Dev, "Overconfidence," *Encyclopedia of Personality and Individual Differences*, 2020, pp. 3382–3386.

_ **, Elizabeth R Tenney, and Uriel Haran**, "Overprecision in judgment," *The Wiley Blackwell handbook of judgment and decision making*, 2015, *2*, 182–209.

_ **, Samuel A Swift, Angela Minster, Barbara Mellers, Lyle Ungar, Philip Tetlock, Heather HJ Yang, and Elizabeth R Tenney**, "Confidence calibration in a multiyear geopolitical forecasting competition," *Management Science*, 2017, *63* (11), 3552–3565.

Murre, Jaap MJ and Joeri Dros, "Replication and analysis of Ebbinghaus' forgetting curve," *PloS one*, 2015, *10* (7), e0120644.

Prims, Julia P and Don A Moore, "Overconfidence over the lifespan," *Judgment and decision making*, 2017, *12* (1), 29–41.

Reinholtz, Nicholas, Philip M Fernbach, and Bart De Langhe, "Do people understand the benefit of diversification?," *Management Science*, 2021, *67* (12), 7322–7343.

Sharpe, William F, Daniel G Goldstein, and Phil W Blythe, "The distribution builder: A tool for inferring investor preferences," *preprint*, 2000.

Soll, Jack B and Joshua Klayman, "Overconfidence in interval estimates.," *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 2004, *30* (2), 299.

Speirs-Bridge, A, F Fidler, M McBride, L Flander, G Cumming, and M Burgman, "Reducing overconfidence in the interval judgments of experts. Risk Analysis30: 512-523," 2010.

Teigen, Karl Halvor and Magne JØrgensen, "When 90% confidence intervals are 50% certain: On the credibility of credible intervals," *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition*, 2005, *19* (4), 455–475.

Tourangeau, Roger and Kenneth A Rasinski, "Cognitive processes underlying context effects in attitude measurement.," *Psychological bulletin*, 1988, *103* (3), 299.

Tversky, Amos and Derek J. Koehler, "Support theory: A nonextensional representation of subjective probability," *Psychological Review*, 1994, *101* (4), 547–567.

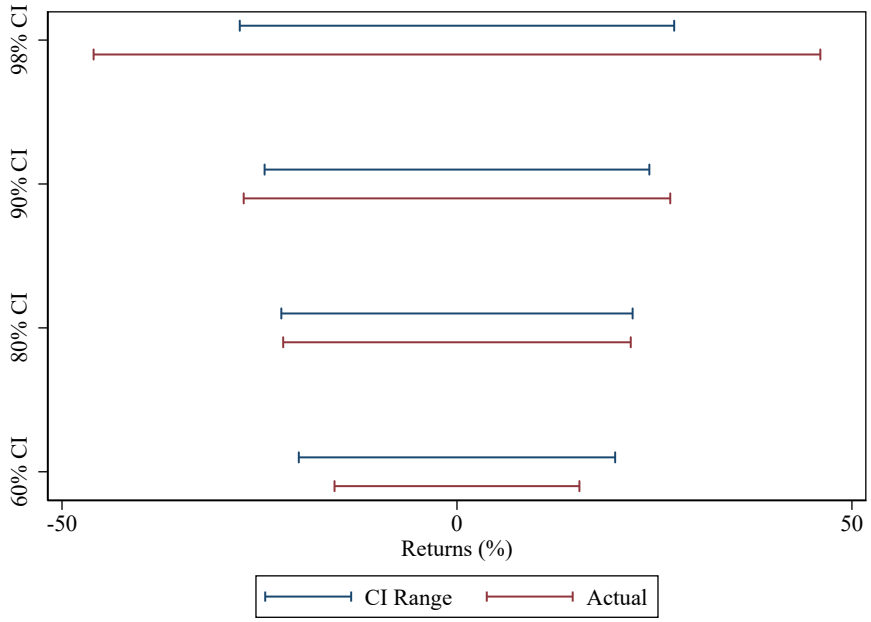
Vissing-Jorgensen, Annette, "Perspectives on behavioral finance: Does" irrationality" disappear with wealth? Evidence from expectations and actions," *NBER macroeconomics annual*, 2003, *18*, 139–194.

Weber, Michael, Francesco D'Acunto, Yuriy Gorodnichenko, and Olivier Coibion, "The subjective inflation expectations of households and firms: Measurement, determinants, and implications," *Journal of Economic Perspectives*, 2022, *36* (3), 157–184.

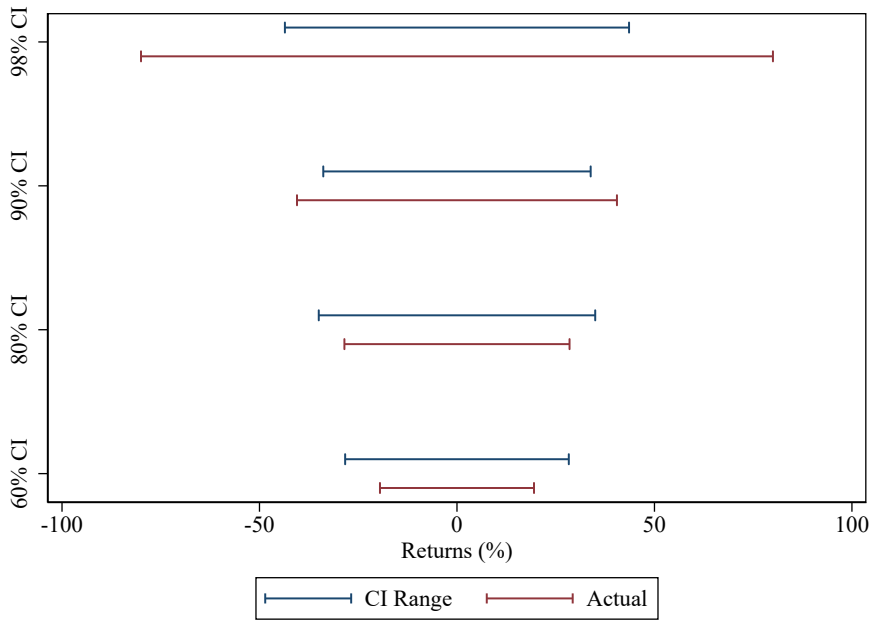
Whitney, David and Allison Yamanashi Leib, "Ensemble perception," *Annual review of psychology*, 2018, *69*, 105–129.

