Expectations: Point-Estimates, Probability Distributions, Confidence, and Forecasts[*]

David Rothschild
Microsoft Research
Applied Statistics Center at Columbia
David@ResearchDMR.com
www.ResearchDMR.com

Abstract

In this paper I test a new graphical, interactive interface that captures both "best estimate" point-estimates and probability distributions from non-experts. When supplementing an expectation, a standard data point is directly stated confidence of the respondent or a confidence range. In contrast to those data points, my method induces the respondents to reveal a level of precision, and there is a sizable and statically significant positive relationship between the respondents' revealed precision and the accuracy of their individual-level expectations. Beyond creating a more meaningful individual-level estimates, researchers can use this positive correlation between precision and accuracy to create precision-weighted aggregated forecasts that are more accurate than the standard "consensus forecasts". Varying financial incentives does not affect these findings.

This draft: September 20, 2012
Latest draft: www.ResearchDMR.com/RothschildConfidence
Keywords: Polling, information aggregation, belief heterogeneity
JEL codes: C52, C53, C83, D03, D8

[*] Thank you to Alex Gelber, Daniel G. Goldstein, Andrew Leonard, Katherine Milkman, Neil Malhotra, James Rothschild Jr., Uri Simonsohn, Jeremy Tobacman, Justin Wolfers, and the two anonymous reviewers for the Workshop on Social Computing and User Generated Content. Further thanks to the Wharton Risk Center's Ackoff grant program that helped me fund this project (they had no direct involvement research).

1. Introduction

This paper tests a new graphical, interactive interface that eases the creation of probability distributions by non-expert respondents. The method captures a "best-estimate" point-estimate and then a full probability distribution from non-experts on upcoming events. Specifically, the interface creates a series of bins, around the respondent's point-estimate, and allows the respondent to distribute 100 probabilities into the bins. These 100 probabilities form a probability distribution and I determine a revealed precision of the respondent as the inverse of the width of various relevant ranges; the smaller the width of a range, the more precise the respondent.

The responses demonstrate a slight over-precision. Throughout this paper I compare my method of deriving confidence from the probability distributions to the two most attractive alternatives: a simple four point scale of stated confidence and a standard, transparent method of collecting a confidence range around a point-estimate. Where comparable, data from the probability distributions derived in my method and confidence range provide a similar slight over-precision; this over-precision is consistent with the literature (Soll and Klayman, 2004; Teigen and Jorgensen, 2005; and Speirs-Bridge et al., 2010).[1]

There is a sizable and statistically significant correlation between precision, estimated as the inverse of the variance, derived from the respondents' probability distributions and the accuracy of the corresponding expectations. This correlation is true both within questions (i.e., between respondents) and within respondents (i.e., between questions). This correlation is weaker and/or non-significant in the two simplified methods. Engleberg et al. (2009) determines that the distribution of probabilities, created by experts, widens as the event horizon lengthens. Yet, that paper does not address the correlation with the size of a distribution and the accuracy for a given event. Other papers conclude that tighter confidences ranges correlate with increased over-precision (i.e., at best, the width of ranges is uncorrelated with accuracy for non-

[1] Over/under-precision is the percentage of answers that come to fruition within a range; a set of perfectly calibrated 80% ranges would have 80% of the answers occur within the range. Stated confidence cannot provide a measure and the confidence range provides only the one range it asks, where the probability distribution reveals an infinite set of confidence ranges.

experts).[2] There is a large literature on confidence and precision, and buried in some papers are positive correlations between some measure of accuracy and some measure of precision or confidence, but this outcome is neither common nor publicized when it is occurs.[3]

On an individual-level I am gathering more useful data about expectations from the respondents. Psychology, political science, and marketing are just a few literatures where there are standard survey practices that ask respondents to state their: confidence, likelihood, or agreement with their previously stated estimations. Researchers use this data to qualify or moderate the main response or as the main data point itself; either way, my method of revealing this information can provide more meaningful data.

Precision can be a weight in the creation of aggregated forecasts of the level of the outcome; precision-weighted forecasts are more accurate than forecasts from the point-estimates alone. There is consensus in the literature that aggregating expectations creates more accurate forecasts, on average, than picking individual-level expectations; Pennock and Reeves (2007) demonstrates that with a self-selected group of respondents making NFL predictions only 6 of the 2231 players are more accurate than the collective average. Further, there are numerous studies showing that the simplest aggregation techniques are the most efficient (Bates and Granger, 1969; Stock and Watson, 2004; Smith and Wallis, 2009). The simplest weight considered in the literature, other than an average, is the inverse of the mean square error; this occasionally provides a smaller error than simple averages. Non-experts answer the questions in this paper concurrently, so there is no opportunity for historically derived weights; yet, if the inverse of the variance of their probability distributions is well calibrated, it could serve as a proxy for mean square error.[4]

[2] Kuklinski (2000) concludes that for objective political questions confidence can actually be negatively correlated with accuracy.

[3] For example Dunning et al. (1990) has a small positive correlation in the paper, but does not even mention it in the abstract. The focus, as usual, is on the precision of the respondents.

[4] Without historical data, I cannot update on risk profile as in Chen et al (2003) or more complex Bayesian methods like those cataloged in Clemen (1989) or suggested in Clemen and Winkler (1993).

Prediction markets weigh their individual-level data by how much money people are willing to gamble, a proxy for confidence or precision; this second finding of the paper illustrates the benefits of a new method of directly weighing individual-level polling data by precision. There is growing consensus that prediction markets aggregate individual-level expectations into more accurate forecasts than standard polling methods. Rothschild (2009) confirms the accuracy of prediction market-based forecasts and outlines some of the differences between them and standard polls as methods of capturing and aggregating individual-level responses: the nature of their sample of respondents, the question asked to the respondents, the aggregation method of the individual-level data, and the incentives for the respondents. Rothschild and Wolfers (2012) addresses the question asked of the respondents; that paper concludes that eliciting expectations, as prediction markets do, captures more information from the respondent than the questions asked in standard polls. This paper addresses another of the differences between prediction markets and polls, the weighting of the individual-level data. Traditional methods of gathering information from individual respondents fail to capture large quantities of the respondents' relevant information and the quality of the information they reveal; this new method gathers more relevant information and provides a new rubric for weighing its quality.

In this paper I test a graphical, interactive interface to advance the ability of researchers to collect and utilize individual-level expectations. I capture both point-estimates and probability distributions from non-expert respondents. On an individual-level, the results provide insight into expectations, revealing the nature of their: accuracy, precision, and calibration of probabilities. Better understanding of non-expert expectations will allow researchers to learn more about the absorption and transformation of information. On an aggregate-level, researchers can use these expectations as individual-level data for aggregated forecasts and, in the future, to understand heterogeneity in revealed behavior under uncertainty. Better forecasts help researchers connect shocks with changes in the underlying values of the world and investors make more efficient use of their time and money.

