Running Head: Combining Probability Forecasts

Combining Probability Forecasts:
60% and 60% Is 60%, but Likely and Likely Is Very Likely

Robert Mislavsky
Johns Hopkins University
Carey Business School
mislavsky@jhu.edu

Celia Gaertig
University of Chicago
Booth School of Business
celia.gaertig@chicagobooth.edu

Abstract

How do we combine others' probability forecasts? Prior research has shown that when advisors provide numeric probability forecasts, people typically average them (i.e., they move closer to the average advisor's forecast). However, what if the advisors say that an event is "likely" or "probable?" In 7 studies (N = 6,732), we find that people "count" verbal probabilities (i.e., they move closer to certainty than any individual advisor's forecast). For example, when the advisors both say an event is "likely," participants will say that it is "very likely." This effect occurs for both probabilities above and below 50%, for hypothetical scenarios and real events, and when presenting the others' forecasts simultaneously or sequentially. We also show that this combination strategy carries over to subsequent consumer decisions that rely on advisors' likelihood judgments. We find inconsistent evidence on whether people are using a counting strategy because they believe that a verbal forecast from an additional advisor provides more new information than a numerical forecast from an additional advisor. We also discuss and rule out several other candidate mechanisms for our effect.

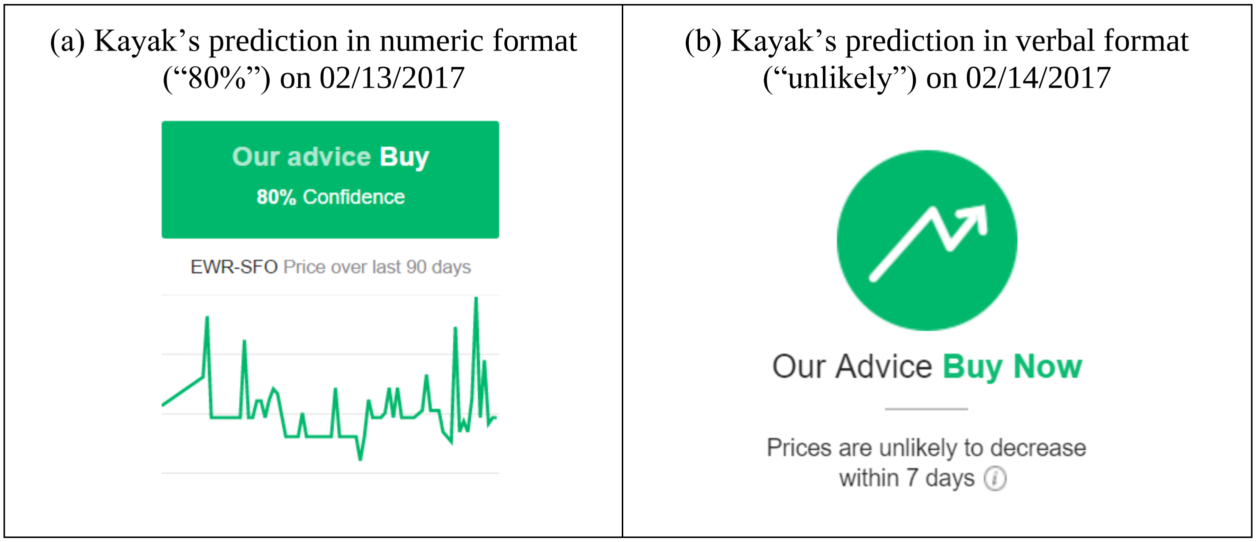
Word count: 179

Keywords: uncertainty, forecasting, verbal probabilities, combining judgments, combining forecasts, predictions

People must navigate uncertainty to make informed decisions. For example, they must judge the likelihood that they would prefer Product A to Product B, that an investment will provide an adequate return, or that the benefits of a new policy will outweigh its costs. In doing so, they might rely solely on their own judgments. They may, however, also consider others' opinions. They may ask their friends, relatives, colleagues, external experts, or any combination of these. They may also consult one of the growing numbers of websites that offer predictions as a service to customers or as businesses unto themselves. For instance, Kayak and Hopper estimate how likely it is that flight prices will change in the coming days. Fuelcaster and Cruisewatch provide similar predictions for gas and cruise prices, respectively. Investment websites such as Betterment predict whether a customer will reach their savings goals by a certain date. Fivethirtyeight and the Upshot became two of the most popular websites for election coverage, in large part because they created prediction models and aggregated forecasts from other sources.

In addition to making accurate forecasts, these websites must decide *how* to present their forecasts to readers. Specifically, they may choose to present forecasts numerically (e.g., "There is a *60% chance* that prices will increase") or verbally (e.g., "It is *likely* that prices will increase"). In fact, Kayak has used both types of formats to present predictions to their customers, see Figure 1 for screenshots of Kayak's predictions in different formats on different dates.

A substantial body of research has shown that numeric and verbal probabilities are interpreted differently in isolation (e.g., Beyth-Marom, 1982; Lichtenstein & Newman, 1967; Windschitl & Wells, 1996). However, given that people often prefer to get multiple opinions before acting (Sah & Loewenstein, 2015; Sarvary, 2002; Schwartz, Luce, & Ariely, 2011; West & Broniarczyk, 1998) and that there are substantial benefits to aggregating multiple opinions (Yaniv, 2004; Yaniv & Milyavsky, 2007)**,** surprisingly little is known about how people actually combine multiple probability estimates to form their own judgments. Where research has examined this question, it has tested how people combine numeric probabilities (Budescu & Yu, 2006, 2007) and optimal strategies for doing so (e.g., Ariely et al., 2000; Ashton & Ashton, 1985; Baron, Mellers, Tetlock, Stone, & Ungar, 2014; Wallsten, Budescu, Erev, & Diederich, 1997). However, although verbal probabilities are much more commonly used (Budescu, Weinberg, & Wallsten, 1988; Erev & Cohen, 1990; Zimmer, 1983), there has been no research on how people combine multiple verbal probabilities.


(a) Kayak’s prediction in numeric format (“80%”) on 02/13/2017

(b) Kayak’s prediction in verbal format (“unlikely”) on 02/14/2017

Our advice Buy
80% Confidence
EWR-SFO Price over last 90 days
Our Advice Buy Now
Prices are unlikely to decrease within 7 days

Figure 1. Prediction formats used by Kayak’s price prediction tool on February 13, 2017 (a) and February 14, 2017 (b).

In this paper, we demonstrate that people use different strategies to combine verbal probabilities than they do to combine numeric probabilities. Replicating past research, we find that participants use an averaging strategy to combine two numeric probabilities. However, we find that they use a “counting” strategy to combine two verbal probabilities. For example, imagine that you are purchasing a plane ticket for your next vacation and you check two websites, Kayak and Hopper, to see if they predict any future price changes. If both websites say that there is a 60%

chance that prices will increase, you would typically average the two and also believe there is a 60% chance. However, if both sites say that it is “likely” that prices will increase, you would act as if you are “counting” each prediction as a positive signal, becoming more confident in your prediction and believing that a price increase is “very likely.”

Our findings are theoretically and practically important. We show that people’s tendencies to use certain combination strategies for verbal vs. numeric probability forecasts differ, contributing to literature on combining forecasts (e.g., Budescu & Yu, 2006, 2007) and on understanding differences in verbal versus numeric probabilities (e.g., Windschitl & Weber, 1999; Windschitl & Wells, 1996). Researchers who study advice-taking and decision-making under uncertainty should consider these differences in people’s combination strategies when designing experiments or creating more formal models of decision making under uncertainty. Practically, our research is relevant to anyone who either seeks or provides advice from multiple sources, and who would be well-served to take differences in how people combine verbal vs. numeric probabilities into account when determining how to use or present multiple forecasts.

NUMERIC PRECISION AND VERBAL EVALUABILITY

That numeric and verbal probabilities have different uses and interpretations is well known to many who rely on probability judgments to make decisions. Within the past decade, both the Intergovernmental Panel on Climate Change (IPCC; Mastrandrea et al., 2011) and the United States Director of National Intelligence (DNI; 2015) issued official guidance on how verbal probabilities must be used in their reports. To ensure that report authors do not use ambiguous terms, both sets of guidelines provide a list of acceptable verbal probability phrases as well as their corresponding numeric probabilities. For example, an event described as “likely” in an IPCC report must have a 66% to 100% likelihood of occurring. On the other hand, an event described as “likely” in a DNI report must have a 55% to 80% likelihood.

The need for such guidance signals several things. First, verbal probabilities are sufficiently common to deserve official comment. Second, verbal probabilities are sufficiently vague to need official conversion charts. They are so vague, in fact, that the IPCC and DNI only have about a 25% overlap in their definitions of “likely.” Finally, it shows that the IPCC and DNI may believe that providing such conversion charts will improve consistency in how verbal probabilities are used and interpreted, although there is mixed evidence on the effectiveness of these guidelines (Budescu, Broomell, & Por, 2009; Budescu, Por, & Broomell, 2012; Budescu, Por, Broomell, & Smithson, 2014).