Figure 1: Elicited Confidence Intervals

(a) Panel A: US Confidence Intervals



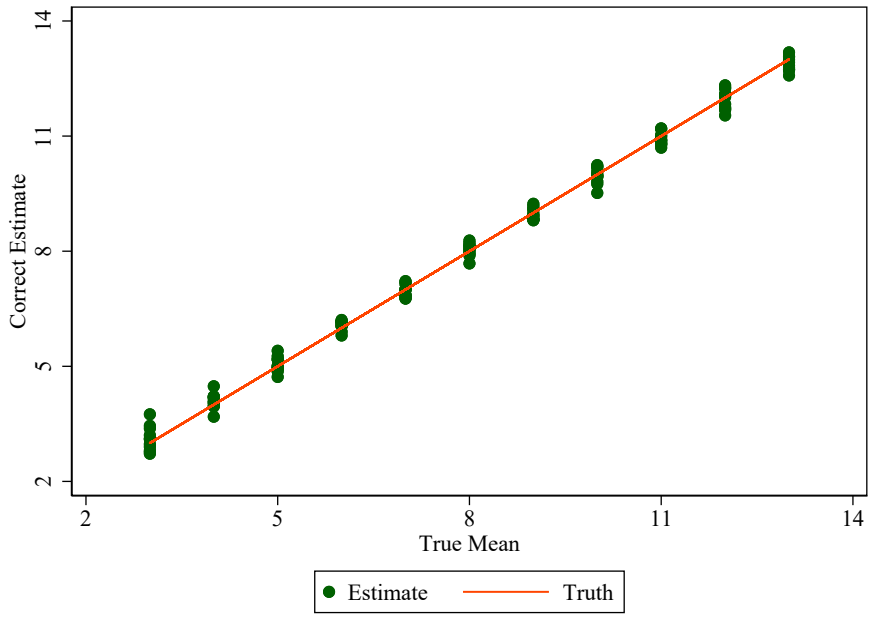
(b) Panel B: Japan Confidence Intervals



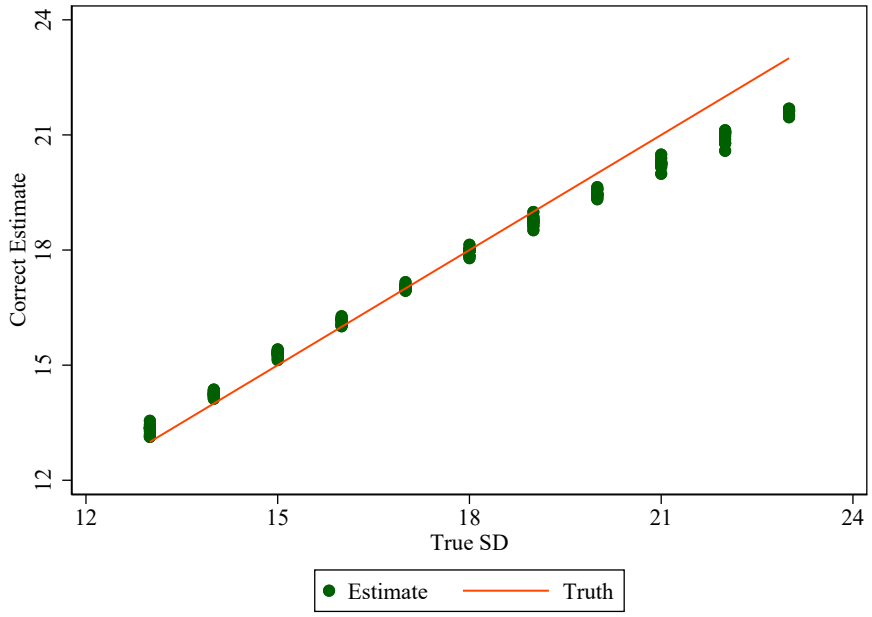
This graphs shows ranges for confidence intervals. For every person we take the upper number minus the lower and average this for a given interval indicated on the y-axis. This value is graphed centered at zero. The Navy line indicates the survey response while the maroon line indicates the sample average. Panel A reports participants who viewed the US data and Panel B shows participants who viewed the Japanese data. The maroon line indicates the actual range.

Figure 2: Implied Value vs. Correct Estimate using Simulated Data

(a) Panel A: Mean Values



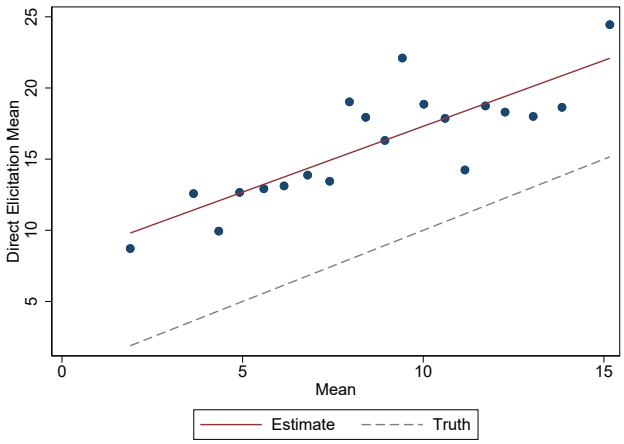
(b) Panel B: Standard Deviation Values



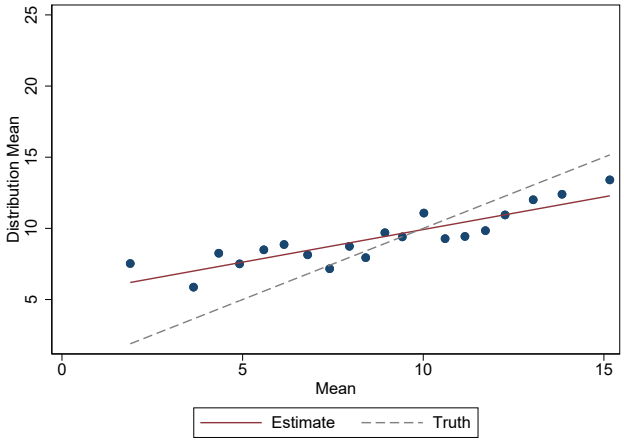
These graphs show scatter plots of the distribution value used (on the x-axis) relative to the implied value based on the distribution method on the y-axis. Each dot is based on a simulation of 10,000 draws from a distribution N(M,S). M ranges from 3 to 13 and S ranges from 13 to 23. Each integer in this range is used. The orange line is the 45 degree line. Panel A shows results for mean estimates and Panel B shows results for the standard deviation.

Figure 3: Estimates vs. Actual Values in the Random Distribution Treatment

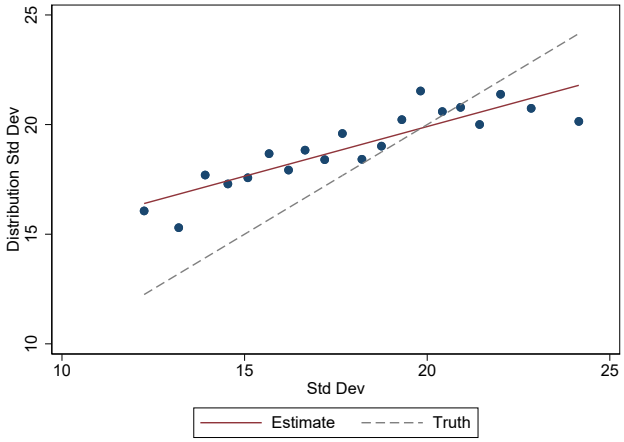
(a) Panel A: Direct Elicitation Means



(b) Panel B: Distribution Method Means



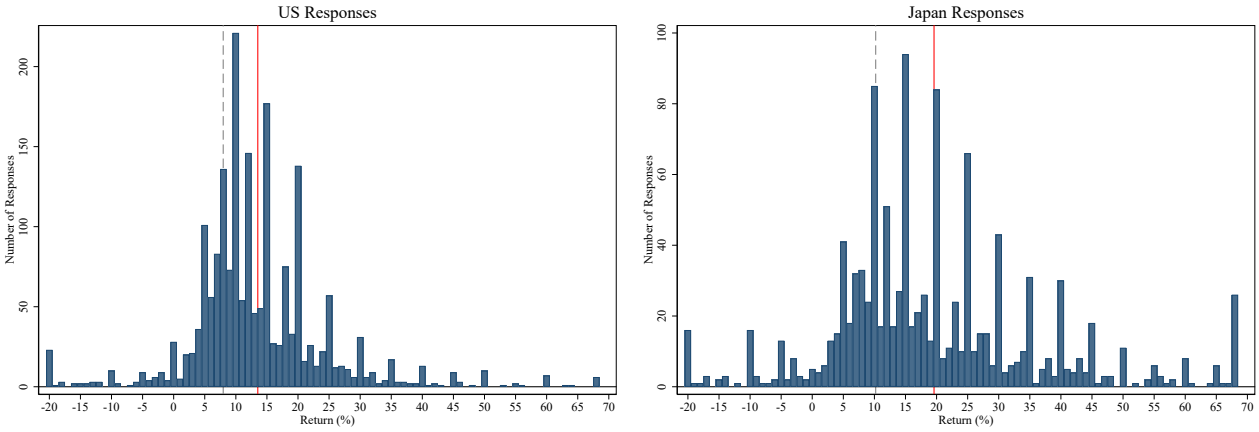
(c) Panel C: Distribution Method Standard Deviations



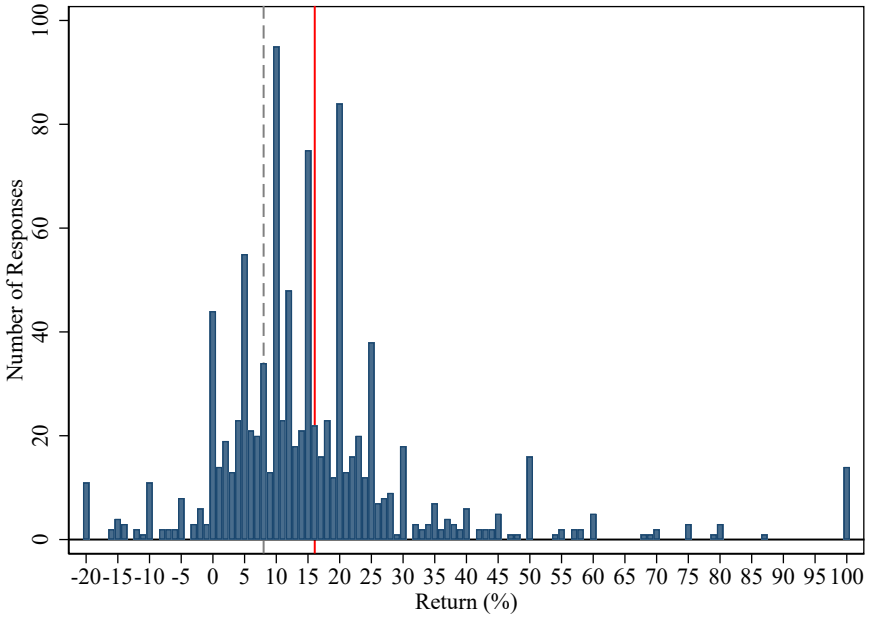
These graphs show binscatter plots of elicited values vs the true values of N(M,S) used to create observed data. Blue dots show the mean in each bin, the maroon line shows the fitted line and the gray dashed line shows the correct true answer of what was displayed. Panel A shows the direct elicitation means, Panel B shows means from the distribution method and Panel C shows standard deviations from the distribution method.