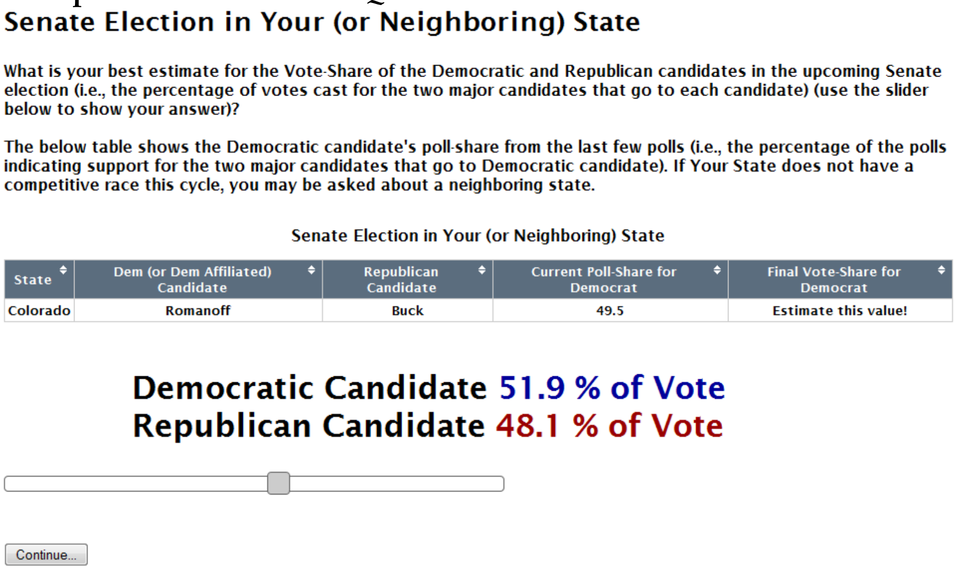
2. Method

I build on the most recent methods of surveying expectations to create a graphical, interactive interface that gathers expectations: point-estimates and probability distributions. One influence on my method is Delavande and Rohwedder (2008), which asks respondents for a point-estimate and then uses a visual screen that asks the respondents to distribute 20 balls (each representing a 5% probability of the outcome) in a series of bins that signify possible value-ranges of the outcome. I enhance their method with lessons from the literature involving the generation of confidence ranges around point-estimates. A few key innovations of my method: it distributes probabilities as small as 1% into upwards of nine bins of value-ranges, forces the respondent to consider the question from multiple angles, expands the ranges to if the respondent is placing more than 50% of her probability in one range, and uses new graphical tools to efficiently clarify the procedure for the respondent. Further, after I collect these expectations (i.e., the point-estimate and probability distribution) I probe other characteristics of the respondents that may be correlated with biases or varying information levels.[5]

The first piece of data I recover from the respondent for any specific question is a "best estimate" point-estimate. While maintaining overall consistency regardless of the specifics of the questions, the interface is adjusted enough to ensure that the respondents provide valid responses. The appropriate interface allows the respondent to understand exactly what the question asks, without taking up too much time and effort. Further, it takes measures to avoid any anchoring or suggestive examples for the respondent; it provides enough background information to ensure that the user can apply her information rationally without anchoring her to a specific estimate. As an example, the graphical design for vote share questions uses a slider that shows the two-party vote share for both candidates, which makes it easy to understand the meaning of two-party vote share.

[5] Delavande and Rohwedder (2008) tested their method against verbal methods of gathering probability distributions on questions involving potential returns on social security when the respondents reach the eligible age; the paper has no calibrated outcomes with which to test the efficiency of the responses.

Example of Point-Estimate Question:



Senate Election in Your (or Neighboring) State

What is your best estimate for the Vote-Share of the Democratic and Republican candidates in the upcoming Senate election (i.e., the percentage of votes cast for the two major candidates that go to each candidate) (use the slider below to show your answer)?

The below table shows the Democratic candidate's poll-share from the last few polls (i.e., the percentage of the polls indicating support for the two major candidates that go to Democratic candidate). If Your State does not have a competitive race this cycle, you may be asked about a neighboring state.

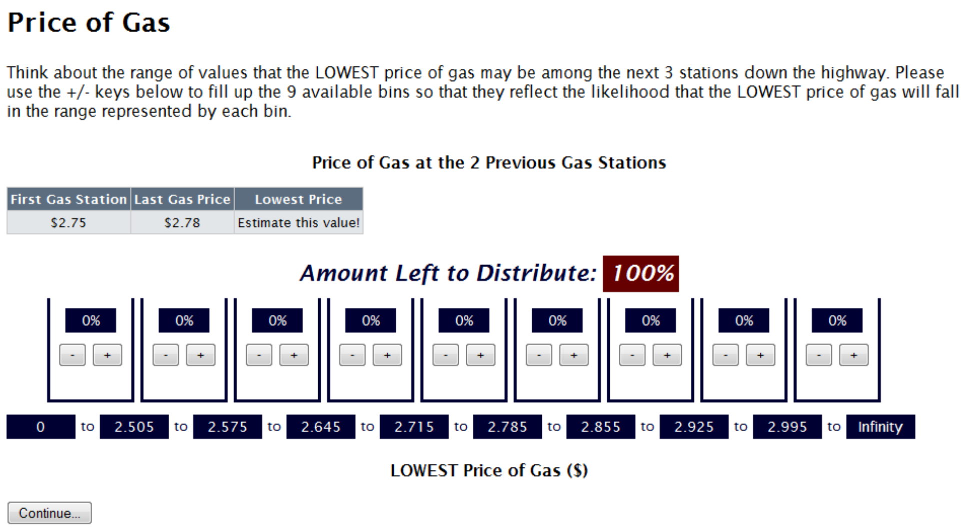
Senate Election in Your (or Neighboring) State

State	Dem (or Dem Affiliated) Candidate	Republican Candidate	Current Poll-Share for Democrat	Final Vote-Share for Democrat
Colorado	Romanoff	Buck	49.5	Estimate this value!

Democratic Candidate 51.9 % of Vote
Republican Candidate 48.1 % of Vote

The second piece of data elicited is a probability distribution. A non-interactive poll would ask a series of questions to the respondent to create a distribution of probabilities or it could provide a series of pre-set value ranges for the respondent to distribute her probabilities. My method creates a series of value ranges centered on their point-estimate. The respondent distributes “100 likelihoods” into the 9 different bins (each representing a value range). Further, if the ranges prove too wide and the respondent places more than 50% of the probability in one bin, the program creates a new set of bins within the aforementioned range. The respondent must answer fully and cannot be internally inconsistent (i.e., the overall probability must equal 100% and there cannot be contradicting probabilities), thus all responses can be used. By adjusting the bins’ ranges so that they are centered on the point-estimate, the method eliminates irrelevant thresholds and anchoring derived from pre-determined thresholds. Respondents can answer more questions faster than the traditional method of gathering distributions, because instead of a series of questions, there is just one question to get a distribution. Consistent with the literature, I assume that probabilities are distributed uniformly within a bin.

Example of Probability Distribution Question:



Price of Gas

Think about the range of values that the LOWEST price of gas may be among the next 3 stations down the highway. Please use the +/- keys below to fill up the 9 available bins so that they reflect the likelihood that the LOWEST price of gas will fall in the range represented by each bin.

Price of Gas at the 2 Previous Gas Stations

First Gas Station	Last Gas Price	Lowest Price
$2.75	$2.78	Estimate this value!

Amount Left to Distribute: 100%

0%
0%
0%
0%
0%
0%
0%
0%
0%
- +
0 to 2.505 to 2.575 to 2.645 to 2.715 to 2.785 to 2.855 to 2.925 to 2.995 to Infinity
LOWEST Price of Gas ($)
Continue...

There are two additional sources of data from the respondents. First, before the first task, the respondent is asked a series of questions on observable characteristics. Second, within each question, the two main questions are followed by a question which examines the respondents' personal partiality related to the question. As an example, if the questions were about movies, I would ask the respondent about their intention to see the movie.