Although this sort of guidance may improve consistency, verbal and numeric probability phrases differ in several additional ways that may impact how they are interpreted and used. An overview of the academic literature shows that, in general, numeric probabilities are precise but lack direction, whereas verbal probabilities have direction but lack precision.

Precision

Numeric probabilities are more precise than verbal probabilities, as numeric probabilities have an unambiguous interpretation. For example, “a 65% chance” can only mean that an event will occur 65 out of 100 times on average. In contrast, “likely” can have many interpretations (see also IPCC and DNI guidance). Lichtenstein and Newman (1967) illustrated this in a seminal study in which they asked 188 people what they thought “likely” (among other phrases) meant in terms of likelihood and received answers ranging from 25% to 99%.

This difference in the precision between numeric and verbal probabilities is a primary driver of whether people prefer to use or receive verbal or numeric probabilities. Because people typically make relatively coarse probability judgments (Mellers et al., 2015, pp. 275–276) and can comprehend only a few different levels of probability (Zimmer, 1983), they generally prefer to *give* verbal predictions (Erev & Cohen, 1990). These verbal predictions map more easily on to their internal representations of probability (Zimmer, 1983). In addition, for those in the role of an advisor, giving advice in form of verbal probabilities also provides more cover, allowing the advisors to claim accurate predictions for many different outcomes (Erev & Cohen, 1990). Consumers, on the other hand, consider numeric probabilities to be more informative, and thus prefer to *receive* numeric probabilities from others (Erev & Cohen, 1990).

Direction

Although verbal probabilities are less precise than numeric probabilities, it is easier to tell whether a given verbal probability is a "positive" or "negative" sign that an event will occur (Teigen & Brun, 1995, 1999). For example, imagine that a political pundit says that a candidate has a "40% chance" of winning an election. Is that good or bad news for the candidate? If there are only two candidates, a 40% chance is not very good, but if there are 20 candidates, a 40% chance will generally make someone the favorite to win (Teigen, 2001; Windschitl & Wells, 1998). In contrast, a candidate that is "likely" to win will be considered the favorite, regardless of the number of other candidates. The inherent direction of verbal probabilities may also influence how people think of an event: Whereas positive verbal probabilities (e.g., likely, good chance) cue a focus on an event's occurrence, negative verbal probabilities (e.g., unlikely, small chance) cue a focus on an event's failure to occur (Bagchi & Ince, 2015; Teigen & Brun, 1995). Numeric probabilities, in contrast, cue a focus on an event's occurrence regardless of its direction (Bagchi & Ince, 2015; Teigen & Brun, 1995).

COMBINING OTHERS' PROBABILITY ESTIMATES

Although past research has significantly advanced our understanding of the differences that exist between numeric and verbal probabilities, surprisingly little is known about how people *combine* multiple probability estimates in either format to form their own judgments, and this is particularly true for verbal probabilities. In this article, we focus on differences in people's combination strategies for numeric vs. verbal probabilities.

Multiple Numeric Probabilities

In the current article, we focus on numeric probabilities in the form of percentages. Budescu and Yu (2006, 2007) find that people typically average others' forecasts when these forecasts are provided to them as percentages. Sometimes people also use what can be described as a "naïve Bayesian" approach, where they become more confident than the average prediction. This is the case, for example, when multiple advisors give relatively extreme estimates (Budescu & Yu, 2006) or when participants are making predictions about knowledge-based (i.e., epistemic) uncertainty (Wallsten, Budescu, & Tsao, 1997). Such a Bayesian strategy may be normative if advisors have imperfectly correlated information (Ariely et al., 2000; Baron et al., 2014; Wallsten, Budescu, Erev, et al., 1997), but there is little evidence that people take this into consideration when making their own judgments (Budescu & Yu, 2007).

Multiple Verbal Probabilities

No research to date has examined how people combine others' verbal probability estimates for a single event. However, there has been research on how people process multiple verbal probabilities when considering the occurrence of multiple events. When considering the conjunction of two events (i.e., the occurrence of both) or their disjunction (i.e., the occurrence of only one), people act as if they are using a "signed sum" strategy for conjunctions (Yates &

Carlson, 1986) and a “signed average” strategy for disjunctions (Carlson & Yates, 1989), leading to both more conjunction and disjunction errors when using verbal probabilities. Relatedly, Teigen and Brun (1999) find that conjunction errors occur more for positive verbal probabilities (e.g., “likely”), while disjunction errors occur more for negative probabilities (e.g., “unlikely”). In addition, when provided with identical verbal probability phrases, people tend to choose a more extreme and less fuzzy phrase from a separate list to represent the combination of the two, although this is tested with known probabilities (i.e., space on a spinner; Budescu, Zwick, Wallsten, & Erev, 1990).

Optimal Strategies

Given that we are chiefly interested in understanding how people combine others’ probability forecasts, the natural question to ask is: “How *should* we combine probability forecasts?” The unsatisfying answer is, “it depends.” When there are few advisors using similar information to make their forecasts, averaging is typically most accurate, since it reduces the impact of advisors’ idiosyncratic errors (Ashton & Ashton, 1985; Wallsten, Budescu, Erev, et al., 1997; Wallsten & Diederich, 2001).[1] Additionally, if the advisors are using *different* information, then counting positive or negative forecasts may be a good approximation of a Bayesian optimal strategy (Baron et al., 2014; Wallsten & Diederich, 2001). In these cases, the decision-maker has more information than any individual advisor and therefore has “a right to much higher confidence” (Baron et al., 2014, p. 134). Although the purpose of our paper is not to determine what is optimal, we test whether participants consider this when making their own predictions.

[1] However, when combining forecasts from many advisors, a counting strategy may be advantageous, since individual forecasts are often too conservative, particularly for hard-to-predict events (Ariely et al., 2000; Baron et al., 2014; Wallsten, Budescu, Erev, et al., 1997).

COMPARING COMBINATION STRATEGIES

In this article, we examine how people combine multiple advisor forecasts that are presented to them either in numeric (e.g., “60-69%”) or verbal format (e.g., “somewhat likely”). Throughout this paper, we refer to participants using “averaging” or “counting” strategies when combining forecasts, and we compare the proportion of participants using either strategy across conditions.

To illustrate how we measure these strategies, imagine that a participant is tasked with making a forecast on a 10-point subjective likelihood scale, where points 1 to 5 indicate that the event is unlikely (i.e., below 50% chance) to occur, and points 6 to 10 indicate that the event is likely (i.e., above 50% chance). Now imagine that the participant receives two advisor forecasts of 7 and 9 on the 10-point scale, both “positive” signals.

We use the term “averaging” to refer to a combination strategy by which the participant takes a weighted average of the advisors’ forecasts, placing their own forecast in between the advisors’ forecasts. Since the advisors’ forecasts were 7 and 9, a participant who averages these forecasts would answer somewhere at or between the two advisor forecasts, i.e., 7, 8, or 9. What exact forecast the participant will make will depend, for example, on her own prior beliefs, whether she trusts one advisor more than the other, or some other external factor.

In contrast, we use the term “counting” to refer to a combination strategy by which participants interpret each of the advisors’ forecasts as a positive signal and adjust their own forecasts in the direction of the signal, thus moving their own forecast closer to certainty than each of the advisor’s forecasts. Importantly, however, in order to meaningfully distinguish between the counting and averaging strategies, we preregistered to test specifically for the proportion of participants whose forecasts are *above* each of the advisor’s forecasts. For instance, what if a participant sees the first advisor’s forecast (i.e., 7), makes a forecast of 7, sees the second advisor’s forecast (i.e., 9), and then makes a second forecast of 8? This is consistent with both averaging (making a forecast between the two advisors' forecasts) *and* counting (moving closer to 10). Therefore, we can only distinguish between the two strategies when participants move from at or below the most extreme advisor's forecast (i.e., 9 in this scenario) to above it (i.e., 10). A forecast above the most extreme advisor's forecasts is consistent with a counting strategy, but not with an averaging strategy.

OVERVIEW OF STUDIES

We conducted 7 studies in which we asked participants to forecast the likelihood of an uncertain event (Studies 1-3, 5a-b, and 6) or to make a decision that is informed by others' likelihood forecasts (Study 4). In all studies, participants were shown either one or two expert predictions, and we manipulated whether these expert predictions were provided to them numerically (e.g., "60-69%") or verbally (e.g., "somewhat likely"). In all studies, the experts all qualitatively agreed on the outcome (e.g., both advisors predicted that a stock had a greater than 50% chance of increasing). We focus on forecasts in the same direction because it is not possible to distinguish between counting and averaging strategies when experts disagree. For example, if one experts forecasts that an event is "likely" to occur, and another expert forecasts that an even is "unlikely" to occur, then both a counting and averaging strategy would cause a participant to say something like "neither likely nor unlikely" or "even chance." Focusing on forecasts that are in the same direction instead allows us to distinguish between counting and averaging strategies. Participants always made their own predictions on the same scale that the advisor used (i.e., either on a numeric or verbal probability scale).