Figure 4: Direct Elicitation: Distribution of Mean Responses

(a) Panel A: US and Japan Responses



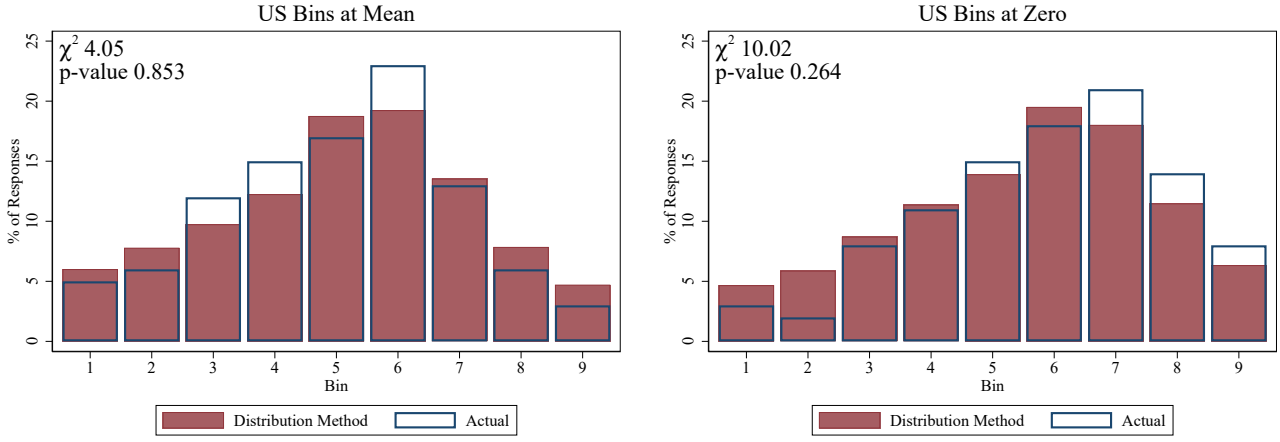
(b) Panel C: Responses from Continuous Treatment



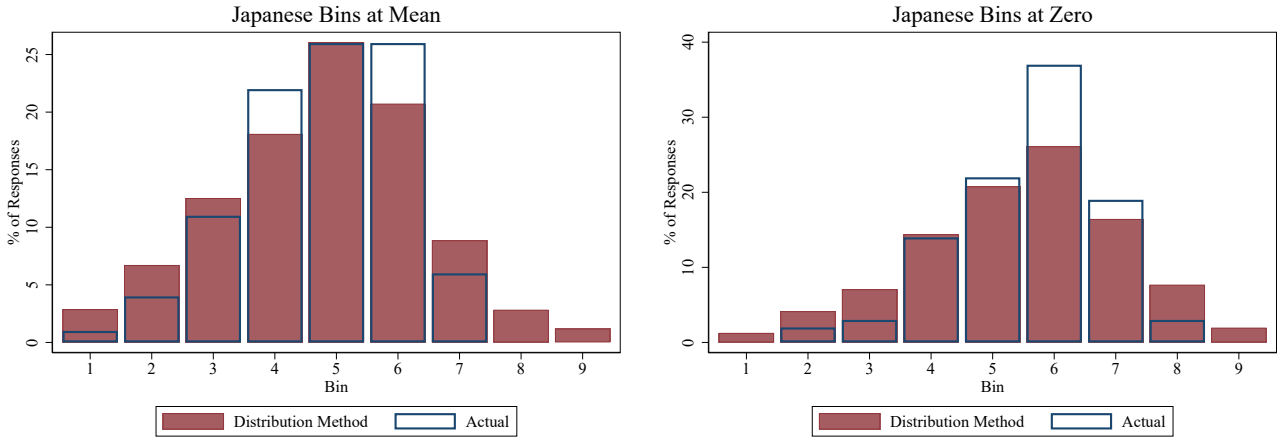
This graphs show the number of responses when elicited directly. Panel A shows responses when subjects viewed the US data and Panel B shows responses based on the Japanese data. The red line indicates the sample average of elicited responses and the gray dashed line indicates the true historical average.

Figure 5: Distribution Treatment Histogram of Responses to US Data

(a) Panel A: US Responses with US Buckets



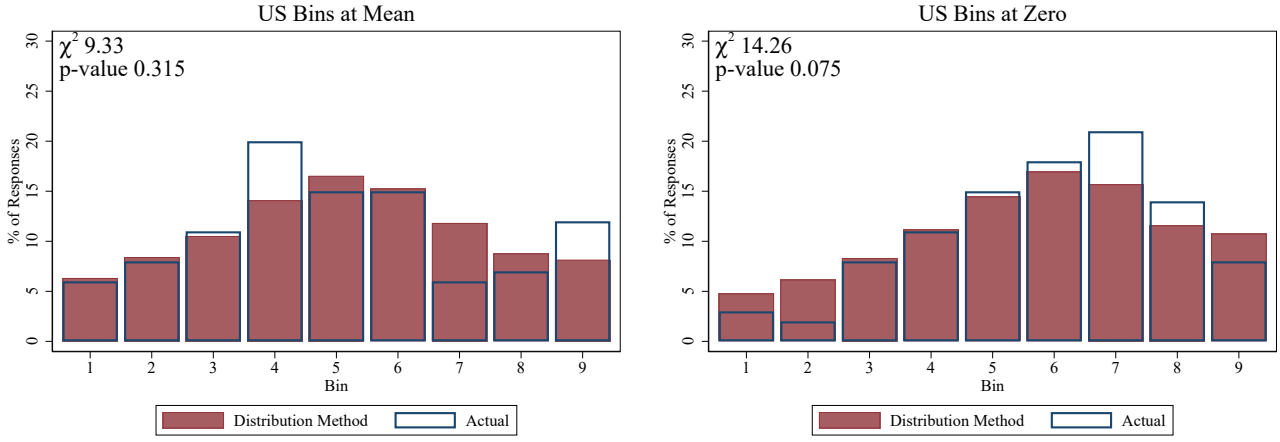
(b) Panel B: US Responses with Japanese Buckets



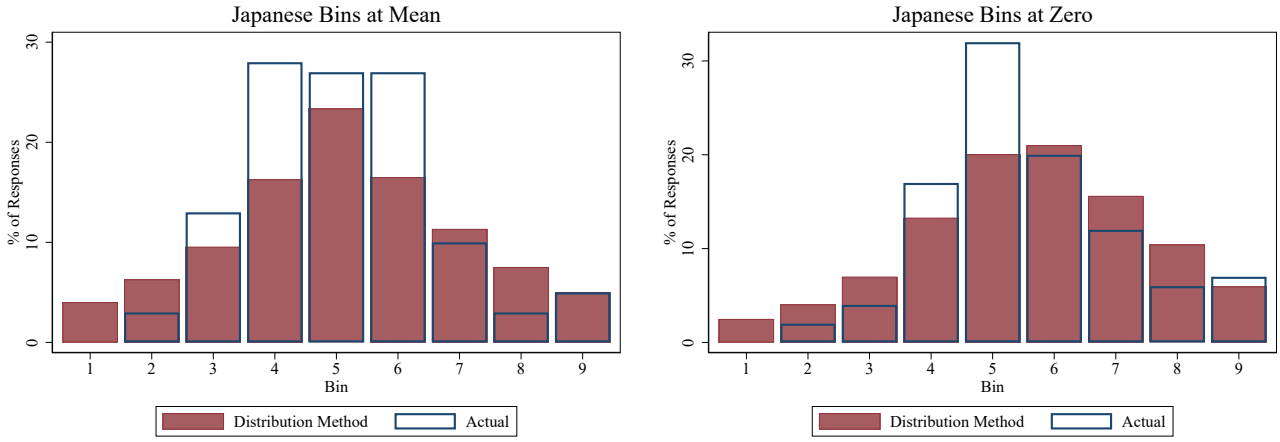
This graphs shows the average probability given to each bin of returns for participants viewing US data. Panel A shows responses created using US data and Panel B shows responses using Japanese data. Figures on the left show responses where bins were centered around historical mean returns and figures on the right shows responses where bins were centered at zero.