The data for this paper was gathered in two rounds of studies. In both of the studies there were five categories of questions, with each category having either nine or ten unique questions. Each respondent was in just one of the studies and answered just one question from each category. Everything is randomized so that each respondent sees the categories in a random order, is randomly assigned to a unique question in a category, and, if there is variation in the information level for a given question that is also randomly assigned. Both studies are a mixture of two groups of people, the Wharton Behavioral Lab (mainly students and staff) and respondents from around the United States with Mechanical Turk. [6]

Study I concentrates on comparing my method to the methods of stated confidence and confidence intervals. Each respondent answered one question in the following five categories: calories, concert ticket prices, gas prices, movie receipts, and unemployment rate (for question details, see Appendix A). Half of the respondents were randomly assigned to create the full probably distribution and half were asked state their confidence and create a confidence

[6] Paolacci, et al (2010) explains Mechanical Turk in detail.

interval. The stated confidence is recovered with the standard polling question: "How confident are you of your answer?" with a drop down menu of four choices: very confident, somewhat confident, not very confident, and not at all confident. The confidence range is recovered in the most efficient method, asking the respondent the combination of: "I am 90% sure answer is greater than ________" and then "I am 90% sure the answer is less than ________". Soll and Klayman (2004) demonstrates that respondents reduce overconfidence when they are asked to formulate an upper and lower limit in separate steps. The theory is also demonstrated in Vul and Pashler (2008); that paper shows that when respondents are forced to think about the same question in two different ways, they add new information that is not apparent when they just consider the question from one angle.[7] This method of creating the confidence range exists between having the respondents directly state it and having them unwittingly reveal it. 129 respondents answered the stated confidence and confidence range questions, while 120 respondents answered the probability distribution questions.

Study II concentrates on comparing my method under different financial incentives. Each respondent answered one question in each of the following five categories: calories, gas prices, and unemployment rates (as in Study I), and then voter turnout and Senate election results (for question details, see Appendix A).[8] While the first three categories are the same, all of the questions are new. Half of the respondents are paid with standard flat fee and half of the respondents are paid a flat fee and then incentivized with a bonus; the five respondents with

[7] Teigen and Jorgensen (2005) concludes that overconfidence is decreased when respondents assign probabilities to confidence ranges, rather than confidence ranges for a set probability and similarly Speirs-Bridge et al. (2010) argues that it is best to ask the respondent to re-calibrate his own range (i.e., show them their 80% range they just created and ask them how wide that really is). The full probability distribution method requests the respondent to input their own probability to set ranges, rather than input a range to a set probability, but that is not possible with the confidence ranges. I do not want to make ungrounded assumptions on how to translate a 60% range into an 80% range in order to create uniform ranges. And, without uniform ranges, there is no method of comparing the respondents' on their calibration, as they would all have different size ranges.

[8] The voter turnout and Senate election results are conducted with the respondents' home states, so due to lopsided draws, are not included in this paper, but the basis for further research. The questions replaced movies and concerts, because Study II was done in the early fall, where the time frame was too short for movies and concerts are not frequent enough.

the lowest weighted square error over all of their responses were given the bonuses.[9] Since Study II is unbalanced in its categories, I only use it in the full data when I am making comparisons concerning the effect of the incentives. 103 respondents were non-incentivized, while 99 respondents were incentivized.

3. Estimation/Results

Between-respondent disagreement is much larger than within-respondent uncertainty; this demonstrates both the over-precision of the responses and issues involving the accuracy of the stated point-estimates. This comparison is illustrated in the coefficients of variation in Table 1, where between-respondent disagreement is the coefficient of variation of the stated point-estimates for a unique question and within-respondent uncertainty is the average coefficient of variation of the individual-level probability distributions for that unique question. Within any given category, coefficient of variation is a consistent measure of the variability of the responses. Gurkaynak and Wolfers (2006) studied an economic derivative market and show that between-respondent disagreement of point-estimates (submitted by experts) is less dispersed than within-forecast uncertainty, illustrated by an efficient market. One reason that the between-respondent disagreement is relatively larger in this paper is that I am not studying the uncertainty in an efficient market, but within individuals, where the variance of the individual-level probability distributions reflects individual uncertainty of the outcome. In two paragraphs I will demonstrate that these probability distributions are too narrow (i.e., over-precise). The second reason is because the respondents are providing point-estimates that extend all over their distributions, not just the mean or median. The standard deviation of the point-estimates for one question should be similar to the standard deviation of the most likely outcome perceived by the respondents of that question. Yet, for some categories of questions the average absolute log difference between the mean and median of a respondent's distribution and their

[9] I note in parenthesis in the directions the bonus will go to "the most accurate distributions", because very few respondents are going to know what a mean square error is, or what type of response minimizes it. This inability to comprehend the scoring rule is a problem for incentives noted in Artinger (2010). The goal of the score rule was to be intuitive to the user and cover all five responses.

point-estimate approaches 10%.[10] Table 4 provides further insight into the accuracy of the stated point-estimates.

	Study I		Study II Incentivized / non-Incentivized	
Category	**Uncertainty**	**Disagreement**	**Uncertainty**	**Disagreement**
Calories	0.221	0.373	0.175/0.183	0.392
Concert Tickets	0.222	0.384	-	-
Gas Prices	0.026	0.015	0.027/0.026	0.027
Movie Receipts	0.314	0.549	-	-
Unemployment	0.013	0.039	0.018/0.018	0.095

Note: Study I is 120 respondents. Study II is 103 respondents non-incentivized and 99 respondents incentivized. Coefficient of variation for uncertainty is (standard deviation)/(mean of distribution) and for disagreement is (standard deviation)/(mean of point-estimates).

Table 1: Coefficients of variation of individual-level probability distributions and coefficients of variation of point-estimates

The data from Study II shows that the alignment of the incentives has a negligible influence the individual-level distributions. The bonus pay rewarded respondents extra for well calibrated distributions. The mean (median) respondent did spend 1.5 (1.1) more minutes answering their five questions, although there were slightly longer directions for these respondents included in this time. Yet, Table 1 demonstrates that the coefficients of variation are very similar, regardless of incentives.[11] Regardless of the outcome, incentives are not ideal for this project, so it is comforting they have a negligible impact on the responses. First, if the researcher outlines the type of response that maximizes the payout rule, she is manipulating the response, but, in this project, I want the response to be whatever the respondent thinks is the "best estimate" not what I define as the "best estimate". Further, in future papers in this project where I connect the expectations to decisions, I want the true expectations, not expectations created to fulfill my scoring rule. [12] Not stating any goals for best estimate does not guarantee

[10] While very few point-estimates that occur on the tails of the probability distributions, for calories, concert tickets, and movie receipts the mean or median is larger than the point-estimate by a statistically significant amount.

[11] This holds for every major variables discussed in this paper.

[12] There are other contexts where I would want to get the optimal information for a forecast, but in this paper I am as interested in learning about individual-level expectations as I am in creating aggregate-level forecasts.

that the users provide their best estimate, but it is logically more likely than explicitly outlining an incentive compatible definition of best estimate. Second, non-payment or flat fees are standard in polling and I want the results of this project to be relevant.