In our studies, we test differences in combination strategies by comparing the *proportion of extreme forecasts* (i.e., the proportion of participant forecasts that are more extreme than the most extreme advisor's forecast) made by participants across conditions.[2] If participants are using a counting strategy, we predict that the proportion of participants making extreme forecasts will *increase* when they see the second advisor's forecast. However, if participants are using an averaging strategy, we predict that the proportion of participants making extreme forecasts will *decrease* when they see the second advisor's forecast. Specifically, since we predict that participants in the verbal condition being more likely to use a counting strategy and those in the numeric condition being more likely to use an averaging strategy, we predict an interaction between forecast format (i.e., numeric vs. verbal) and the number of advisors (i.e. one vs. two).

In general, we expect that this effect would largely persist for any number of advisors, although it is likely that the size of the effect would decrease or eventually level off, since the amount of new information provided by the thirtieth advisor is likely to be much less than the amount of new information provided by the second advisor. As a result, we only report studies in the main manuscript where participants see up to two advisors. Supplemental Study S1 shows that the effect is largely consistent (although slightly reduced) when participants see up to five advisors.

Studies 1-3 demonstrate that consumers indeed use different strategies to combine numeric versus verbal probabilities. We find that, although people average multiple numeric forecasts, they "count" multiple verbal probability forecasts, resulting in forecasts that are more extreme than each advisor's forecast. These results hold for probabilities above (Studies 1-3, 5-6) and below the midpoint (Study 2), for hypothetical scenarios (Studies 1-2, 4-6) and real events (Study 3), and when presenting multiple probability forecasts simultaneously (Studies 1 and 3) and sequentially (Study 2, 4-6). Furthermore, Study 4 demonstrates that these different strategies influence

[2] Study 4 uses a choice measure.

subsequent consumer decisions. Studies 5-6 and S2-S5 test several potential mechanisms for the effect.

We preregistered all studies, except for Study 1. All deviations from preregistrations are mentioned in footnotes. Analyses preregistered as "secondary" are included in Supplement 4 if not discussed in the main manuscript. For all studies, we report all data exclusions, all manipulations, and all measures. Sample sizes were determined before data collection. Supplementary materials, including data, analysis code, preregistrations, and survey materials, are available at https://bit.ly/2k2ruZR. Links to preregistrations are also listed in Appendix 1.

STUDY 1: LIKELY AND LIKELY IS VERY LIKELY

In Study 1, we examine whether people use different strategies to combine others' forecasts when these forecasts are presented as numeric probabilities or verbal probabilities. We asked participants to predict the likelihood that a stock's price would be higher in one year and presented them with forecasts from two advisors given either numerically (e.g., "60%") or verbally (e.g., "rather likely").

Method

We recruited 205 participants (35.0% female, M_{age} = 33.7 years) on Amazon's Mechanical Turk (MTurk). Participants who completed the survey were paid $0.35.

Participants in this study predicted whether or not a stock's price would be higher one year from the day the study was run. We presented participants with information about a stock (ticker symbol, company name, and most recent closing price), randomly selected from a list of ten. Before making their own forecast, participants saw forecasts made by two (fictional) advisors. We told them, "To help you make your decision, we have provided estimates from two financial analysts."

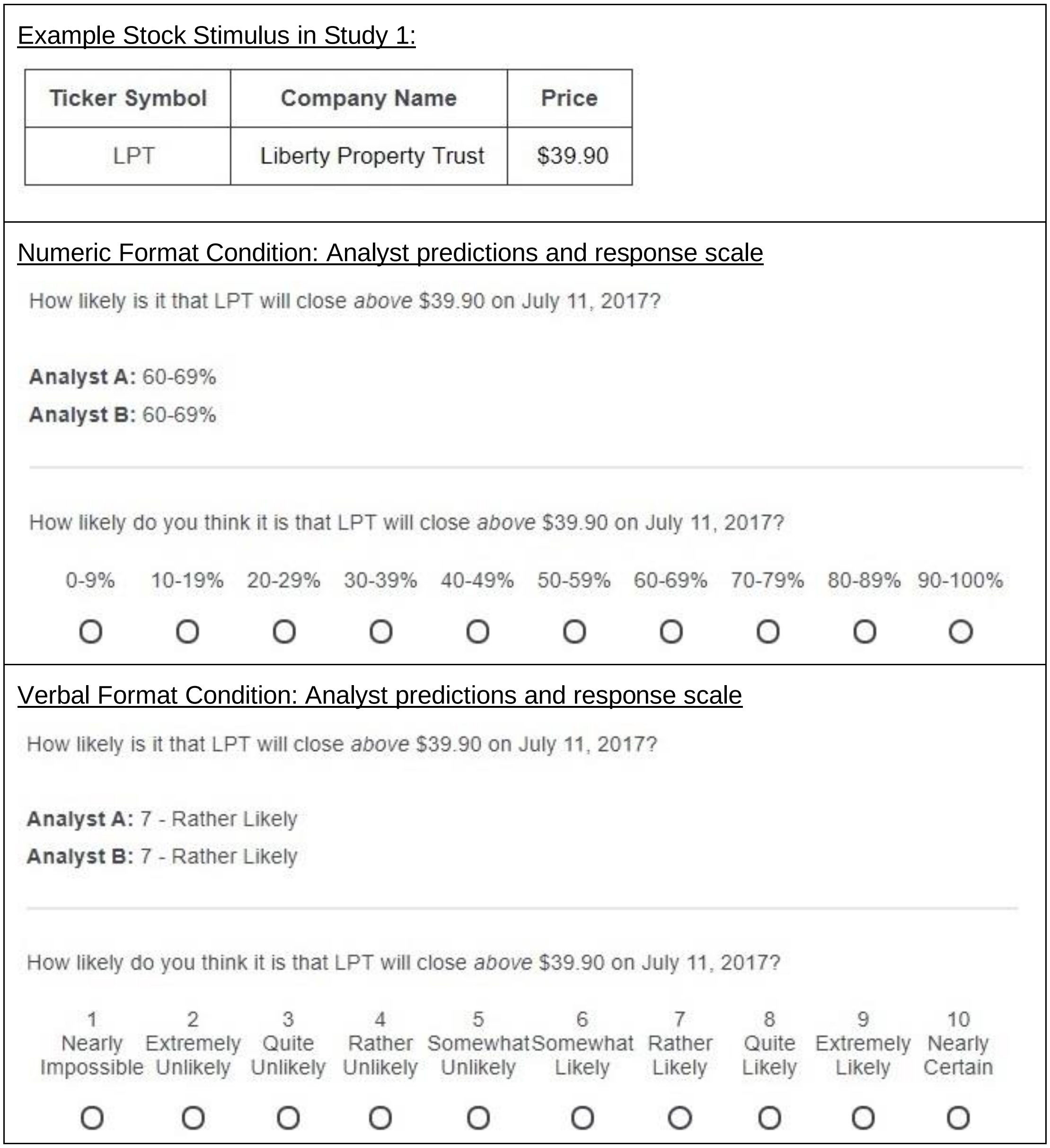
We randomly assigned participants to one of two between-subjects conditions (*numeric* vs. *verbal*). Figure 2 presents both conditions. In the *numeric* condition, both advisors' forecasts were "60-69%," and participants made their forecasts on a 10-point *numeric* probability scale (1 = "0-9%"; 10 = "90-100%"). In the *verbal* condition, both advisors' forecasts were "7 – Rather Likely," and participants made their forecasts on a 10-point *verbal* probability scale (1 = "1 - Nearly Impossible"; 10 = "10 - Nearly Certain"; adapted from Windschitl & Weber, 1999). The advisors' forecasts in both the numeric and the verbal conditions corresponded to the 7th point on participants' response scales, keeping the extremity of advisors' forecasts constant across conditions.

Results and Discussion

We determine participants' combination strategies by assessing the proportion of participants that made *extreme* forecasts across the two conditions. An *extreme* forecast is a participant forecast that is closer to certainty than any individual advisor's forecasts. In this study, since both advisors always provided forecasts that corresponded to the 7th point on the response scale, an extreme forecast is any forecast that is greater than 7 on the 10-point response scale (i.e., 8, 9, or 10). We used a probit regression[3] to regress whether participants made an extreme forecast (1 = yes, 0 = no) on the advice format (1 = verbal, 0 = numeric), including fixed effects for stock.

More participants in the *verbal* condition made extreme forecasts (29.4%) than in the *numeric* condition (10.9%), Z = 3.27, p = .001. This result suggests that participants are more likely to use a counting strategy when combining verbal forecasts than when combining numeric forecasts.

[3] We use probit regressions in all studies where our dependent variable is a binary outcome (e.g., whether or not a participant made an extreme forecast). Our results are nearly identical when using a binary logistic or OLS regression. We report full regression results for all three models in Supplement 4.



Example Stock Stimulus in Study 1:

Ticker Symbol	Company Name	Price
LPT	Liberty Property Trust	$39.90

Numeric Format Condition: Analyst predictions and response scale

How likely is it that LPT will close *above* $39.90 on July 11, 2017?

Analyst A: 60-69%
Analyst B: 60-69%

How likely do you think it is that LPT will close *above* $39.90 on July 11, 2017?

0-9%	10-19%	20-29%	30-39%	40-49%	50-59%	60-69%	70-79%	80-89%	90-100%
○	○	○	○	○	○	○	○	○	○

Verbal Format Condition: Analyst predictions and response scale

How likely is it that LPT will close *above* $39.90 on July 11, 2017?