Figure 6: Distribution Treatment Histogram of Responses to Japanese Data

(a) Panel A: Japanese Responses with US Buckets



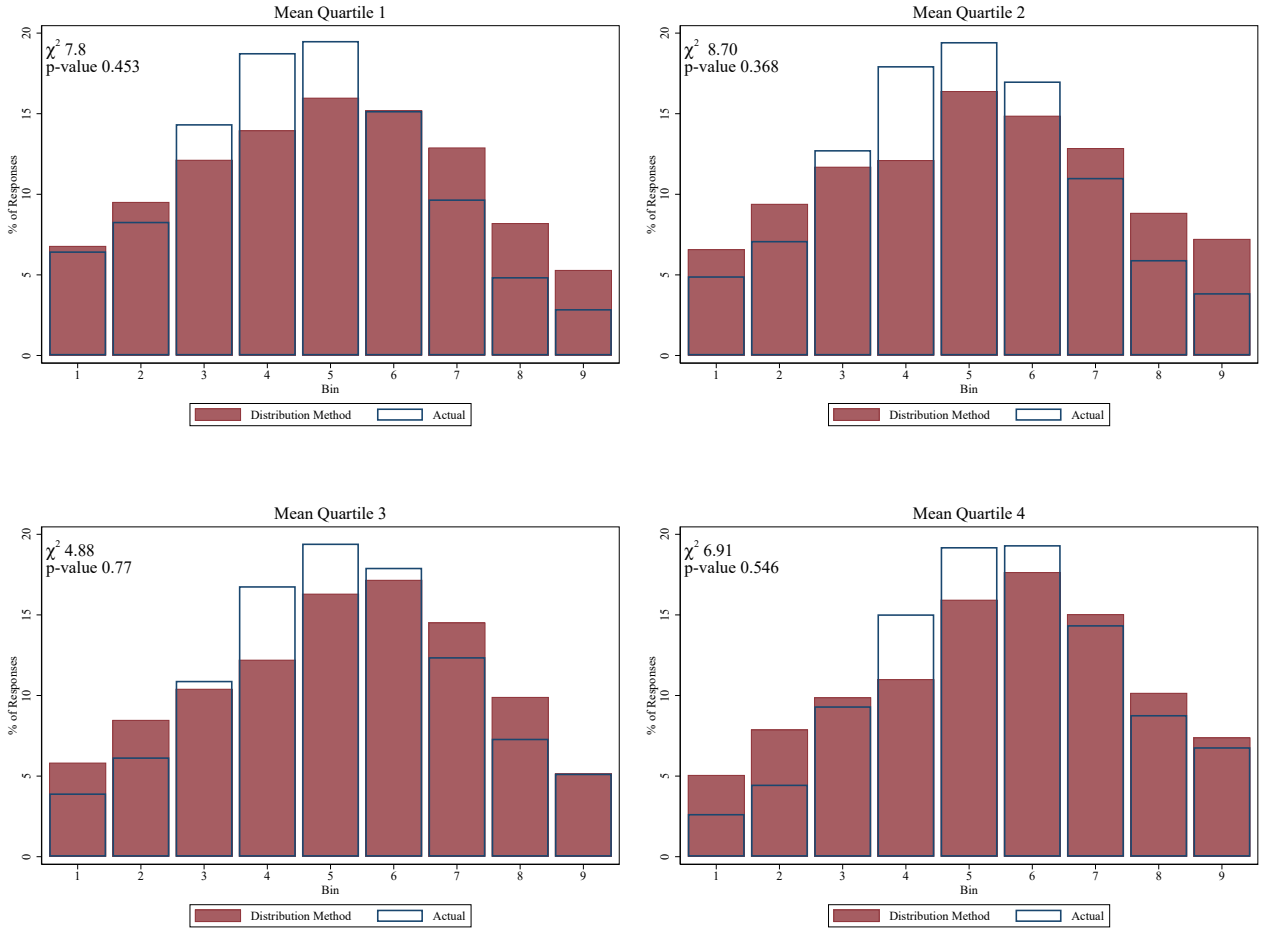
(b) Panel B: Japanese Responses with Japanese Buckets



This graphs shows the average probability given to each bin of returns for participants viewing Japanese data. Panel A shows responses created using US data and Panel B shows responses using Japanese data. Figures on the left show responses where bins were centered around historical mean returns and figures on the right shows responses where bins were centered at zero.

Figure 7: Continuous Treatment Histogram of Responses Based on Means

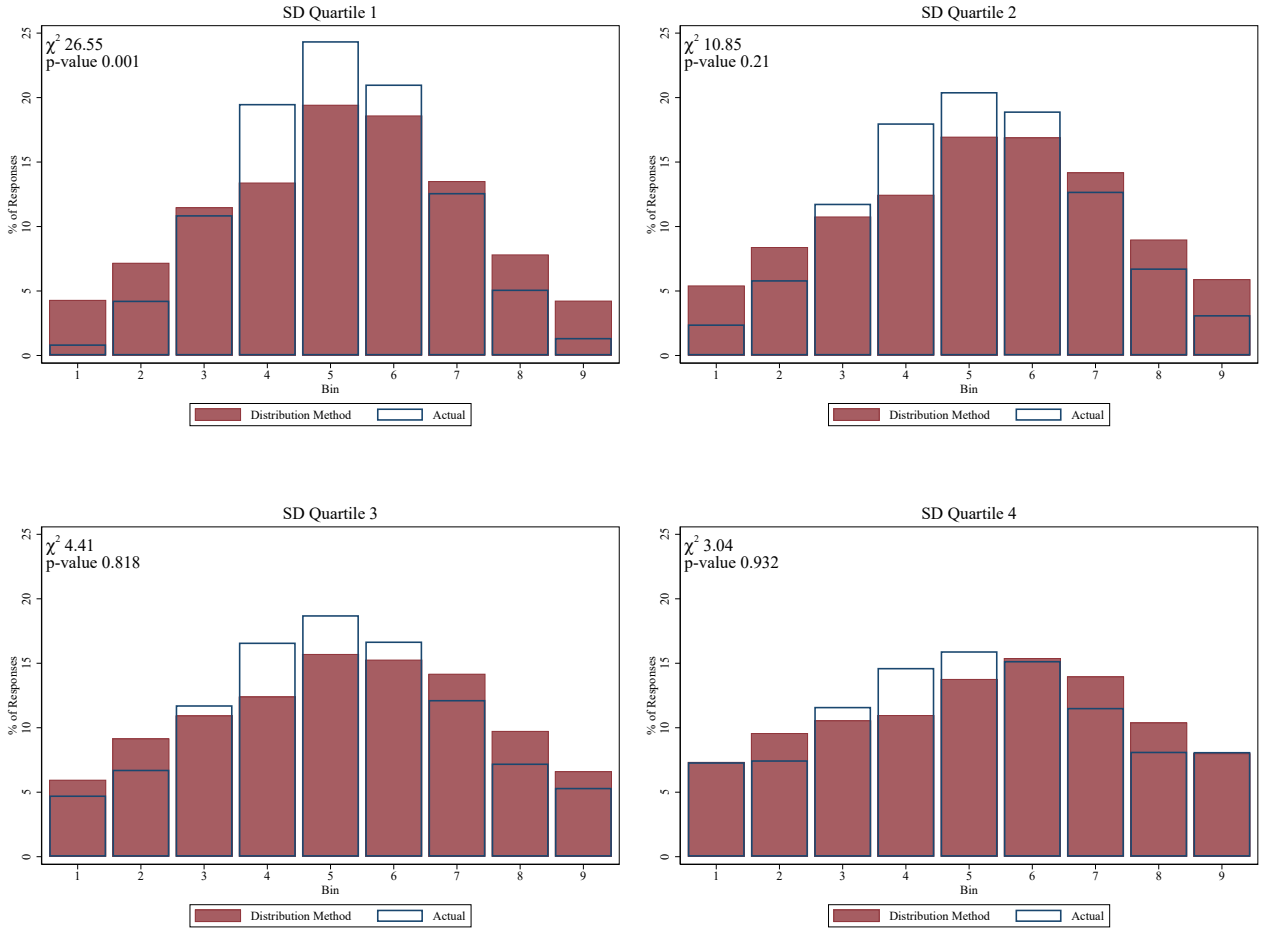
(a) Panel A: Differences in Means



This graphs shows the average probability given to each bin of returns for participants in the continuous treatments where subjects see draws from a distribution N(M,S). Each figure shows responses based on quartiles of M. χ^2 values are calculated as the squared difference between the red and clear bar scaled by the actual value and summed across the bars. p-value is calculated based on 8 degrees of freedom.

Figure 8: Continuous Treatment Histogram of Responses Based on Standard Deviation

(a) Panel A: Differences in Standard Deviations



This graphs shows the average probability given to each bin of returns for participants in the continuous treatments where subjects see draws from a distribution N(M,S). Each figure shows responses based on quartiles of S. χ^2 values are calculated as the squared difference between the red and clear bar scaled by the actual value and summed across the bars. p-value is calculated based on 8 degrees of freedom.