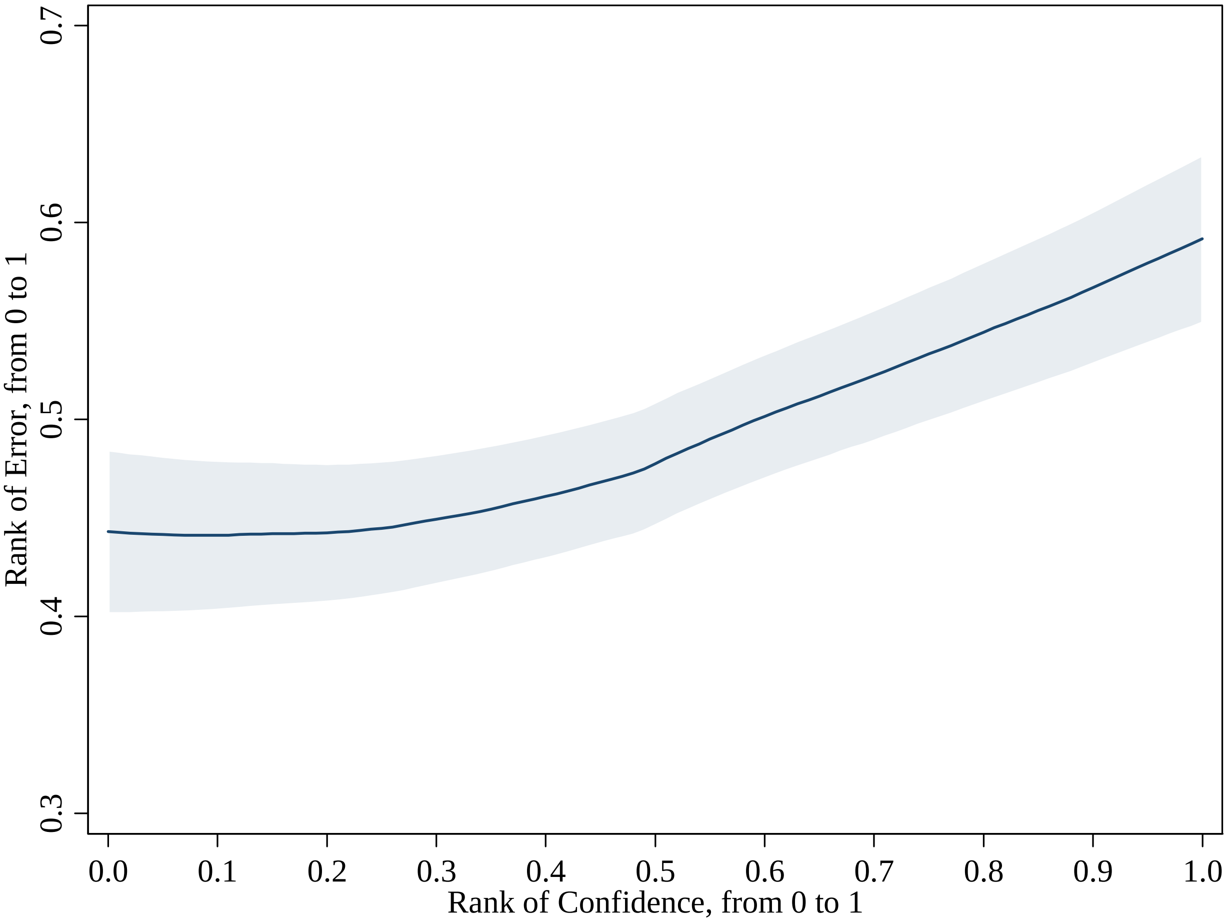
The calibration of certainty for the individual-level probability distribution is very similar to the confidence ranges (where it is comparable); both are slightly over-precise. Utilizing the data from Study I, the answer lies within the 80% confidence range 60% of the time and within the middle 80% range of the probability distribution 58% of the time, a statistically insignificant difference. These are both in line with most comparable studies, where non-experts were asked simple knowledge questions (e.g., the ranking of college or the winning % of NBA teams).[13] The probability distribution forces the respondent to consider the answer for more time, but the confidence range implores them to consider the question from more angles. This systematic over-precision is not a problem for this paper, as I care more about the relationship between precision or confidence and accuracy.[14]

Precision derived from the probability distributions correlates positively with accuracy; this correlation is more sizable than the correlation from the confidence ranges, and the correlation from the stated confidence is positive, but statistically insignificant. In order to make comparisons across the different types of data and the different categories, I simplify the data. Within each question I rank the responses from the least to most precise according to their standard deviation (for probability distribution), width of confidence range, or stated confidence from 0 to 1; this determines the responses position relative to all of the responses to its unique question. The smallest standard deviation, narrowest confidence range, or most confident answer is ranked 0, where the highest, widest, and least confident is ranked 1. I do the same for the error of the point-estimate. Figure 1 illustrates the relationship between confidence and accuracy in the probability distribution data. Relatively low confidence correlates with a lower than average, average rank of error, and relatively high confidence correlates with a higher than average, average rank of error. Table 2 shows that the correlations are positive and

[13] Soll and Klayman (2004), Teigen and Jorgensen (2005), and Speirs-Bridge et al. (2010).

[14] Again, in future work the expectations will be used to inform decision making, so I prefer authentic representation of the expectations relative to forcing more accurately calibrated confidence ranges.

significant within unique questions for both the probability distribution and confidence range, but the correlation is nearly twice as large for the probability distribution. The correlation is positive, but not significant, for stated confidence.



Note: Local linear regression estimates, using Epanechnikov kernel and rule-of-thumb bandwidth. Shaded area shows 95% confidence interval.

Figure 1: Correlations between precision or confidence derived from probability distributions and accuracy of point-estimate

	Stated Confidence	**Confidence Range**	**Probability Distribution**	**R²**
$\boldsymbol{Rank(Error) =}$ $\boldsymbol{\alpha + \beta * Rank(\sigma)}$ **OLS** **(Within Question)**	0.035 (0.038)	-	-	0.000
	-	0.151*** (0.040)	-	0.023
	0.006 (0.038)	0.150*** (0.041)	-	0.023
	-	-	0.231*** (0.040)	0.053
$\boldsymbol{Rank(Error) =}$ $\boldsymbol{\alpha + \beta * Rank(\sigma)}$ **Fixed-Effect** **(Within Respondent)**	0.103** (0.050)	-	-	0.001
	-	0.233*** (0.051)	-	0.023
	0.070 (0.050)	0.222*** (0.052)	-	0.022
	-	-	0.260*** (0.052)	0.053

Note: ***, **, and * denote statistically significant coefficients at the 1%, 5%, and 10% level, respectively. (Standard errors in parentheses). The errors and standard deviations are normalized by their rank within the unique question. The stated confidence and confidence range questions were answered by 129 respondents and the probability distribution by 120. There are a total of 48 unique questions accross 5 categories; each respondent answered 5 questions, one in each category.

Table 2: Correlations between precision or confidence and accuracy of point-estimate

Precision is not only correlated with accuracy within questions, but within respondents. Respondents adjust their precision for given questions, relative to the other respondents. It is relatively easy for a respondent to shift her confidence between questions when she is using a confidence scale, as the values remain constant between questions. Yet, for a confidence range or probability distribution, the relative precision is uncertain, as average range sizes and standard deviations vary widely between questions categories. This inability to consciously gauge relative precision makes it difficult for respondents to manipulate their response for any other incentives than their truthful response.

The wisdom of the crowd is working; the mean or median of the question's responses is more accurate than a random respondent for the majority of individual responses and has a significantly smaller error. This is shown in Table 3. For example, in Study I, just 24.3% of respondents' stated point-estimates are more accurate than the median point-estimate of the respondents. Further, the error for the median is less than that of the mean, on average and for the vast majority of questions. It is likely that the outliers are bigger issue in non-expert point-estimates than in expert point-estimates; either way, the median of the point-estimates is a standard method to use for aggregating point-estimates into forecast.[15]

	Study I	**Study II**
Categories	5	3
Questions per Category	9.6	10
Observations per Question	25.8	20.1
% of Individual-Level Point-Estimate Absolute Errors < Mean Point-Estimate of Question Absolute Errors	36.7 %	38.8 %
% of Individual-Level Point-Estimate Absolute Errors < Median Point-Estimate of Question Absolute Errors	24.3 %	27.9 %

Note: Point-estimates are all recorded prior to the probability distributions. Study I is randomized between probability distribution method and confidence questions, with 249 respondents. Study II is randomized between flat pay and incentive compatible pay for probability distribution method, with 202 respondents.