Analyst A: 7 - Rather Likely
Analyst B: 7 - Rather Likely

How likely do you think it is that LPT will close *above* $39.90 on July 11, 2017?

1 Nearly Impossible	2 Extremely Unlikely	3 Quite Unlikely	4 Rather Unlikely	5 Somewhat Unlikely	6 Somewhat Likely	7 Rather Likely	8 Quite Likely	9 Extremely Likely	10 Nearly Certain
○	○	○	○	○	○	○	○	○	○

Figure 2. Sample Stimuli and Response Scale in Study 1. Advisor forecasts always corresponded to the 7th point on the response scale.

STUDY 2: SEQUENTIAL EVALUATION (AND PROBABILITIES BELOW THE MIDPOINT)

Study 1 showed that people use different strategies to combine others' forecasts depending on whether these forecasts are presented numerically or verbally. When probabilities are presented verbally, a higher proportion of participants made extreme forecasts, suggesting that participants are more likely to “count” verbal probabilities. However, Study 1 has a key limitation—we do not know what participants’ own forecasts would have been if they had only seen one advisor forecast. Thus, it is possible that participants’ own forecasts in the verbal condition are already more extreme when seeing only one advisor forecast. In Study 2, we rule out this possibility by showing participants two forecasts sequentially and comparing the proportion of participants that make extreme forecasts after seeing only one advisor forecast to the proportion of participants that make extreme forecasts after seeing two advisor forecasts. If people use a counting strategy more often for verbal probabilities, we would predict an interaction between advice format and the number of advisors, such that the proportion of participants making extreme forecasts increases as the number of advisors increases from one to two in the verbal condition but not in the numeric condition. In addition, in this study we extend our investigation to probability forecasts below the scale midpoint (i.e., below 50%).

Method

We recruited 854 participants on MTurk, of which 806 (39.0% female, M_{age} = 33.4 years) passed an attention check that was embedded at the beginning of the study. In this study (and subsequent studies with an attention check), participants who failed the attention check were not able to continue with the rest of the study (and are therefore not included in our data). Participants who completed the survey were paid $0.35.

The design of Study 2 was similar to that of Study 1. As in Study 1, all participants saw information about a stock and estimated how likely it was that the stock’s price would be higher in one year. Participants again saw forecasts from fictional advisors, given either numerically or verbally, and made their own forecast on a corresponding scale. However, in this study, we manipulated the number of advisors within subjects. That is, we showed participants the two advisor forecasts one at a time, and participants made two predictions, one after seeing the first advisor's forecast and another after seeing the second advisor's forecast. In addition, we manipulated whether the advisors' forecasts were above or below the midpoint of participants' response scale to test whether our effect holds when "extreme" means "closer to impossibility." Specifically, we randomized whether advisors' forecasts were on the 7th point on the response scale (i.e., "60-69%" or "Rather Likely") or the 4th point on the scale (i.e., "30-39%" or "Rather Unlikely").

In summary, participants were assigned to one of eight conditions in a 2 (number of advisors: one vs. two; within subjects) x 2 (format: numeric vs. verbal; between subjects) x 2 (direction: above vs. below midpoint; between subjects) mixed design.

After participants made their forecasts, we asked them two exploratory mechanism questions about their perceptions of advisor consensus. First, we asked, "If you asked 100 analysts, how many do you think would predict the stock would close above [current stock price] on [date]?" (0-100 slider bar). Second, we asked, "Do the analysts both think that the stock will [go up/not go up]?" (1 = they both definitely do not; 7 = they both definitely do).

Results

As in Study 1, in the *above midpoint* conditions, we classify participants' forecasts as extreme if they are closer to certainty than each advisor's forecast (i.e. above 7 on the 10-point response scale). In the *below midpoint* conditions, we classify participants' forecasts as extreme if they are closer to impossibility than each advisor's forecast (i.e., below 4 on the 10-point response scale). As preregistered, we ran separate analyses for the *above* and *below midpoint* conditions. For each analysis, we used probit regressions to regress whether participants made an extreme forecast (1 =

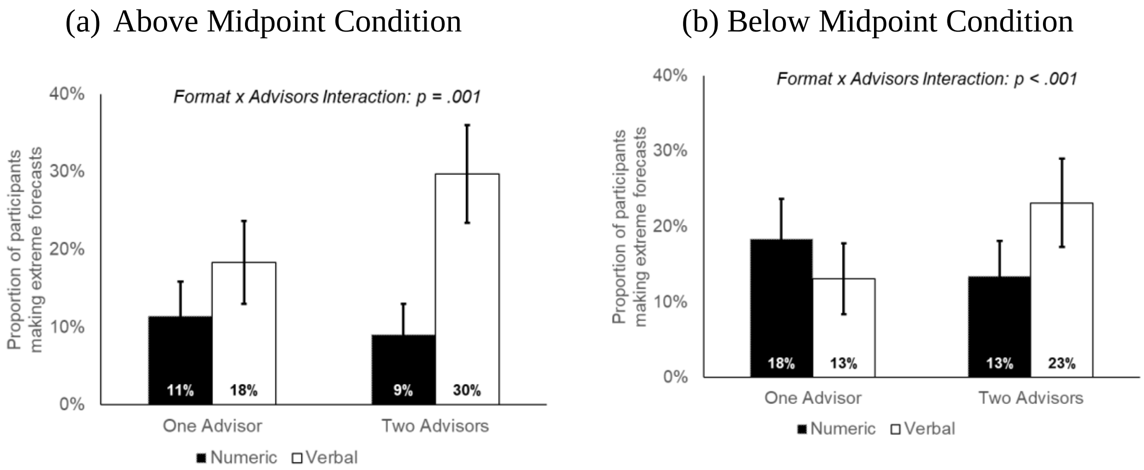
yes, 0 = no) on (1) the advice format condition (1 = verbal, 0 = numeric), (2) the number of advisors (1 = two, 0 = one), and (3) their interaction. These analyses included fixed effects for stock and clustered standard errors by participant.

Figure 3 shows the results for Study 2. For both the above and below midpoint conditions, participants became more likely to make extreme forecasts as they saw additional advisor forecasts in the *verbal* condition but less likely to do so in the *numeric* condition. For probabilities above the midpoint, there is a significant interaction between advice format and number of advisors, Z = 3.62, p < .001 (Figure 3, Panel A). When the advisors made *verbal* forecasts, the proportion of participants making extreme forecasts increased from 18.3% after the first advisor to 29.7% after the second advisor, Z = 4.14, p < .001. In contrast, when the advisors made *numeric* forecasts, the proportion of extreme forecasts directionally decreased from 11.4% to 9.0% after the second advisor, Z = 1.24, p = .21.

This pattern also held for forecasts *below the midpoint* of the response scale (Figure 3, Panel B). Specifically, the interaction between format and number of advisors was again significant, Z = 4.14, p < .001. When the advisors made *verbal* forecasts, the proportion of participants making extreme forecasts increased from 13.1% after the first advisor to 23.1% after the second advisor, Z = 3.28, p = .001. In contrast, when the advisors provided their forecasts as *numeric* probabilities, the proportion of extreme forecasts decreased from 18.3% to 13.4%, Z = 2.46, p = .014.

Combining the *above* and *below* midpoint conditions into one analysis, the interaction between format and number of advisors is also significant, Z = 3.99, p < .001. There is no significant 3-way interaction between format, number of advisors, and above/below midpoint, Z = -.41, p = .68, indicating that the effect is approximately the same size for forecasts above or below the scale midpoint.

Finally, participants perceived more consensus in the *verbal* condition (*p*s < .003), but this did not mediate our effect. When controlling for consensus in our regression analysis, the two-way interaction between format and number of advisors remains significant for the above midpoint, below midpoint, and combined conditions, all Zs = 3.76, *p*s < .001. See Supplement 1 for more details.


(a) Above Midpoint Condition

(b) Below Midpoint Condition

Figure 3. Results from Study 2. Participants' forecasts become more extreme when they see an additional verbal forecast (white bars), but not when they see an additional numeric forecast (black bars). Extreme forecasts are considered to be those closer to certainty than both advisors (i.e., higher than 7 out of 10 in the *above midpoint* condition and below 4 out of 10 in the *below midpoint* condition). Error bars represent 95% confidence intervals.

Discussion

Study 2 rules out the possibility that people simply make more extreme judgments when provided with verbal probabilities, regardless of the number of forecasts they receive. Rather, we find that people using different combination strategies for verbal versus numeric probabilities, resulting in them being more likely to make more extreme forecasts when combining multiple verbal probabilities than when combining multiple numeric probabilities.

STUDY 3: REAL EXPERTS

The preceding studies have demonstrated that people are more likely to use a “counting” strategy when combining verbal probabilities than when combining numeric probabilities. As a result, those combining verbal probabilities make extreme forecasts more often than those combining numeric probabilities. So far, we presented participants with forecasts from fictional advisors who always provided identical forecasts. In Study 3, we provide participants with advice from real experts. Participants predicted the outcomes of a series of baseball games and received advice from well-calibrated sports prediction websites (Fivethirtyeight.com, Fangraphs.com, or VegasInsider.com). This also introduced natural variation across advisor forecasts, meaning that the advisors did not provide identical forecasts, and more granular predictions (on a 0% to 100% scale). In addition, in this study, we manipulated the number of advisors between subjects.