Table 1: Summary Statistics

(a) Panel A: Lay Sample

	Obs	Mean	P25	Median	P75
Age	5712	39.93	29	37	49
Income	5709	77448	35000	65000	112500
Male	5712	0.49	0	0	1
No Degree	5712	0.32	0	0	1
College Degree	5712	0.51	0	1	1
Graduate Degree	5712	0.17	0	0	0

(b) Panel B: Mathematically Sophisticated Sample

	Obs	Mean	P25	Median	P75
Stats Knowledge	493	5.28	5	5	6
Stat Major	493	0.09	0	0	0
Engineering Major	493	0.61	0	1	1
Math Major	493	0.26	0	0	1
Physics Major	493	0.11	0	0	0
Age	491	53.90	40	53	68
Income	493	118722	75000	112500	175000
Male	493	0.75	1	1	1
Graduate Degree	493	0.47	0	0	1

(c) Panel C: High Asset Value Sample

	Obs	Mean	P25	Median	P75
Portfolio Value	509	401326	175000	375000	600000
Trades Per Year	509	9.33	5	10	15
ETF or Mutual Fund	509	0.83	1	1	1
Hold Stocks	509	0.93	1	1	1
Hold Bonds	509	0.54	0	1	1
Trades Options	509	0.63	0	1	1
Age	509	47.13	35	44	59
Income	509	125668	85000	112500	175000
Male	509	0.76	1	1	1
No Degree	509	0.11	0	0	0
College Degree	509	0.48	0	0	1
Graduate Degree	509	0.41	0	0	1

(d) Panel D: Participants per Treatment Type

	Obs
Total	6714
US Only	804
US and Japan	2371
Different Number of Bins (Numbers Only)	1443
Numbers Only	100
Random M and S (Numbers Only)	994
Math Sample (Numbers Only)	493
Finance Sample	509

This table shows summary statistics about participants in our treatments. Panel A shows all participants from the lay sample, Panel B shows participants in our sample of people with degrees in math, statistics, physics or engineering. Stats knowledge reports the average response to the question "How would you rate your understanding of statistics relative to the general US population?" on a scale from (1) Much lower than the general US population to (7) Much higher than the general US population. Panel C describes our sample of people with a financial portfolio of at least $100,000. Panel D shows counts of participants in each type of our treatments.

Table 2: Differences between the US and Japan

(a) Panel A: Mean Estimates

	All		Direct	Distribution
	(1)	(2)	(3)	(4)
Japan	4.585***	4.596***	6.969***	2.257***
	(9.81)	(10.56)	(9.31)	(5.87)
Treatment FE	No	Yes	No	No
R^2	0.0366	0.195	0.0637	0.0249
Observations	3102	3102	1593	1509

(b) Panel B: SD Estimates

	All		Conf Int	Distribution
	(1)	(2)	(3)	(4)
Japan	5.559***	5.020***	7.311***	3.704***
	(12.45)	(11.45)	(8.99)	(11.10)
Treatment FE	No	Yes	No	No
R^2	0.0569	0.148	0.0593	0.0830
Observations	3102	3102	1593	1509

This table shows regressions of reported means (Panel A) or reported standard deviations (Panel B) regressed on a dummy variable equal to 1 if the participant viewed data from the Japanese distribution. t-staistics based on heterscedasticity robust standard errors are in parenthesis.

Table 3: Mean Estimates

(a) Panel A: US Mean Estimates

	Actual Value	Direct Elicitation	Distribution: Mean	Distribution: Zero
Mean	7.97	13.53	8.11	5.18
t-stat (value=actual)		22.29	0.51	-9.48
t-stat (value=correct)		22.29	1.77	-10.71
Obs		1952	475	468

(b) Panel B: Japan Mean Estimates

	Actual Value	Direct Elicitation	Distribution: Mean	Distribution: Zero
Mean	10.21	19.59	10.48	7.21
t-stat (value=actual)		19.12	0.61	-6.03
t-stat (value=correct)		19.12	0.42	-2.31
Obs		1150	286	280

This table shows estimates of means. Panel A shows responses from participants who view US data and Panel B shows responses from participants who viewed Japanese data. Column 1 shows the mean of the sample participants were shown. Column 2 shows responses to the direct elicitation method when it was the first question asked. Column 3 shows estimates from the distribution method with bins centered at the sample average and the average calculated using the bin midpoint. Column 4 is the same as column 3, but the bins used were centered at zero. The second row shows a t-test that the indicated value is equal to the actual value in column 1.

Table 4: Within Person Difference in Means

(a) Panel A: Within Person Difference

	All	US	Japan
	(1)	(2)	(3)
Within Person Difference	7.586***	6.570***	9.279***
	(21.54)	(16.99)	(13.70)
Observations	1509	943	566

This table shows averages of the direct elicitation mean minus the distribution method mean. Column 1 shows this for all estimates, column 2 shows respondents who saw US data and column 3 is from respondents who saw Japan data. t-statistics are in parenthesis.

Table 5: Mean Estimates by Question Order

(a) Panel A: Direct Elicitation First

	Direct Elicitation CI Treatment	Direct Elicitation Distribution Treatment	Distribution Historic Mean	Distribution Zero
Mean	13.78	13.15	8.00	5.02
t-stat (value=actual)	13.04	9.90	0.08	-6.95
Obs	512	465	232	233

(b) Panel B: Direct Elicitation Second

	Direct Elicitation CI Treatment	Direct Elicitation Distribution Treatment	Distribution Historic Mean	Distribution Zero
Mean	13.86	13.30	8.21	5.34
t-stat (value=actual)	11.56	10.21	0.67	-6.44
Obs	497	478	243	235

This table shows estimates of means based on whether the direct elicitaion question was asked first (Panel A) or second (Panel B). Column 1 shows direct elicitation means from the treatment where the other question was about confidence intervals and column 2 shows where the other question was to build a distribution. Column 3 shows means from distributions centered at the historic mean and column 4 shows means from distributions centered at zero. Responses are from participants who view US data. The second row shows a t-test that the indicated value is equal to the actual value.

Table 6: Standard Deviation Estimates

(a) Panel A: US Confidence Interval Method

	Actual Value	98% CI	90% CI	80% CI	60% CI
Standard Deviation	18.35	13.36	14.76	17.19	23.54
t-stat (value=actual)		-10.95	-6.67	-1.78	5.42
Obs		255	251	247	256

(b) Panel B: Japan Confidence Interval Method

	Actual Value	98% CI	90% CI	80% CI	60% CI
Standard Deviation	27.24	21.10	20.21	27.09	29.91
t-stat (value=actual)		-5.17	-5.81	-0.09	-1.63
Obs		145	148	147	144

(c) Panel C: US Distribution Method

	Actual Value	Dist: Mean	Dist: Zero
Standard Deviation	18.35	19.78	19.84
t-stat (value=actual)		5.58	6.01
t-stat (value=correct)		5.22	5.91
Obs		475	468

(d) Panel D: Japan Distribution Method

	Actual Value	Dist: Mean	Dist: Zero
Standard Deviation	27.24	24.00	23.02
t-stat (value=actual)		-7.87	-10.98
t-stat (value=correct)		3.48	1.90
Obs		286	280

This table shows estimates of standard deviations. Panel A and B show responses from participants who viewed US data and panels C and D show responses from those that viewed Japanese data. Panel A and B shows estimates derived from confidence interval elicitation. Panel B and D shows estimates from the distribution method. The top row show the estimated standard deviation. The second row shows a t-test that the indicated value is equal to the actual value in column 1.

Table 7: No Mention of Stock

(a) Panel A: Mean Estimates

	Actual Value	Direct Elicitation	Distribution: Mean	Distribution: Zero
Mean	7.97	13.74	7.81	3.98
Obs		100	47	53

(b) Panel B: Standard Deviation Estimate

	Actual Value	Distribution: Mean	Disttribution: Zero
Standard Deviation	18.35	19.41	20.21
Obs		47	53

This table shows responses from a treatment where numbers were not described as stock market returns, but instead only as numbers. Panel A displays mean values and panel B displays standard deviations. All respondents viewed US data.

Table 8: Estimates From Random Distributions

(a) Panel A: Error

	Direct Mean		Distribution Mean		Distribution SD	
	(1)	(2)	(3)	(4)	(5)	(6)
Error	7.411***		0.646***		1.000***	
	(13.58)		(3.16)		(7.24)	
Abs(Error)		11.58***		4.602***		3.464***
		(24.82)		(31.80)		(38.64)
Observations	994	994	994	994	994	994

(b) Panel B: Regression of Estimates on Underlying Values

	Direct	Distribution	
	(1)	(2)	(3)
Mean	0.925***	0.459***	
	(6.00)	(8.30)	
SD			0.453***
			(11.77)
R^2	0.0350	0.0649	0.123
Observations	994	994	994

This table shows how elicited means and standard deviations vary with different underlying parameters. Panel A shows the mean forecast error (odd columns) and the absolute value of that forecast error (even columns). The first two columns show direct elicitation of the mean, the next two show distribution based estimates of the mean and the final two columns show distribution based estimates of the standard deviation. Errors are calculated relative to the sample of values displayed to participants. Panel B shows regressions of elicited expectations on underlying parameters. Mean and SD are calculated from the 100 numbers viewed for the direct method and to the implied mean from correctly filling out the bins from the distribution method.

Table 9: High Asset Value Sample: Estimates From Random Distributions

(a) Panel A: Error

	Direct Mean		Distribution Mean		Distribution SD	
	(1)	(2)	(3)	(4)	(5)	(6)
Error	10.02***		0.371		1.342***	
	(10.54)		(1.08)		(5.14)	
Abs(Error)		14.65***		5.228***		4.484***
		(17.79)		(20.49)		(25.07)
Observations	509	509	509	509	509	509

(b) Panel B: Regression of Estimates on Underlying Values

	Direct	Distribution	
	(1)	(2)	(3)
Mean	-0.105	0.227**	
	(-0.38)	(2.37)	
SD			0.259***
			(3.56)
R^2	0.000285	0.0109	0.0244
Observations	509	509	509

This table shows how elicited means and standard deviations vary with different underlying parameters for our sample of investors with a portfolio value greater than $100,000. Panel A shows the mean forecast error (odd columns) and the absolute value of that forecast error (even columns). The first two columns show direct elicitation of the mean, the next two show distribution based estimates of the mean and the final two columns show distribution based estimates of the standard deviation. Errors are calculated relative to the sample of values displayed to participants. Panel B shows regressions of elicited expectations on underlying parameters. Mean and SD are calculated from the 100 numbers viewed for the direct method and to the implied mean from correctly filling out the bins from the distribution method.