Table 3: Individual-level point-estimates

Point-estimates derived from the probability distributions are more accurate than the stated point-estimate. I test three simple point-estimates from the probability distributions: mean, median, and mode. On the top of Table 4 I run a Fair-Shiller (1989 and 1990) regression, which includes the point-estimate and the probability distribution's point-estimates, with fixed-effects by category of question. When compared directly with the stated point-estimates, the mean and the median of the probability distribution are both significant at the 10% level where the stated point-estimate is not significant, and the coefficients for the mean and median are substantially larger than the stated point-estimates. I cannot rule out that the mode provides no information that is not in the stated point-estimate. By switching to OLS with no constant and constraining the coefficients to sum to 1, I can determine the optimal weights of the different variables if I was forced to put them together for a best estimate. Again, compared directly, the mean and the median are both significant and the stated point-estimate is not. There are two

[15] Galton (1907) recommends the median for non-experts guessing the weight of a cow (a point-estimate) and this is the inspiration shown repeatedly in Surowiecki's *Wisdom of the Crowd*. Engleberg, et al (2009) also uses the median for GDP and inflation with experts.

plausible explanations for this finding. First, I cannot rotate the order of the stated point-estimate and the probability distribution, thus the respondents may be making a more accurate estimate in the probability distribution versus the stated point-estimate, because it is their second chance to consider the question. Second, the respondents may be ignoring long asymmetric tails of the probability distribution when they state their point-estimates, to the detriment of the accuracy of their stated point-estimate.

$\boldsymbol{DistVar}$ **in regression:**	**Mean**	**Median**	**Mode**	**Point-Estimate**
$\boldsymbol{ans = \alpha + \beta_1 DistVar + \beta_2 PointEst}$	0.311*	-	-	0.187
	(0.181)			(0.179)
	-	0.287*	-	0.210
		(0.170)		(0.168)
	-	-	0.018	0.472***
			(0.142)	(0.147)
	-0.310	0.824	-0.314	0.307
	(0.824)	(0.840)	(0.221)	(0.198)
$\boldsymbol{ans = \beta DistVar + (1-\beta) PointEst}$	67.8%***	-	-	32.2%
	(22.3)			(22.3)
	-	61.9%***	-	38.1%*
		(0.211)		(0.211)
	-	-	-0.228	1.228***
			(0.179)	(0.179)
	-1.423	2.872***	-1.211***	0.763***
	(0.962)	(1.020)	(0.267)	(0.241)

Note: ***, **, and * denote statistically significant coefficients at the 1%, 5%, and 10% level, respectively. (Standard errors in parentheses). There are fixed-effects by category in top regression. If I run separate OLS regressions with a constant on each category: the mean has a much larger coefficient (and more significance) in 3 categories, the point-estimate in 1 category, and similar in 1 category. If I run a regression for each expectation type by itself, the mean has the highest R^2. There are total of 590 observations in the five categories.

Table 4: Comparing the point accuracy of the individual-level expectations

Weighing the point-estimates by the precision from the probability distributions produces more accurate forecasts of the outcome level than other standard methods. Table 3 illustrates how without a probability distribution the most accurate consensus forecast of the outcome level is the median of the point-estimates for any question. Table 4 shows that on an individual-level, the mean of the probability distribution is the most informative of its point-estimates (and more informative than the median of the stated point-estimates). Following the literature's use of the inverse of the mean square error, I use the simplest and most transparent comparable method of aggregating the means of the probability distributions to create a consensus forecast; within a unique question, I weigh each mean by the inverse of its standard deviation squared (i.e., the variance):

$$w_i = \frac{{}^{1}/_{\sigma_i^2}}{\sum_{j=1}^{n} {}^{1}/_{\sigma_i^2}}$$

Any response, i, has the weight of ${}^{1}/_{\sigma_i^2}$ divided by the sum of all of the inverse variances of the responses to its unique question. This method is efficient only if the responses are efficient, but I have already shown they are overconfident. Yet, this is the most transparent and universal method available, and it does provide a more accurate answer.

Illustrated in Table 5, the confidence-weighted forecasts have more weight and/or significance more often than the median of the point-estimates. There is certainly fluctuation between the categories, but the precision-weighted mean demonstrates itself to be meaningful in relation to the standard method.

Category	**Weight**	**Median of Point-Estimate**	**Precision-Weighted Mean**	**Median of Point-Estimate**	**Precision-Weighted Mean**
		$\boldsymbol{ans = \alpha + \beta_1 PointEst + \beta_2 ConEst}$		$\boldsymbol{ans = \beta PointEst + (1 - \beta) ConEst}$	
Calories	${}^{1}/_{\sigma_i^2}$	0.059 (0.286)	1.146*** (0.281)	0.052 (0.245)	0.948*** (0.245)
Concert Tickets	${}^{1}/_{\sigma_i^2}$	0.730 (0.822)	0.282 (0.677)	0.390 (0.564)	0.610 (0.564)
Gas Prices	${}^{1}/_{\sigma_i^2}$	-0.315 (0.398)	-0.021 (0.425)	-0.405 (1.133)	1.405 (1.133)
Movie Receipts	${}^{1}/_{\sigma_i^2}$	0.805** (0.319)	-0.791* (0.348)	0.458 (0.453)	0.542 (0.453)
Unemployment	${}^{1}/_{\sigma_i^2}$	-1.052 (1.786)	2.097 (1.808)	-0.480 (1.553)	1.480 (1.553)

		$ans = \alpha + \boldsymbol{\beta_1} PointEst + \boldsymbol{\beta_2} ConEst$		$ans = \boldsymbol{\beta} PointEst + (\mathbf{1} - \boldsymbol{\beta}) ConEst$	
Calories	$1/\sigma_i^{1.5}$	0.007	1.186***	-0.020	1.020***
		(0.286)	(0.279)	(0.246)	(0.246)
Concert Tickets	$1/\sigma_i^{1.5}$	0.726	0.303	0.272	0.728
		(1.064)	(0.940)	(0.861)	(0.861)
Gas Prices	$1/\sigma_i^{0.3}$	0.342	-0.633	0.345	0.655
		(0.846)	(0.823)	(2.463)	(2.463)
Movie Receipts	$1/\sigma_i^{0.1}$	-0.499	0.792	1.947**	-0.947
		(0.829)	(0.773)	(0.729)	(0.729)
Unemployment	$1/\sigma_i^{2.7}$	-1.017	2.055	-0.738	1.738
		(1.653)	(1.666)	(1.515)	(1.515)

Note: ***, **, and * denote statistically significant coefficients at the 1%, 5%, and 10% level, respectively. (Standard errors in parentheses). There are 48 question total: 10 for calories, 10 for gas prices, and 10 for unemployment, 9 for concert tickets, and 9 for movie receipts.

Table 5: Comparing the point accuracy of the most promising forecasts

In the second part of Table 5 I explore more efficient weighting for the precision-weighted forecasts. Rather than raising the standard deviation to the exponent two (i.e., squared), I allow the exponent to fluctuate to whatever generates the smallest root mean square error for the predicted forecasts from the precision -weighted forecast. There would be a different set of exponents if I take the exponents that minimize the mean square error of the raw confidence-weighted forecasts or what minimizes the jointly predicted mean square error of the confidence-weighted forecast and the median of the point-estimates. Yet, regardless of which set of efficient exponents I use to create the precision-weighted forecasts, the weighting between the median point-estimate and the precision -weighted forecast are similar.