Method

We recruited 626 participants (42.7% female, M_{age} = 35.6 years) from MTurk, each paid $0.50.

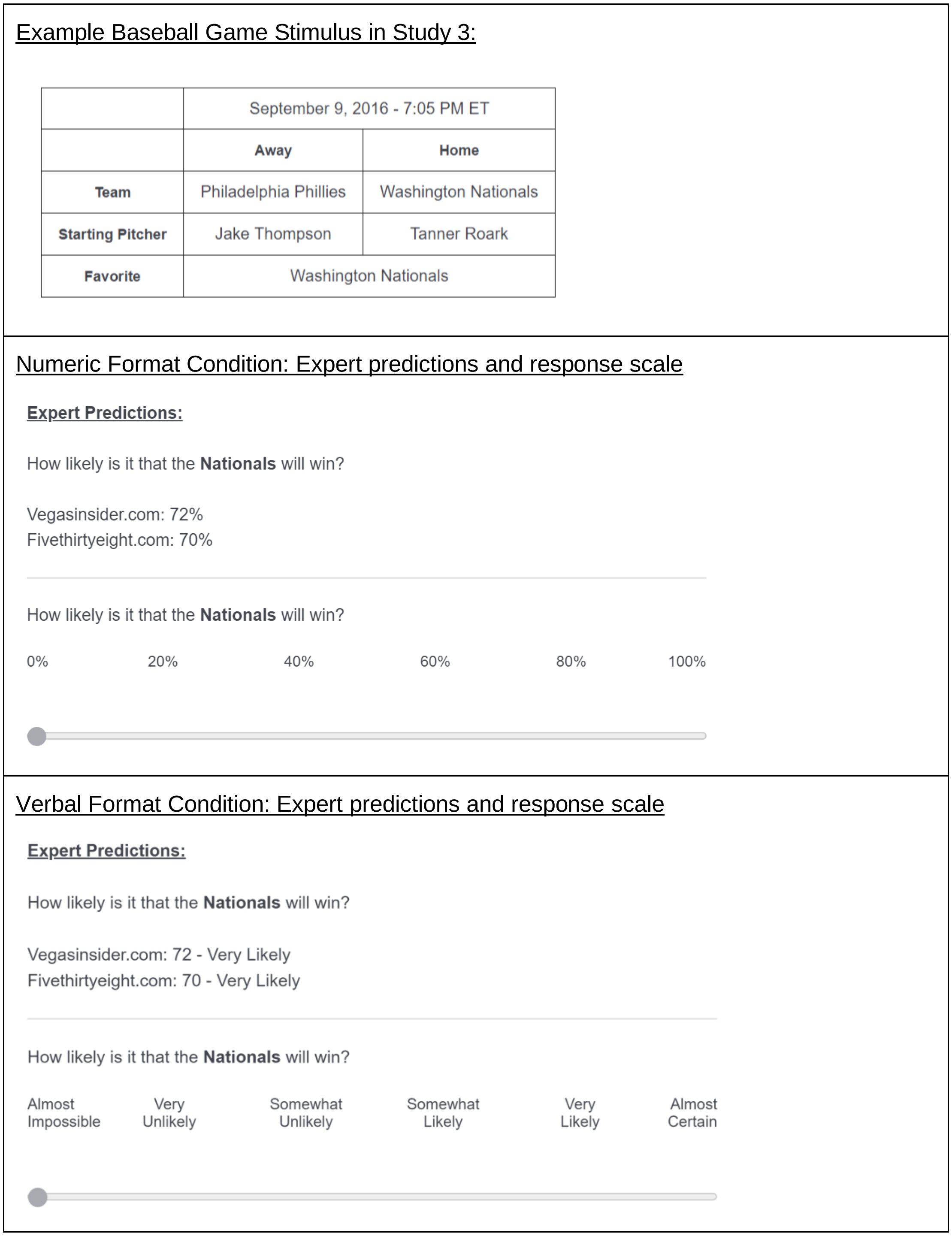
Participants were assigned to one of four between-subjects conditions in a 2 (advice format: numeric vs. verbal) x 2 (number of advisors: 1 vs. 2) design. All participants were shown information about ten Major League Baseball games (randomly selected from the 15 games played on that day) and were asked to predict how likely it was that the favorite would win each game. For each game, participants saw the game’s start time, team names, starting pitchers, and the consensus expert favorite (defined as the team chosen by two of three expert websites). Additionally, they either saw one or two forecasts of how likely it was that the favorite would win the game,randomly selected from Fivethirtyeight.com, Fangraphs.com, or VegasInsider.com. The first two websites give predictions ranging from 0% to 100%, and the third provides “money line” predictions, which we converted to percentages. All expert forecasts were rounded to the nearest percent. Each game was presented to participants on a separate page, in random order.

In the *numeric* condition, advisor forecasts were given as percentages (e.g., “55%”), and in the *verbal* condition, they were given as a whole number with a verbal label (e.g., “55 – Somewhat Likely”). The number was added to the verbal condition to keep the extremity of the advice constant across conditions. Participants made their own predictions on a 0 to 100 slider scale with numeric or verbal labels, depending on condition. Participants either saw advice from one or two advisors before making their own prediction for each game. See Figure 4 for example stimuli.

After participants made their predictions, we asked them six baseball knowledge questions. We also asked participants how motivated they were to make accurate predictions (1 = not at all; 7 = extremely), and how much they trusted each of the three different websites that provided our forecasts (1 = completely distrust; 7 = completely trust). In addition, we asked participants what their favorite team was.

Results and Discussion

Participants made forecasts for all games played on the day of the experiment. However, we can only then meaningfully distinguish extreme from average forecasts when both advisors agreed on the likely winner (i.e., when *both* advisors predicted that there is a greater than 50% chance that the favorite will win the game). As a result, we preregistered that we would only analyze games where the forecasts from the two websites shown agreed on the winner. That is, in the two advisor conditions, we only included games where the websites both gave the same team a greater than 50% chance of winning (67.8% of games). In the one advisor condition, we include only games where the website gave the favorite a greater than 50% chance of winning (84.5% of games). Our results do not change if we include all games in the analyses (see Footnote 5).



Example Baseball Game Stimulus in Study 3:

	September 9, 2016 - 7:05 PM ET	
	Away	**Home**
Team	Philadelphia Phillies	Washington Nationals
Starting Pitcher	Jake Thompson	Tanner Roark
Favorite	Washington Nationals	

Numeric Format Condition: Expert predictions and response scale

Expert Predictions:

How likely is it that the **Nationals** will win?

Vegasinsider.com: 72%
Fivethirtyeight.com: 70%

How likely is it that the **Nationals** will win?

Verbal Format Condition: Expert predictions and response scale

Expert Predictions:

How likely is it that the **Nationals** will win?

Vegasinsider.com: 72 - Very Likely
Fivethirtyeight.com: 70 - Very Likely

How likely is it that the **Nationals** will win?

Figure 4. Sample Stimuli and Response Scales in Study 3. Stimuli shown correspond to *2 advisor* condition. Participants in *1 advisor* condition saw nearly identical stimuli, except with only one expert prediction listed. Expert predictions were randomly selected from list of three.

As in Studies 1 and 2, we classify participant forecasts as "extreme" if they are closer to certainty than each advisor's forecast for each game (i.e., closer to 100 than each advisor).[4] We used probit regression to regress whether participants made an extreme forecast (1 = yes, 0 = no) on (1) the advice format (1 = verbal, 0 = numeric), (2) the number of advisors (1 = two, 0 = one), and (3) their interaction, including fixed effects for game and clustered standard errors by participant. As preregistered, we included participant motivation, baseball knowledge, and average advisor forecast as control variables in the regression. Our results do not change if these controls are not included (see Supplement 4).

When participants saw only one advisor forecast, there were no differences between the proportion of extreme forecasts in the *numeric* (50.0%) and *verbal* (55.5%) conditions, Z = 1.39, $p = .17$. However, participants who saw two advisor forecasts made many more extreme forecasts in the *verbal* condition (46.6%) than in the *numeric* condition (29.8%), Z = 5.11¸ $p < .001$. The interaction between format and number of advisors was significant, Z = 2.93, $p = .003$, and remains so when we exclude control variables, Z = 2.56, $p = .011$.[5] Thus, when extending our investigation to probability forecasts provided by real experts for real events, we again see that participants use different strategies to combine numeric vs. verbal probability forecasts.

Note that, unlike in Study 2, the number of extreme forecasts decreased in both conditions when participants saw two advisor forecasts. We believe that this is due to the granularity of the scale in this study. That is, it is more difficult to give an exactly average forecast in this study (when it is 1 out of 101 points) than in Study 2 (when it is 1 out of 10 points). Indeed, in this study,

[4] We originally preregistered that we would look at the proportion of forecasts above the average expert's forecast. We ultimately decided to use the proportion of "extreme" forecasts because it is a more precise test of our hypothesis. Using the proportion of above average forecasts yields the same results, where the interaction format and number of advisors is significant, with controls (Z = 2.72, $p = .007$) or without controls (Z = 2.44, $p = .015$).