Table 10: Mathematically Sophisticated Sample: Estimates From Random Distributions

(a) Panel A: Error

	Direct Mean		Distribution Mean		Distribution SD	
	(1)	(2)	(3)	(4)	(5)	(6)
Error	4.982***		0.370		2.304***	
	(6.23)		(1.39)		(11.81)	
Abs(Error)		11.06***		4.188***		3.733***
		(16.66)		(22.16)		(26.02)
Observations	493	493	493	493	493	493

(b) Panel B: Regression of Estimates on Underlying Values

	Direct	Distribution	
	(1)	(2)	(3)
Mean	0.547**	0.422***	
	(2.46)	(6.04)	
SD			0.528***
			(9.27)
R^2	0.0121	0.0691	0.149
Observations	493	493	493

This table shows how elicited means and standard deviations vary with different underlying parameters for our sample of people with degrees in mathematics, statistics or engineering. Panel A shows the mean forecast error (odd columns) and the absolute value of that forecast error (even columns). The first two columns show direct elicitation of the mean, the next two show distribution based estimates of the mean and the final two columns show distribution based estimates of the standard deviation. Errors are calculated relative to the sample of values displayed to participants. Panel B shows regressions of elicited expectations on underlying parameters. Mean and SD are calculated from the 100 numbers viewed for the direct method and to the implied mean from correctly filling out the bins from the distribution method.

Table 11: High Statistical Knowledge Sample: Estimates From Random Distributions

(a) Panel A: Error

	Direct Mean		Distribution Mean		Distribution SD	
	(1)	(2)	(3)	(4)	(5)	(6)
Error	3.697***		0.175		1.791***	
	(3.24)		(0.47)		(5.97)	
Abs(Error)		9.713***		3.945***		3.458***
		(10.14)		(15.48)		(15.66)
Observations	213	213	213	213	213	213

(b) Panel B: Regression of Estimates on Underlying Values

	Direct	Distribution	
	(1)	(2)	(3)
Mean	0.412	0.387***	
	(1.31)	(4.11)	
SD			0.576***
			(6.56)
R^2	0.00811	0.0742	0.169
Observations	213	213	213

This table shows how elicited means and standard deviations vary with different underlying parameters for our mathematically sophisticated sample with a self rated statistical knowledge of 6 or 7 out of 7. Panel A shows the mean forecast error (odd columns) and the absolute value of that forecast error (even columns). The first two columns show direct elicitation of the mean, the next two show distribution based estimates of the mean and the final two columns show distribution based estimates of the standard deviation. Errors are calculated relative to the sample of values displayed to participants. Panel B shows regressions of elicited expectations on underlying parameters. Mean and SD are calculated from the 100 numbers viewed for the direct method and to the implied mean from correctly filling out the bins from the distribution method.

Table 12: Distribution Estimates Relative to Correct Answer

(a) Panel A: US Mean Estimates

Center Bin Country	Mean US	Zero US	Mean Japan	Zero Japan
Estimate	8.83	4.56	6.49	6.59
Correct	8.12	8.15	6.53	8.70
t-stat (value=correct)	2.46	-11.70	-0.08	-3.22
Obs	329	326	146	142

(b) Panel B: US Sigma Estimate

Center Bin Country	Mean US	Zero US	Mean Japan	Zero Japan
Estimate	18.32	18.48	23.06	22.96
Correct	17.96	18.34	19.73	18.49
t-stat (value=correct)	1.58	0.65	5.83	7.84
Obs	329	326	146	142

(c) Panel C: Japan Mean Estimate

Center Bin Country	Mean US	Zero US	Mean Japan	Zero Japan
Estimate	9.15	5.63	11.75	8.79
Correct	8.71	7.64	11.80	9.04
t-stat (value=correct)	0.81	-3.84	0.08	-0.31
Obs	139	140	147	140

(d) Panel D: Japan Sigma Estimate

Center Bin Country	Mean US	Zero US	Mean Japan	Zero Japan
Estimate	19.66	19.55	28.10	26.49
Correct	21.62	20.49	23.69	24.27
t-stat (value=correct)	-5.79	-2.76	8.02	4.03
Obs	139	140	147	140

This table shows estimates of means and standard deviations in relation to the implications to the correct answer from the bins. Panel A and B show responses from particpants who viewed US data and panels C and D show responses from those that viewed Japanese data. Panel A and C shows mean estimates and Panels B and C show standard deviation estimates. The first and third columns show estimates from bins centered at the sample average. The first two columns show bins using US data and the last two show bins using Japanese data. The top row shows the estimate. The second row shows the implied mean or standard deviation from the correct answer from the historical data. The third row shows a t-test that the estimate is equal to this correct value.

Table 13: Different Number of Bins

(a) Panel A: Mean

	5 Bin		9 Bin		13 Bin	
	(1)	(2)	(3)	(4)	(5)	(6)
Error	0.962***		-0.0136		0.464*	
	(3.08)		(-0.05)		(1.83)	
Abs(Error)		5.139***		4.529***		4.116***
		(24.29)		(21.73)		(24.36)
Observations	486	486	487	487	470	470

(b) Panel B: SD

	5 Bin		9 Bin		13 Bin	
	(1)	(2)	(3)	(4)	(5)	(6)
Error	3.315***		1.151***		-0.391**	
	(15.57)		(6.00)		(-2.18)	
Abs(Error)		4.435***		3.283***		3.014***
		(26.77)		(24.90)		(26.33)
Observations	486	486	487	487	470	470

(c) Panel C: Sensitivity

	5 Bin		9 Bin		13 Bin	
	(1)	(2)	(3)	(4)	(5)	(6)
Mean	0.571***		0.437***		0.409***	
	(7.02)		(5.84)		(6.37)	
SD		0.432***		0.453***		0.402***
		(7.12)		(8.87)		(8.18)
R^2	0.0924	0.0948	0.0658	0.140	0.0798	0.125
Observations	486	486	487	487	470	470

This table shows how elicited means and standard deviations using different numbers of bins vary with different underlying parameters. The first two columns show values from 5 bin elicitation, the next two use 9 bins and the final two columns use 13 bins. Panel A shows the mean forecast error (odd columns) and the absolute value of that error (even columns). Panel B shows the standard deviation forecast errors errors absolute value of that forecast error (odd columns) and the absolute value of that error (even columns). Panel C shows regressions coefficeients of the elicited mean (odd columns) or standard deviation (even columns) on the given parameter that was displayed. Mean and SD are calculated from the 100 numbers viewed for the direct method and to the implied mean from correctly filling out the bins from the distribution method.

Online Appendix for

Elliciting Expectaions

Samuel M. Hartzmark Abigail B. Sussman

Appendix A: Experimental Treatments

A. Baseline survey text

Below is the text of the baseline survey. Every survey respondent responded to Screen 5mean as well as randomly responding to one of 5CI, 5DistBuilder or 5 DistBin. Participants were randomized into seeing 5mean before or after the other screen 5.

Screen 1: Consent Form

Screen 2:

In this survey, we are interested in understanding how people recall and summarize information that is presented to them.

On the following pages, you will see 100 numbers. Each number represents the **stock market return** from one of the last 100 years. The returns will be presented in random order and cover all 100 years.

Screen 3: Attention Check (Order of options randomized)

In this survey, you will be shown information about _ _ _ _ _ _ _ over the past 100 years:

- ○ stock market returns
- ○ baseball statistics
- ○ weather patterns
- ○ political outcomes
- ○ none of the above
- ○ do not recall

Screen 4:

You have passed the comprehension check!

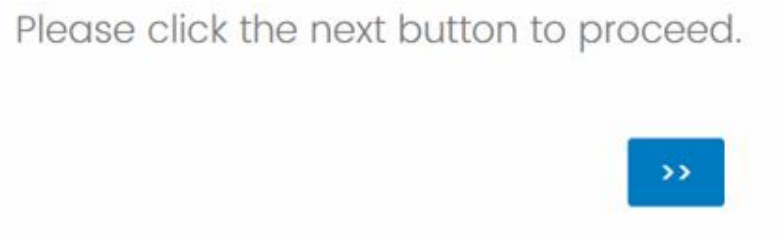
On the next pages, you will see these numbers. Each number will be presented for only half a second, so please pay attention. You will be asked questions about these numbers once they are done being displayed.

Your HIT may be rejected if we detect that you have left the current tab while these numbers were being displayed.