Table 6 looks at the R^2 from the regression of the answer on just the median of the point-estimate, on just the precision -weighted forecast, and on both the median of the point-estimate and the confidence-weighted forecast. For two of the categories, the growth in R^2 from median of point-estimate to having both forecasts is minimal, but in the other three there are substantial gains (i.e., there is explanatory power in the confidence-weighted forecast).

Category	R^2 with only Median of Point-Estimate	R^2 with only Precision-Weighted Forecast	R^2 for Joint Forecast
Calories	0.585	0.884	0.884
Concert Tickets	0.880	0.873	0.882
Gas Prices	0.308	0.347	0.362
Movie Receipts	0.131	0.129	0.534
Unemployment	0.985	0.987	0.988

Note: The confidence-weighted forecast is optimized by category as in the lower half of Table 5. The table is nearly identical regardless of which efficient weighting scheme I utilize.

Table 6: Comparing the point accuracy of the most promising forecasts, with R^2 from $\boldsymbol{ans = \alpha + \beta * Forecast}$

Two categories of questions have follow-up questions that better calibrate individual-level responses. First, for the calorie question, I ask "How closely do you follow calories?" with options that run from "not very closely" to "regularly, and I know the item in the question." There is no correlation between self-reported information and the rank of absolute error within the question.[16] This conforms to results of Table 2, where similar to stated information, stated confidence levels on a four point scale fails to gain significance in its correlation with accuracy. Second, for the movie question, I ask "Are you interested in seeing 'MOVIE'?" with options that run from "Definitely not going to watch it" to "Have already been excited about it/Definitely watch it in a theater." Thus, this question is combining two dimensions: information and partiality. There is a positive and statistically significant correlation between information/intent and the point-estimate; if a respondent likes a movie, she projects it to earn more money. A fixed-effect regression of information/intent (which runs from 1 to 5) on the point-estimate yields a coefficient of 19.47 (5.33) for information/intent (i.e., respondents estimate the movie to earn $19.5 million more for each degree of intent to see the movie). With repeated data, I would be able to inflate or deflate responses to debias them for the information and partiality of the respondents.

[16] Study I is slightly negative and insignificant and the non-treated Study II is slightly positive and insignificant. Together the coefficient from OLS of intent on rank of error is 0.001 (0.022).

4. Discussion

There are several reasons why a graphical, interactive interface, utilized in a web-based setting, can collect individual-level information information that is difficult to attain in standard telephone or in-person settings. First, information can be revealed, rather than stated, which makes it much harder for the respondents to manipulate the answer to fulfill incentives other than their best estimate.[17] Second, the interface makes the revelation part of the main question, whereas asking a respondent to state confidence after she supplied the main point-estimate, may not seem as serious to the respondent or operate under different incentives. Third, Ariely, et al (2003) shows that people can incorporate new information into their understanding of the world; the problem is that they sometimes appear arbitrary, because they are not sure where with what baseline they should start. A graphical interface can provide some subtle baselines for the respondent without providing too much anchoring. For example, while there was no tested variation in this paper, good example of utilizing this principle is the question regarding calories of fast food includes pictures and descriptions of a few different foods. Similarly, the presentation of the questions themselves are providing subtle information about point-estimates and probability distributions that teach people how to provide information they have, but do not know how to elucidate.

Polls and predication markets are just two methods for gathering individual-level information and aggregating it into forecasts; both methods have benefits and negatives, and my method is one attempt to harness the better aspects of both of them. One of the key problems with polls is the reluctance of researchers to ask the question they are trying to answer, which is usually the question that gathers the most relevant information from the respondent. The graphical and interactive nature of this method allows me to ask questions that do not gather consistent and meaningful responses in a telephone or in-person setting. Polls'

[17] Barber and Odean (2001) shows that men are more overconfident than women, as demonstrated with more aggressive behavior in the stock market, regardless of knowledge. My method reveals confidence and precision without the other confounding factors of utility in a decision such as investing. One reason for the results in Kuklinski (2000) is that respondents perceive an incentive to be strategic about stated confidence levels in political questions, but this is not an issue when people are revealing, rather than stating their additional information.

aggregation does not take advantage of disparities in information of the respondent and prediction markets' aggregation does not record massive amounts of information; participants in prediction markets are creating constant subjective probabilities, but only the aggregated price and a few bids and asks are recorded for researchers. Further, prediction markets are susceptible to manipulation.[18] With my method I can capture all of the information and aggregate it, transparently, with measure of certainty. Further, I can not only create accurate forecasts of the level of the outcome, but also, I can explore full probability distributions on both the individual and aggregate level.

The full method proves itself meaningful in absolute terms and trumps simpler confidence ranges in information, but it does take up more time, which can be important in polling. The mean (median) length of time from start to finish for the five questions with the full method is 13.1 (12.0) minutes, while the confidence range variation is 7.6 (6.7) minutes. Further, while it is not nearly as significant or meaningful as the full probability distribution, the confidence range responses provide a small, but statistically significant positive correlation between confidence and accuracy that can be utilized for the creation of certainty-weighted forecasts. The goal of this paper is to provide validation of my method versus the best and most practical of the other possible options on information and utility, but if time/cost is an issue, there will definitely be scenarios where the confidence range is the right option.

Turning to decision making, there is consensus in the literature on the importance of expectations in decision making. Manski (2004) demonstrates that playing a simple economic game, a subject with one of three different expectations and one of two different utility functions will make the same move (i.e., revealed behavior) in four out of six possible scenarios. He outlines many empirical examples of subjects having faulty expectations, but emphasizes the gap in the literature in understanding expectations separated from utility.

[18] There is evidence that the prediction markets may suffer from manipulation by people motivated to gain publicity for their chosen candidate. The aggregation is over willingness to invest money, not confidence!

The follow-up question for the gas question hints at the usefulness of my method in decoupling expectation from utility in revealed decisions. The main question asks the respondent to imagine that she is driving down a major highway and she notes the price of gas at last few consecutive stations. She is running low and can hold out only long enough to stop at one of the next three stations; she is asked to create a probability distribution of the lowest price of gas among these next three stations. The follow-up question asks what price would induce the driver to stop at the first station she sees, rather than keep going and try one of the following two stations. The median response was at the 30% point of the probability distribution of what they expect the lowest price of the next three gas stations to be. That means that the median driver would stop where they believe that there is only a 30% chance that one of the next two stations would be less. Just 6% of respondents said they would stop at station in the 80% percentile or higher. Most importantly, there is a statistically significant positive correlation between the point-estimate expectation and price in which the driver would stop for gas. Thus, the higher the driver expects the lowest price to be, the higher price the driver will stop and pay. Further, precision demonstrates a meaningful role in the decision making; if two drivers have the same point-estimate, the driver with the larger standard deviation (i.e., a less precise estimate) will stop at a gas station with a higher price. Expectation matters, but so does precision!

5. References

Ariely, Dan, et al., 2003. “Coherent Arbitrariness.” Quarterly Journal of Economics, 118(1):73-105.

Artinger, Floriean, et al., 2010. “Applying Quadratic Scoring Rule Transparently in Multiple Choice Settings: A Note.” Jena Economic Research Papers 2010-021.