[5] Including all games, the interaction is still significant, with controls (Z = 3.23, $p = .001$) or without controls (Z = 2.80, $p = .005$). We also preregistered that we would run our analyses excluding games involving participants' favorite teams. The interaction is still significant with controls (Z = 2.86, $p = .004$) or without controls (Z = 2.51, $p = .012$).

only 13.0% of forecasts were exactly "average" when participants only saw one advisor forecast, compared to 44.2% in Study 2. It is also possible that participants had stronger prior beliefs in this study, leading to more initially extreme estimates.

STUDY 4: DECISIONS BASED ON FORECASTS

In Studies 1-3, we find that, when participants see multiple *verbal* probability forecasts from advisors, they are much more likely to use a "counting" strategy than when they see multiple *numeric* probability forecasts. However, it may be that this is caused by differences in how participants use the respective response scales rather than, as we hypothesize, actually becoming more certain of an event's outcome. In Study 4, we test whether our findings extend to participants' decision-making informed by their beliefs about the event's likelihood.

Method

We recruited 809 participants (44.8% female, M_{age} = 35.3 years) on MTurk, each paid $0.40.

Participants in this study indicated whether they would buy a certain product now or wait to buy it, due to some uncertainty regarding the product. Participants were randomly assigned to one of four conditions in a 2 (scenario: plane ticket vs. cell phone) x 2 (advice format: numeric vs. verbal) between-subjects design. Participants were either assigned to read a scenario about buying a plane ticket or a scenario about buying a new cell phone. In both scenarios, participants learned about uncertainty related to their purchase and were told that they sought expert predictions to inform their decisions. Participants then saw advice from two advisors who provided probability estimates about the product's uncertainty (e.g., whether there would be a price change) either in numeric or verbal format. We describe one of the scenarios, the plane ticket scenario, in detail below, and we provide the exact wording of both the plane ticket scenario and the cell phone scenario in Appendix 2.

In the plane ticket scenario, participants read that they were considering buying a plane ticket for an upcoming vacation and that the price could change in the coming weeks. They then read that they checked a price prediction website, which forecasted that the price of the plane ticket would drop, giving either a *verbal* (e.g., “somewhat likely”) or *numeric* (e.g., “55% chance”) probability forecast, essentially recommending that the participant wait to make the purchase. Participants then indicated whether they would buy the plane ticket or wait to make the purchase on a 7-point scale (1 = Definitely buy; 7 = Definitely wait). After making their decision, participants read that they checked a second website, which also forecasted that the price of the plane ticket would drop, and confirming the first website’s recommendation. After seeing the forecast from the second website, participants again indicated whether they would buy the plane ticket or wait with the purchase on the same 7-point scale.

For the sake of stimulus sampling, in each scenario, we included multiple probability levels (i.e., 55% and 65% in the *numeric* condition and “rather likely” and “somewhat likely” in the *verbal* condition). However, the second website always made the same forecast as the first website, thus confirming the prediction, and we preregistered that we would collapse results across probability levels. In addition, in order to keep the stimuli realistic, we used slightly different wording for the first and second forecast that participants saw, and we randomized which wording participants saw first or second.

The cell phone scenario shared the same basic structure and dependent measures as the plane ticket scenario. Participants read that they were buying a cell phone, that there was some uncertainty about whether a new model would be released shortly, and then received forecasts about the likelihood of the new model being released (see Appendix 2 for the full scenario).

For both scenarios, after participants indicated their purchase intentions, we collected four additional measures to test for potential mediation. Specifically, we asked the extent to which the websites used the same or different information to make their predictions (1 = the same information; 7 = completely different information), the extent to which the second website provided new information (1 = definitely did not; 7 = definitely did), how useful the second website was (1 = not at all; 7 = extremely), and which prediction participants weighed more (1 = only the first prediction; 7 = only the second prediction)

Results

Main analysis. Following our preregistration, we collapsed the data across both scenarios and probability levels. We then constructed a binary measure indicating whether or not participants became more likely to follow the websites' advice (i.e., became more likely to wait; 1 = yes, 0 = no). We used probit regression to regress whether participants became more likely to wait on the advice format condition (1 = verbal, 0 = numeric), including fixed effects for each stimulus (i.e., scenario and probability level).

After seeing the second website's forecast, more than a third of participants (33.8%) in the *verbal* condition became more likely to follow the websites' advice, compared to 20.5% in the *numeric* condition, Z = 4.31, $p < .001$.[6] This indicates that participants updated their beliefs more when seeing an additional verbal forecast than when seeing an additional numeric forecast. Thus, differences in combination strategies influence internal representations of uncertainty enough to translate into downstream decisions.

As a preregistered secondary analysis, we also considered participants' mean (i.e., untransformed) responses. Overall, participants were more willing to follow the websites' advice

[6] We preregistered that we would also conduct a proportions test. The results are identical, Z = 4.23, $p < .001$.

when it was provided in verbal format than when it was provided in numeric format (M = 5.32 vs. M = 4.92), t(806) = 3.00, *p* = .003. They also became more likely to follow the websites' advice after the second forecast compared to after the first forecast (M = 5.25 vs. M = 4.99), t(806) = 5.76, *p* < .001. Importantly, however, there was also a significant interaction between the advice format and the number of forecasts that participants saw, such that participants increased their answers more after getting a second opinion in the verbal condition than they did after getting a second opinion in the numeric condition, t(806) = 2.26, *p* = .024. That is, confirming the results from our main analysis, participants updated their beliefs more in the verbal condition than in the numeric condition.

Mediation analysis. Considering our four exploratory mediator questions, participants believed that the websites were more likely to have used different information in the *verbal* condition (M = 3.08) than in the *numeric* condition (M = 2.65), t(797) = 4.43, *p* < .001. They also thought that the second website provided more new information in the *verbal* condition (M = 2.92 vs. M = 2.37), t(797) = 4.55, *p* < .001, and was more useful (M = 4.63 vs. M = 4.07), t(798) = 4.64, *p* < .001, in the *verbal* condition than in the *numeric* condition. There was no difference in the weight that participants reported placing on each prediction across the conditions (numeric: M = 3.95, verbal: M = 3.98), t(797) = .54, *p* = .59.

To test whether the three measures that differed across conditions impacted participants' willingness to follow the websites' advice, we performed three separate bootstrapped mediation analyses, using 10,000 samples, one for each of the three mediators. The extent to which participants felt the websites used different information (*b* = .045; 95% CI: .007, .088) and the extent to which the second website provided new information (*b* = .052; 95% CI: .004, .106) partially mediated our effect at the 95% level. The perceived usefulness of the second website did not mediate our effect at the 95% level but did at the 90% level (*b* = .063; 95% CI: -.005, .133; 90% CI: .006, .121).[7] Table 1 shows that the effect of forecast format (i.e., numeric vs. verbal) slightly decreases, but is not eliminated when mediators are included. These results suggest that people's increased tendency to use a counting strategy for verbal probabilities may be driven by their belief that there is less information overlap between advisors, making their forecasts individually more informative.

Table 1. Study 4 results

	(1)	(2)	(3)	(4)	(5)
Constant	-.593*** (.096)	-.954*** (.141)	-.989*** (.123)	-1.93*** (.196)	-2.07*** (.213)
Numeric (0) or Verbal (1) format	.417*** (.097)	.380*** (.099)	.370*** (.010)	.359*** (.102)	.324** (.104)
Do you think the two sites used the same or different information to make their predictions? (1 = the same; 7 = completely different)		.125*** (.035)			.012 (.042)
To what extent did the second site provide you with new information? (1 = definitely did not; 7 = definitely did)			.146*** (.028)		.079* (.033)
How useful was the second site for making your decision? (1 = not at all; 7 = extremely)				.276*** (.034)	.254*** (.036)
Fixed effects	Probability level, scenario				

† p ≤ .1, * p ≤ .05, ** p ≤ .01, *** p ≤ .001

Notes. Probit regression predicting whether a participant became more confident (i.e., increased their answer) after seeing the second website's forecast. The effect of format remains significant, but decreases slightly, when mediator variables are included. Standard errors listed in parentheses.

[7] We highlight this to note that these three results are all qualitatively similar and do not differ from each other (i.e., the 95% CI for the usefulness mediator fully contains the 95% CI for the other mediators), even though they do not all reach traditional levels of significance (Gelman & Stern, 2006; Simonsohn, 2014).

STUDIES 5A-B: ARE PEOPLE BEING BAYESIAN?

As discussed in the Introduction, although it is typically better to average others' forecasts if those forecasts were generated using correlated information (e.g., Ashton & Ashton, 1985), making more extreme estimates is the normatively correct strategy if the forecasts were generated using uncorrelated information (Baron et al., 2014; Wallsten & Diederich, 2001). The mediation results from Study 4 suggest that participants may be behaving like naïve Bayesians. Specifically, when two advisors give similar numeric forecasts, the precision of those forecasts might lead participants to believe the advisors are using similar information. On the other hand, participants may not have this intuition when advisors make similar verbal forecasts but may instead believe the advisors are using different information. This would lead to more extreme participant predictions in the verbal condition, compared to the numeric condition. However, the evidence in Study 4 is not particularly strong, as the mediators only account for between 11% and 16% of our effect. In addition, the task that participants completed in Study 4 is qualitatively different from the one participants completed in Studies 1-3. Thus, we conducted Studies 5a and 5b to test this mechanism more completely, using a scenario similar to that in Study 2. In Study 5a, we measured the potential mediator, and, in Study 5b, we directly manipulated whether or not we told participants that the advisors were using different information/methods to come up with their forecasts.