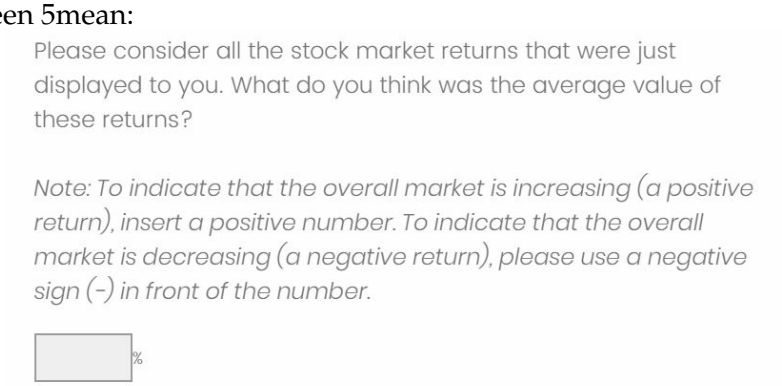
100 numbers displayed one at a time for 1/2 a second in random order. Numbers displayed as integer%. Example below:



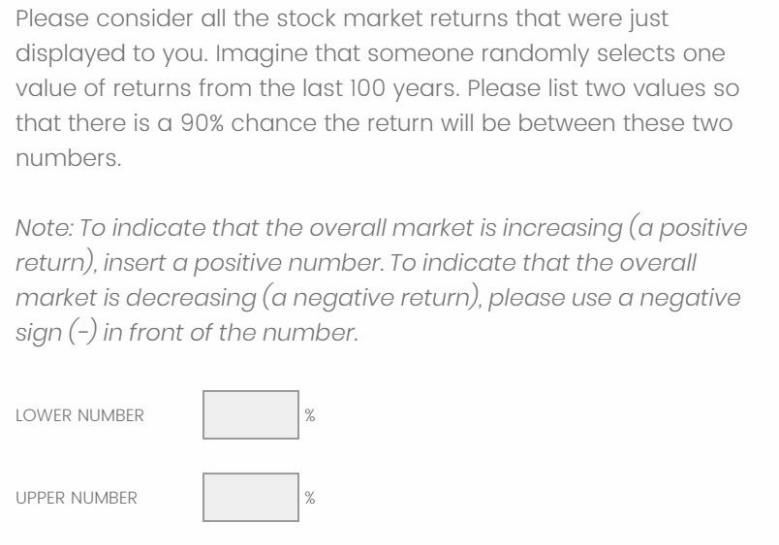
After 1/2 a second subjects saw a screen saying:



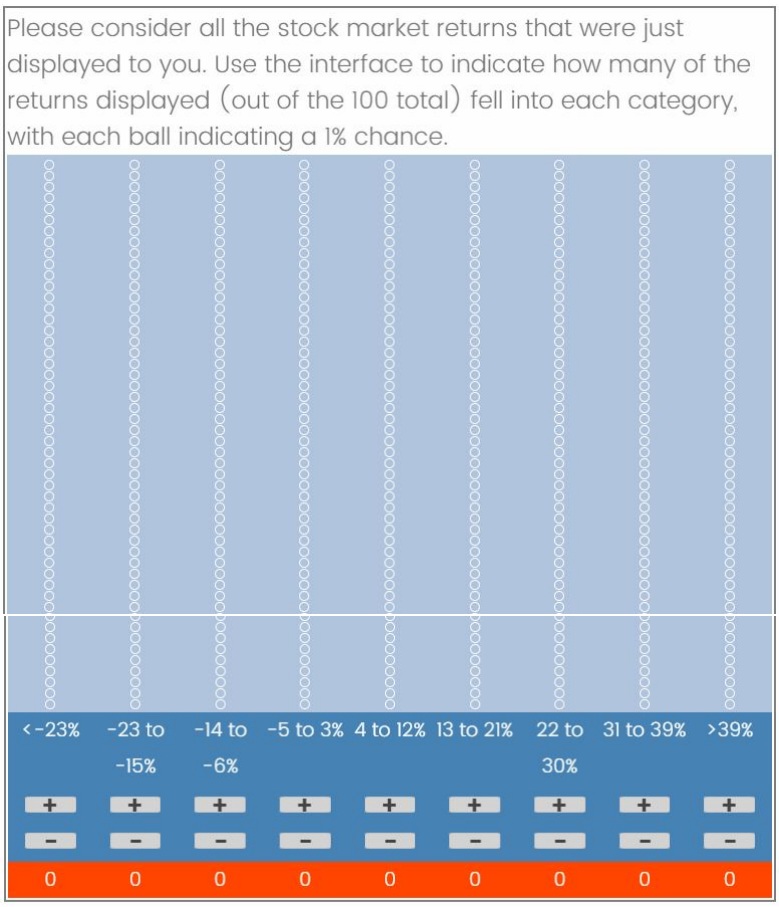
Screen 5mean:



Screen 5CI: (90% (displayed below) was randomized to 98/90/80/60)



Screen 5DistBuilder: (US Bins at mean displayed)



Screen 5DistBin: (US Bins at mean displayed)

Please consider all of the stock market returns that were just displayed to you and indicate how many of the returns displayed (out of the 100 total) fell into each category.

<-23%	0
-23 to -15%	0
-14 to -6%	0
-5 to 3%	0
4 to 12%	0
13 to 21%	0
22 to 30%	0
31 to 39%	0
>39%	0
Total	0

Demographic questions

What is your age?

What is your gender?

- ○ Male
- ○ Female
- ○ Other (insert below)

What is the highest level of education you have completed?

- ○ Less than High School
- ○ High School / GED
- ○ Some College
- ○ 2-year College Degree
- ○ 4-year College Degree
- ○ Masters Degree
- ○ Professional Degree (JD, MD)
- ○ Doctoral Degree

What is your total pretax household income last year (2022)? Please give your best estimation.

○ Less than $10,000	○ $70,000-$79,999
○ $10,000-$19,999	○ $80,000-$89,999
○ $20,000-$29,999	○ $90,000-$99,999
○ $30,000-$39,999	○ $100,000-$124,999
○ $40,000-$49,999	○ $125,000-$149,999
○ $50,000-$59,999	○ $150,000-$199,999
○ $60,000-$69,999	○ $200,000 or greater

Final Screen

Thank you for participating!

Optional: If you have any comments for the researchers, put them in the text box below. If anything in the survey was unclear, we would greatly appreciate your letting us know.

B. Numbers Only and Continuous Treatment survey text

Note: These are the screens that were different from the baseline

Screen 2:

In this survey, we are interested in understanding how people recall and summarize information that is presented to them.

On the following pages, you will see 100 numbers.

Screen 3: Attention Check (Order of options randomized)

In this survey, you will see _ _ _ numbers.

- ○ 50
- ○ 100
- ○ 150
- ○ 200
- ○ do not recall

Screen 5mean:

Please consider all the numbers that were just displayed to you. What do you think was the average value of these numbers?

Note: To indicate that the value is bigger than 0, insert a positive number. To indicate that the value is smaller than 0, please use a negative sign (-) in front of the number.

[ ] %

Screen 5DistBin:

Please consider all of the numbers that were just displayed to you and indicate how many of the numbers displayed (out of the 100 total) fell into each category.

< -23%	0
-23 to -15%	0
-14 to -6%	0
-5 to 3%	0
4 to 12%	0
13 to 21%	0
22 to 30%	0
31 to 39%	0
>39%	0
Total	0

C. Finance Sample

Note: These are the screens that were different from the general continuous treatment. The finance sample saw a version of the continuous treatment that discusses stock returns.

Screen 2:

In this survey, we are interested in understanding how people recall and summarize information that is presented to them.

On the following pages, you will see a series of **100** numbers, which we refer to as "stock market returns". These are randomly selected numbers chosen from a distribution that approximates (though is different from) the annual stock returns in the past 100 years of various large national stock markets.

Screen 4:

You have passed the comprehension check!

On the next pages, you will see these stock market returns. Each number will be presented for only half a second, so please pay attention. You will be asked questions about them once they are done being displayed.

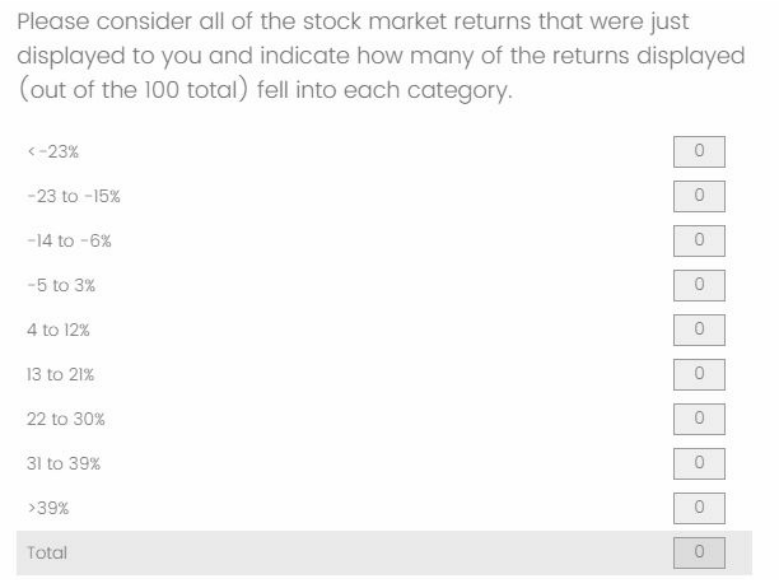
Your HIT may be rejected if we detect that you have left the current tab while these numbers were being displayed.