Avery, Christopher, et al., 2009. “The “CAPS” Prediction System and Stock Market Returns.”Working paper, Harvard University.

Barber and Odean, 2001. “Boys Will Be Boys: Gender, Overconfidence, and Common Stock Investment,” Quarterly Journal of Economics, 116(1):261-292.

Bates J., and Grange C, 1969. “The Combination of Forecasts.” Operational Research Quarterly, 20:451-468.

Chen, Kay-Yut, et al., 2003. “Predicting the Future.” Information Systems Frontiers, 5(1):47-61.

Clemen, Robert, 1989. “Combining Forecasts: A Review and Annotated Bibliography.” International Journal of Forecasting. 5:559-583.

Clemen, Robert and Robert Winkler, 1993. “Aggregating Point Estimates: A Flexible Modeling Approach.” Management Science, 39(4):501-515.

Delavande, A. and Rohwedder, S, 2008, “Eliciting Subjective Probabilities in Internet Surveys.” Public Opinion Quarterly, 72(5):886-891.

Dominitz, Jeff and Charles Manski., 2006. “Measuring Pension-benefit Expectations Probabilistically.” Labour, 20:201-236.

Dunning, David, et al., 1990. “The Overconfidence Effect in Social Prediction.” Journal of Personality and Social Psychology, 58(4):568-581.

Engleberg, Joseph, et al., 2009. “Comparing the Point Prediction and Subjective Probability Distributions of Professional Forecasters.” Journal of Business and Economics Statistics, 27(1):30-41.

Fair, Ray, and Robert Shiller, 1989. “The Informational Content of ex-Ante Forecasts.” Review of Economics and Statistics 71(2):325-31.

-----------, 1990. “Comparing Information in Forecasts From Econometric Models.” American Economic Review 80(3):375-89.

Gurkaynak, Refet and Justin Wolfers, 2006. “Macroeconomic Derivatives: An Initial Analysis of Market-Based Macro Forecasts, Uncertainty, and Risk.” CEPR Discussion Paper No. 5466.

Kahneman and Tversky, 1979. “Prospect Theory: An Analysis of Decision Under Risk.” Econometrica, 47:263-291.

-----------, 1992. “Advances in Prospect Theory: Cumulative Representation of Uncertainty.” Journal of Risk and Uncertainty, 5(4):297-323.

Kaufman-Scarborough, Carol, et al., 2010. “Improving the Crystal Ball: Harnessing Consumer Input to Create Retail Prediction Markets.” Journal of Research in Interactive Marketing, 4(1):30-45.

Klayman, Joshua, et al., 1999. “Overconfidence: It Depends on How, What, and Whom You Ask.” Organizational Behavior and Human Decision Process. 79:216-247.

Kuklinski, James, et al., 2000. “Misinformation and the Currency of Democratic Citizenship.” The Journal of Politics, 62(3) 790-816.

Manski, Charles, 2004. “Measuring Expectations.” Econometrica, 72(5):1329-1376.

Paolacci, Gabriele, Jesse Chandler, Panagiotis G. Ipeirotis, 2010. “Running Experiments on Amazon Mechanical Turk.”Judetment and Decision Making, 5(5) 411-419.

Pennock and Reeves, 2007. “How and When to Listen to the Crowd.” http://www.overcomingbias.com/2007/02/how_and_when_to.html.

Rothschild, David, 2009. “Forecasting Elections, Comparing Prediction Markets, Polls, and Their Biases.” Public Opinion Quarterly 73(5):895-16.

Rothschild, David and Justin Wolfers, 2012. “Forecasting Elections: Voter Intentions versus Expectations.” Working paper, University of Pennsylvania, Available at: http://assets.wharton.upenn.edu/~rothscdm/RothschildExpectations.pdf.

Smith, Jeremy, and Kenneth Frank Wallis, 2009. “A Simple Explanation of the Forecast Combination Puzzle.” Oxford Bulletin of Economics and Statistics, 71(3):3331-355.

Soll, JB and Klayman J, 2004. “Overconfidence in Interval Estimates.” Journal of Experimental Psychology Learning Memory and Cognition, 30(20):299-314.

Sonnemans, Joep and Theo Offerman, 2001. “Is the Quadratic Scoring Rule Really Incentive Compatible?” Working paper, CREED, University of Amsterdam.

Speirs-Bridge, Andrew, et al., 2010. “Reducing Overconfidence in the Interval Judgments of Experts.” Risk Analysis, 30(3):512-523.

Stock, James and Mark Watson, 2004. “Combination Forecasts of Output Growth in a Seven-Country Data Set.” Journal of Forecasting, 23:405-430.

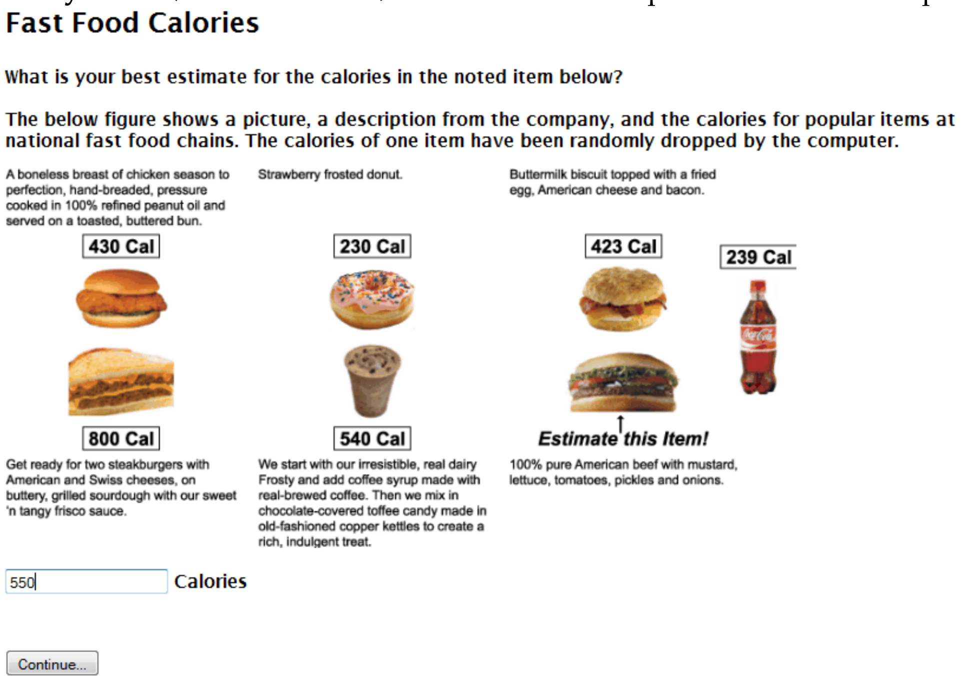
Surowiecki, James, 2004. Wisdom of the Crowds: Why the Many are Smarter Than the Few and How Collective Wisdom Shapes Business, economics, Societies, and Nations, Little, Brown.

Teigen, KH and M. Jorgensen, 2005. “When 90% Confidence Intervals are 50% Certain: On the Credibility of Credible Intervals.” Applied Cognitive Psychology, 19:455-475.

Vul, Edward and Harold Pashler, 2008. “Measuring the crowd within: Probabilistic representations within individuals.” Psychological Science, 19(7):645–647.

6. Appendix:

Study I & II (Calories Count): Here is a full example. It starts with the point-estimate question:



Fast Food Calories

What is your best estimate for the calories in the noted item below?

The below figure shows a picture, a description from the company, and the calories for popular items at national fast food chains. The calories of one item have been randomly dropped by the computer.