STUDY 5A: PERCEIVED CORRELATION OF INFORMATION

Method

We recruited 1,091 participants on MTurk, of which 816 passed an attention check (40.8% female, M_{age} = 36.0 years). Participants who completed the survey were paid $0.40

The design of Study 5A was identical to that of Study 2, except that the advisors' forecasts were all above the midpoint and corresponded to the 7th point on the 10-point response scale (i.e., "60-69%" in the numeric condition and "rather likely" in the verbal condition). That is, participants were randomly assigned to one of four conditions in a 2 (*number of advisors*: one vs. two; within subjects) x 2 (*format*: numeric vs. verbal; between subjects) mixed design.

We included two mediator questions at the end of the survey. The first question is central to our hypothesis that participants are acting in a Bayesian-rational way and measured participants' perceived correlation between the information that advisors were using to generate their forecasts. Specifically, we asked participants, "Do you think the analysts are using the same or different information to make their forecasts?" (1 = exactly the same information; 7 = completely different information). We added another question designed to test the perceived consensus between the two advisors that asked participants, "Do the analysts *both* think that the stock will go up?" (1 = they both definitely do not think that; 7 = they both definitely do think that). We report the results for this question for completeness.

Results

Main analysis. As in our prior studies, we used probit regression to regress whether participants made an extreme forecast (1 = yes, 0 = no) on (1) the advice format (1 = verbal, 0 = numeric), (2) the number of advisors (1 = two, 0 = one), and (3) their interaction, including fixed effects for stock and clustered standard errors by participant.

We again found a significant positive interaction between advice format and number of advisors, Z = 2.60, p = .009. The proportion of extreme forecasts increased from 24.1% to 33.5% when participants saw the second advisor forecast in the *verbal* condition, but did not substantially change when participants saw the second advisor forecast in the *numeric* condition (14.2% vs.

14.7%). Thus, we replicated our main finding that participants average numeric forecasts but "count" verbal forecasts:

Analysis of mediator questions. However, unlike in Study 4, we did not find significant differences in the perceived correlation of advisors' information across conditions (verbal: M = 2.90; numeric: M = 2.78), t(791) = 0.95, *p* = .34. There was also no significant difference in perceived consensus among advisors (verbal: M = 5.92; numeric: M = 5.90), t(791) = .38, *p* = .71. Although this is initial evidence against the idea that participants' combination strategies are driven by their beliefs about the correlation of the advisors' information, it may be that the measure of perceived correlation that we used in this study is imprecise or that the effect is subtle enough that participants cannot meaningfully answer our mediator question in this context. To provide a stronger test of this mechanism, we directly manipulated the sources of information that advisors are using in Study 5b.

STUDY 5B: MANIPULATED CORRELATION OF INFORMATION

Method

We recruited 1,191 participants on MTurk, of which 1,033 passed an attention check (50.3% female M_{age} = 37.9 years). Participants who completed the survey were paid $0.35.

The design of Study 5B was similar to that of Study 5A. However, we also manipulated the sources of information that we told participants the advisors used to generate their forecasts. As a result, participants were randomly assigned to one of eight conditions in a 2 (*number of advisors*: one vs. two; within subjects) x 2 (*advice format*: numeric vs. verbal; between subjects) x 2 (*information source*: different vs. similar; between subjects) mixed design. We manipulated the number of advisors and the advice format as we did in Study 5A. In addition, in the *different information* condition, participants were told that each advisor used different sources of information (i.e., one used a statistical model and one used general knowledge and intuition, in counterbalanced order). In the *similar information* condition, participants were told that both advisors used similar sources of information (i.e., both used a model or both used general knowledge and intuition[8]). After participants made their forecasts, they indicated the extent to which they felt that the advisors were using the same versus different information (1 = exactly the same information, 7 = completely different information) as a manipulation check.

There are three possible outcomes for this study. First, overlap in information source could explain the *entirety* of our effect (i.e., the interaction between advice format and number of advisors that we obtained in our previous studies). If this is the case, the verbal and numeric conditions should behave essentially identically when information source is held constant. That is, in the current study, the interaction between advice format and number of advisors would no longer be significant. Instead, there would be a significant interaction between information source and number of advisors, such that participants become more likely to make an extreme forecast when the first and second advisors are using different information, regardless of whether forecasts are made numerically or verbally. Second, information source could explain *some* of our effect. Here, the interaction between advice format and number of advisors would be still be present, but reduced, when manipulating information source. Specifically, there would be a significant three-way interaction between advice format, number of advisors, and information source. Finally, information source might *not explain any of our effect*. If this is the case, the interaction between

[8] To account for any effects caused by the information source itself (i.e., a statistical model vs. general knowledge), we randomly assigned participants to see whether both advisors used a model or both used general knowledge and collapse them for analysis. There was no difference in the perceived overlap of information between participants who saw that both advisors used a model (M = 2.79) and those who saw that both advisors used general knowledge (M = 2.70), t(505) = .73, *p* = .47.

advice format and number of advisors would remain significant and there would be neither an interaction between information source and number of advisors nor a three-way interaction.

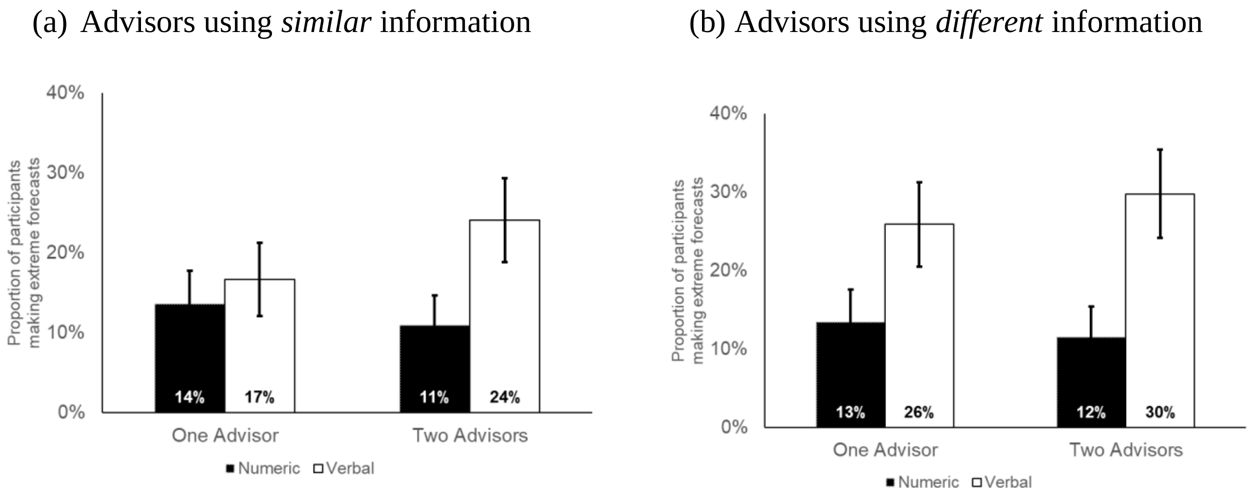
Results

Manipulation check. Consistent with our manipulation, we found that participants were more likely to indicate that the advisors were using similar information when they were told that the advisors used the same method to make their predictions (M = 2.75) than when they were told the advisors used different methods (M = 4.21), t(1,009) = 15.1, $p < .001$.

Main analysis. Our results are consistent with information source *not explaining any of our effect*. The interaction between advice format and number of advisors remains significant, Z = 3.02, $p = .003$. That is, we again find that participants in the numeric format condition become more likely to make an extreme forecast after seeing the second advisor's forecast than those in the verbal format condition do. However, neither the interaction between information source and number of advisors, Z = .27, $p = .79$, nor the three-way interaction between format, number of advisors, and information source, Z = .95, $p = .34$, is significant. Figure 5 illustrates this: The results are nearly identical in both the *similar* (panel a) and *different* (panel b) information conditions, indicating that manipulating the information source does not impact how participants are combining verbal and numeric forecasts.

Discussion

Taken together, the results of Studies 5a and 5b suggest that, unlike what Study 4 suggested, our effect is not driven by participants using a Bayesian updating strategy. In Study 5a, we did not find that participants thought that the advisors were more likely to use different sources of information in the verbal condition than in the numeric condition. In Study 5b, where we directly manipulated whether the advisors were using different or the same sources of information, this did not translate into different combination strategies.


(a) Advisors using *similar* information

(b) Advisors using *different* information

Figure 5. Study 5b results. The overall pattern of results is consistent with prior studies, and does not change whether participants believe that the advisors were using similar information (panel a) or different information (panel b), indicating that this does not influence their combination strategies.