Screen 5mean:

Please consider all the stock market returns that were just displayed to you. What do you think was the average value of these returns? Note: To indicate that the overall market is increasing (a positive return), insert a positive number. To indicate that the overall market is decreasing (a negative return), please use a negative sign (-) in front of the number.

[ ] %

Screen 5DistBin:



Finance questions

What is the total value of your investable assets (for example, money in your 401k, mutual funds, stocks, bonds, ETFs)?

- ○ Less than $100,000
- ○ $100,001 to $250,000
- ○ $250,001 to $500,000
- ○ More than $500,000

Which of the following do you own?

- ☐ Index funds, ETFs, Mutuals funds
- ☐ Individual stocks
- ☐ Individual bonds
- ☐ None of the above

Approximately how frequently do you trade assets in your portfolio?

- ○ Less than once a year
- ○ 1-3 times per year
- ○ 4-6 times per year
- ○ 7-12 times per year
- ○ More than 12 times per year

In the last year, have you traded options?

- ○ Yes
- ○ No

D. Math Sample

Note: The math sample did the standard continuous treatment above with these extra questions at the end.

How would you rate your understanding of statistics relative to the general US population?

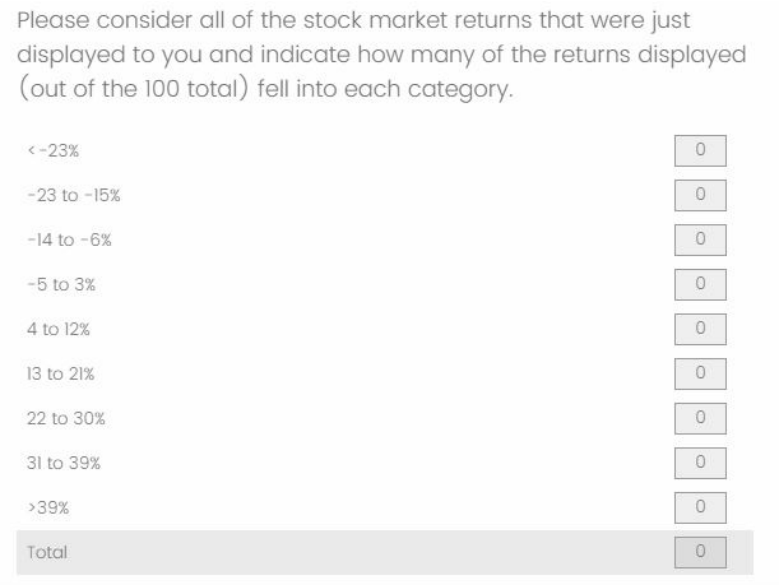
Much lower than the general US population 1	2	3	Same as the general US population 4	5	6	Much higher than the general US population 7
○	○	○	○	○	○	○

What was your college major?

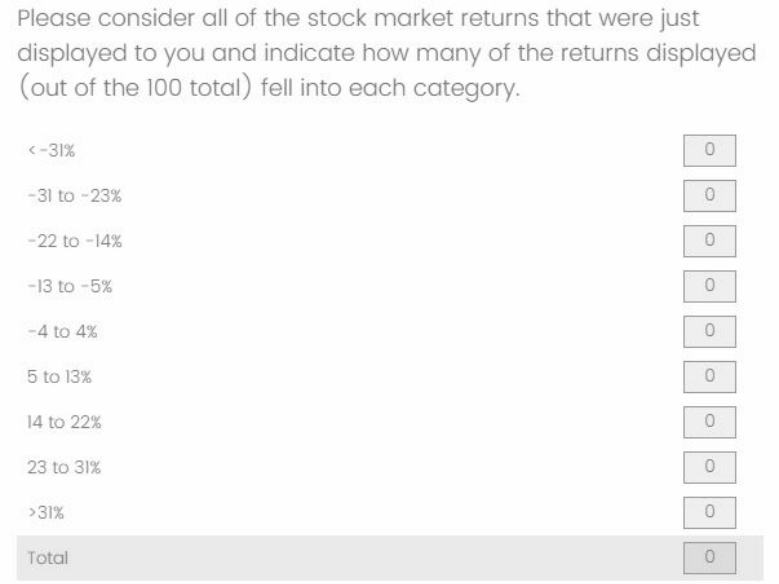
- ☐ Economics
- ☐ Engineering
- ☐ English
- ☐ History
- ☐ Math
- ☐ Physics
- ☐ Political Science
- ☐ Statistics
- ☐ Other (Please State)

E. Bin Variations

US Mean



US Zero



Japan Mean

Please consider all of the stock market returns that were just displayed to you and indicate how many of the returns displayed (out of the 100 total) fell into each category.

< -42%	0
-42 to -28%	0
-27 to -13%	0
-12 to 2%	0
3 to 17%	0
18 to 32%	0
33 to 47%	0
48 to 62%	0
>62%	0
Total	0

Japan Zero

Please consider all of the stock market returns that were just displayed to you and indicate how many of the returns displayed (out of the 100 total) fell into each category.

< -52%	0
-52 to -38%	0
-37 to -23%	0
-22 to -8%	0
-7 to 7%	0
8 to 22%	0
23 to 37%	0
38 to 52%	0
>52%	0
Total	0

US Mean: 5-bin

Please consider all of the numbers that were just displayed to you and indicate how many of the numbers displayed (out of the 100 total) fell into each category.

< -23%	0
-23 to -3%	0
-2 to 18%	0
19 to 39%	0
>39%	0
Total	0

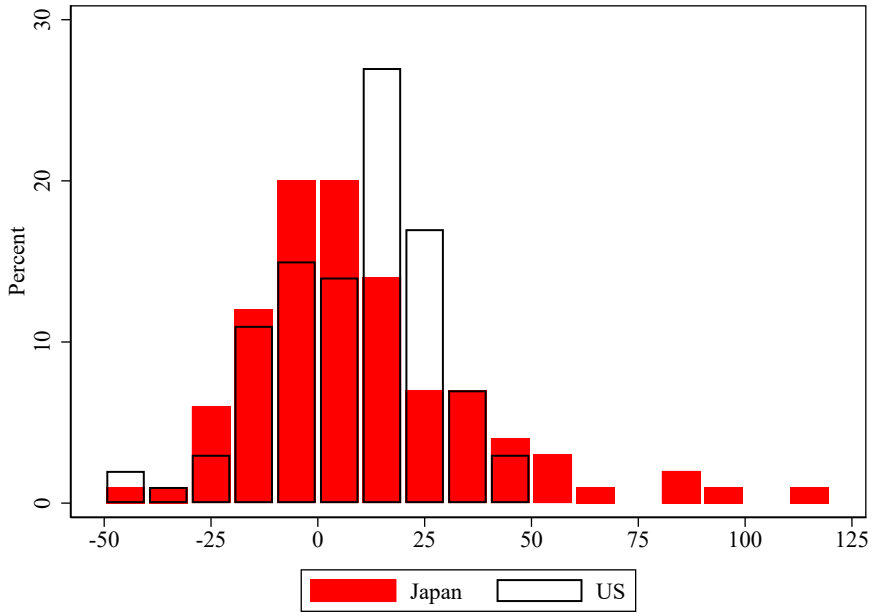
US Mean: 13-bin

< -19%	0
-19 to -15%	0
-14 to -10%	0
-9 to -5%	0
-4 to 0%	0
1 to 5%	0
6 to 10%	0
11 to 15%	0
16 to 20%	0
21 to 25%	0
26 to 30%	0
31 to 35%	0
>35%	0
Total	0

Appendix B: Further Analysis

Figure A1: Historical Distribution of Returns in US and Japan

(a) Panel A: Histogram



This graphs show the distributions of the returns viewed by subjects in the US and Japan.

Table A1: Differences Between Distribution Builder and Simple Bins

(a) Panel A: Distribution Builder vs Simple Bins

	Mean		SD	
	(1)	(2)	(3)	(4)
Dist Builder	-0.424	-0.329	-1.593***	-1.153***
	(-1.10)	(-0.88)	(-4.76)	(-3.95)
Constant	7.629***	7.596***	21.75***	21.59***
	(33.64)	(34.55)	(110.50)	(126.23)
Treatment FE	No	Yes	No	Yes
R^2	0.000799	0.0647	0.0148	0.258
Observations	1509	1509	1509	1509

This table shows how elicited means and standard deviations vary when elicited with the distribution builder graphical interface as opposed to filling in numbers into bins. Means (Columns 1 and 2) or standard deviations (Columns 3 and 4) are regressed on a constant and a dummy variable for whether the estimate comes from the distribution builder method and is zero if it comes from the simple bin method.