A boneless breast of chicken season to perfection, hand-breaded, pressure cooked in 100% refined peanut oil and served on a toasted, buttered bun.

Strawberry frosted donut.

Buttermilk biscuit topped with a fried egg, American cheese and bacon.

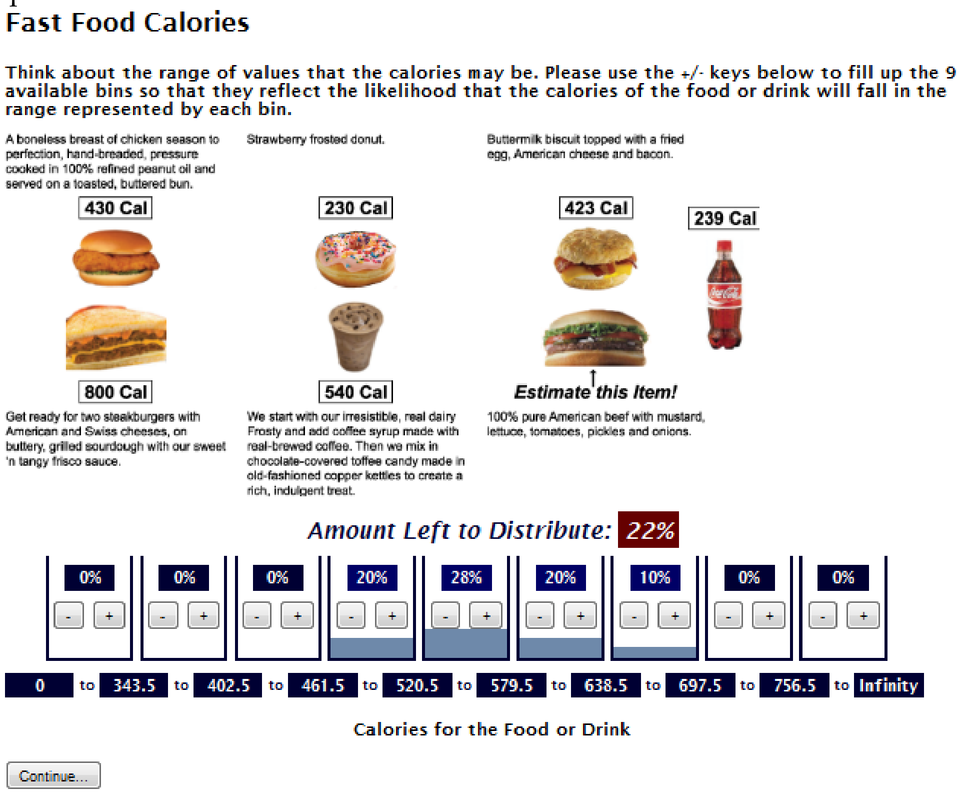
Get ready for two steakburgers with American and Swiss cheeses, on buttery, grilled sourdough with our sweet 'n tangy frisco sauce.

We start with our irresistible, real dairy Frosty and add coffee syrup made with real-brewed coffee. Then we mix in chocolate-covered toffee candy made in old-fashioned copper kettles to create a rich, indulgent treat.

100% pure American beef with mustard, lettuce, tomatoes, pickles and onions.

550 Calories

Calories Count Probability Question: Since I placed 550 calories at my estimate the probability question is centered on 550 calories:



Fast Food Calories

Think about the range of values that the calories may be. Please use the +/- keys below to fill up the 9 available bins so that they reflect the likelihood that the calories of the food or drink will fall in the range represented by each bin.

A boneless breast of chicken season to perfection, hand-breaded, pressure cooked in 100% refined peanut oil and served on a toasted, buttered bun.

Strawberry frosted donut.

Buttermilk biscuit topped with a fried egg, American cheese and bacon.

Get ready for two steakburgers with American and Swiss cheeses, on buttery, grilled sourdough with our sweet 'n tangy frisco sauce.

We start with our irresistible, real dairy Frosty and add coffee syrup made with real-brewed coffee. Then we mix in chocolate-covered toffee candy made in old-fashioned copper kettles to create a rich, indulgent treat.

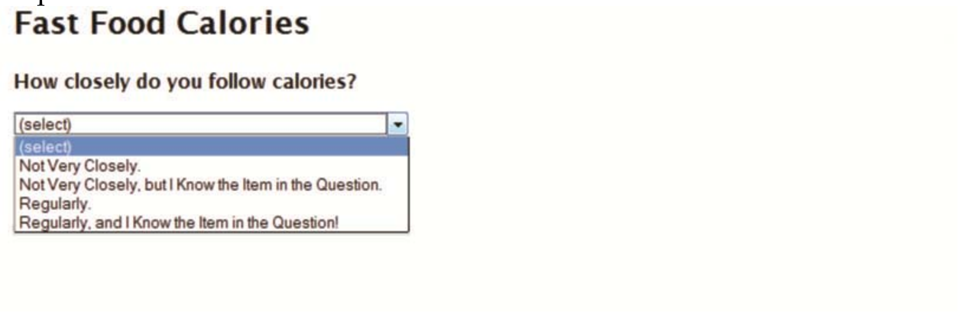
100% pure American beef with mustard, lettuce, tomatoes, pickles and onions.

Amount Left to Distribute: 22%

0%	0%	0%	20%	28%	20%	10%	0%	0%
0 to 343.5	343.5 to 402.5	402.5 to 461.5	461.5 to 520.5	520.5 to 579.5	579.5 to 638.5	638.5 to 697.5	697.5 to 756.5	756.5 to Infinity

Calories for the Food or Drink

Calories Count Follow-up Question: All drop down menus start at select, forcing the respondent to fill a choice.



Fast Food Calories

How closely do you follow calories?

(select)

(select)
Not Very Closely.
Not Very Closely, but I Know the Item in the Question.
Regularly.
Regularly, and I Know the Item in the Question!

Below are the other four questions used in this study:

Study I Only (Concert Ticket Prices): What is your best estimate for the lowest possible price of 2 tickets to the noted concert the day before the show? The top table highlights one randomly selected upcoming concert from the full list of StubHub's top selling concert tours. The bottom table shows the lowest possible price (including all charges) for 2 tickets, from either the box office or StubHub, whichever is lower, that could be purchased 10 days and 1 day from the concert for a random selection of concerts by performers on the same list.

Study I & II (Gas Prices): What is your best estimate for the LOWEST price of gas among the next 3 stations on the highway described in the below table? The below table shows the price of gas at the 3 previous gas stations and the question assumes that you continue down the same highway. All prices are from 8/4/2010 on a major Eastern highway. (9/16/2010 for the Study II.)

Gas Prices Follow-up Question: If you had about enough gas where you felt comfortable driving for up to 3 more stations, what price of gas would induce you to stop at the next station?

Study I Only (Movie Receipts): What is your best estimate for the 4 week gross for *MOVIE* in millions of dollars (i.e. what will *MOVIE* gross domestically through its 4th weekend of wide release)? *DESCRIPTION OF MOVIE*. Nationwide release on *DATE OF RELEASE*. The below table shows the domestic gross for the last 30 wide-release movies through their 4th weekend of release.

Movie Receipts Follow-up Question: Are you interested in seeing *MOVIE*?

Study I & II (Unemployment Rate): What is your best estimate for the August/September Unemployment Rate in the state noted below (use the slider below to show your answer)? The below table shows the Unemployment Rate in a randomly chosen state in a few relevant periods. Unemployment rates are adjusted for seasonal trends.

Unemployment Rate Follow-up Question: How familiar are you with the state in the question?