STUDY 6: MAKING NUMERIC PROBABILITIES EVALUABLE

The results of Studies 5a and 5b suggest that it is unlikely that the observed differences in combination strategies are attributable to participants assuming that advisors are more likely to use different information when they are providing verbal forecasts than when they are providing numeric forecasts. In Study 6, we test another explanation for our effect: It may be easier to infer the *direction* of the forecast from verbal probabilities than it is from numeric probabilities. Such inferences would be consistent with research showing that verbal probabilities have an inherent "direction" (Teigen & Brun, 1995, 1999). That is, they are easily classifiable as a "good sign" or a "bad sign," making it more intuitive to use a signed summation strategy, where all signals that are in the same direction are added together (Yates & Carlson, 1986). We designed Study 6 to test whether adding information about the direction of the forecasts to numeric probabilities increases people's tendency to use a counting strategy for numeric forecasts.[9]

Method

We recruited 3,364 participants on MTurk, of which 2,437 passed an attention check (50.9% female M_{age} = 35.9 years).

The design of Study 6 was similar to those of Studies 2 and 5a-b. Participants predicted the likelihood that a stock's price would be higher in one year, and we manipulated the number of advisors within subjects, such that participants saw forecasts from two advisors sequentially. However, in this study, participants were assigned to one of three between-subjects conditions. As in our prior studies, two of these conditions manipulated whether participants saw *verbal* (i.e., "rather likely") or *numeric* (i.e., "60-69%") forecasts from advisors. In the third condition (*numeric with direction*), participants saw a numeric forecast, but we added a note at the beginning of the survey saying that any advisor prediction above 50% should be considered a positive sign. We also highlighted the advisor's predictions in green during the forecasting task to emphasize this point. We also repeated this note when participants were making their actual predictions. In a separate study with a nearly identical design (Study S3), we included a manipulation check question asking participants to indicate whether the advisors' predictions represented a good or bad sign that the stock's price would be higher in a year (1 = definitely a bad sign; 7 = definitely a good sign), confirming that participants perceive verbal forecasts to be a better sign than numeric forecasts without added direction (5.49 vs. 5.22); t(799) = 4.03, *p* < .001 and that adding direction

[9] Prior to running this study, we ran three nearly identical studies (Studies S2-S4 in Supplement 3) with minor design differences and with results that directionally suggested that our hypothesis was correct (i.e., that adding a signal of direction to the numeric condition would make participants more likely to use a counting strategy). However, we ultimately did not replicate this result and only include the most recent and most highly powered experiment here.

to numeric predictions increases participants' belief that the prediction is a good sign (5.22 without direction vs. 5.66 with direction), t(799) = 6.76, *p* < .001.

Results

We first focus on the comparison between the *numeric* and *verbal* conditions. Replicating the results from our previous studies, we again found that participants were more likely to count *verbal* forecasts than *numeric* forecasts. In the *verbal* condition, the proportion of extreme forecasts increased from 19.9% to 31.0%, Z = 7.04, *p* < .001, after participants saw the second advisor's forecast. But in the *numeric* condition, the proportion of extreme forecasts directionally decreased from 12.9% to 11.7%, Z = 1.27, *p* = .205. The interaction between the *numeric* vs. *verbal* advice format and the number of advisors (*one* vs. *two*) was again significant, Z = 5.60, *p* < .001.

We next focus on the comparison between the *numeric* and *numeric with direction* conditions. If the evaluability of verbal forecasts is what leads people to use a counting strategy, then we would expect that participants in the *numeric with direction* condition also become more likely to use a counting strategy. That is, we would expect the proportion of extreme forecasts in the *numeric with direction* condition to increase after participants see the second advisor's forecast (similar to what happens in the *verbal* condition) and for there to be a significant interaction between the *numeric* vs. *numeric with direction* advice format and the number of advisors (*one* vs. *two*). However, this is not what we see. Participants in the *numeric with direction* condition act similarly to those in the *numeric* condition: The proportion of extreme forecasts only slightly increased from 17.0% to 18.2%, Z = .93, *p* = .35 as participants saw a second advisor forecast. Thus, we also do not see a significant interaction, Z = 1.53, *p* = .126, providing no evidence that making numeric forecasts more evaluable by telling participants that the forecast is a positive sign increases people's tendency to use a counting strategy to combine these forecasts.

GENERAL DISCUSSION

In 7 studies, we show that people combine probability forecasts from multiple advisors differently depending on whether the forecasts are given as numeric or verbal probabilities. Specifically, people are more likely to average numeric probability forecasts and more likely to *count* verbal probability forecasts. As a result, participants' own forecasts become more extreme (i.e., closer to certainty) as they see additional verbal forecasts from advisors, but closer to the average of the advisors' forecasts as they see additional numeric forecasts. The differences in these strategies affect internal representations of uncertainty enough to influence decisions based on this uncertainty, as shown in Study 4. Finally, it does not appear that participants are behaving this way because they believe that a verbal forecast from an additional advisor provides more new information than a numerical forecast from an additional advisor or because verbal probabilities imply an inherent direction. Below we discuss several additional candidate mechanisms that we tested, although we do not find support for any of them. We conclude by discussing theoretical and managerial implications.

Evaluating further potential mechanisms

Although we tested what we felt were the most plausible mechanisms in Studies 5-6, we tested several additional candidate mechanisms in Studies S2 and S5. Below, we summarize the most important findings from these studies. The exact wording of each of the measures and all detailed results for Studies S2 and S5 in Supplement 3.

Intuition vs. reason. It may be that people are able to make quicker and more intuitive Bayesian judgments when probabilities are presented verbally (Gigerenzer & Hoffrage, 1995; although c.f., Biswas et al., 2011), while numeric probabilities trigger rule-based thinking (Windschitl & Wells, 1996). This could move participants in the direction of making deliberate mathematical calculations, such as averaging, in the numeric condition, thus leading to the pattern of results that we see in our studies. However, in Study S2, we find no difference in participants' reported use of intuition versus reason across the verbal (M = 4.88) and numeric conditions (M = 5.01), t(400) = .88, *p* = .38.

Confidence. The inherent direction of verbal probabilities (Teigen & Brun, 1995, 1999) may also lead participants to conclude that advisors who provide verbal probabilities have more strongly held beliefs or that there was wider agreement amongst the experts. This, in turn, may have increased participants' own confidence in their forecasts, leading them to make more extreme forecasts. However, in Study S2, we find no difference in participants' self-reported confidence between the verbal (M = 4.59) and numeric (M = 4.45) conditions, t(400) = .97, *p* = .33. Additionally, in Study S5, participants thought advisors had slightly stronger opinions in the *numeric* condition (M = 4.65) than in the verbal condition (M = 4.36), t(400) = 2.05, *p* = .041, although this does not mediate our effect (see Table S10 in Supplement 3).

Epistemic vs. aleatory uncertainty. Tannenbaum, Fox, and Ülkümen (2016) report that people make more extreme probability judgments when they believe uncertainty to be more epistemic (i.e., knowable in advance) rather than aleatory (i.e., determined by chance). Therefore, if participants in our studies view *verbal* probabilities as indicating more epistemic uncertainty and *numeric* probabilities as indicating more aleatory uncertainty, this may be driving the tendency to make increasingly extreme judgments as participants see more verbal probability forecasts from advisors. However, in Study S5, there was no difference in participants' ratings of epistemicness between the verbal (M = 4.76) and numeric (M = 4.95) conditions, t(401) = 1.35, *p* = .18, making it unlikely that this was a factor in participants' combination strategies.

Inside vs. outside view. To the extent that the use of percentages may have caused participants to think about base rates, it may be that seeing numeric probabilities prompts participants take the "outside view" (i.e., they think about an average stock or the stock market as a whole), while verbal probabilities cause participants to take the "inside view" (i.e., they think about the specific stock; Dunning, 2007). If taking the inside view causes participants to make "overly optimistic" forecasts (Kahneman & Lovallo, 1993), then this might be driving participants to make more extreme forecasts in the *verbal* conditions. However, in Study S5, there was no difference in the extent to which participants felt the advisors were thinking about a specific stock or stocks on average between the verbal (M = 4.80) and numeric (M = 4.83) conditions, t(400) = .08, *p* = .94, again making it unlikely that this is driving our effect.

Advisor thoughtfulness. We also considered the possibility that more precise numeric forecasts indicated that the advisors put more thought into their forecasts, causing participants to weigh the advisors' forecasts more heavily. This would make participants more likely to indicate agreement with the advisors and select their exact forecast in the numeric condition compared to the verbal conditions. In Study S5, participants thought that advisors in the numeric condition (M = 5.42) put in slightly more thought than those in the verbal condition (M = 5.22), t(400) = 1.73, *p* = .08. However, this does not mediate our effect (see Table S10 in Supplement 3).

Facts vs. opinions. Finally, numeric forecasts might appear to represent a more objective truth about the stock, while verbal forecasts might represent the advisors' subjective opinions. If participants believe that advisors' forecasts are based on facts, any additional advisor's forecast would not provide much additional information on top of the first. In Study S5, participants rated on a 7 point scale whether the advisors based their predictions more on facts (1) or opinions (7).

There was no difference in ratings across the verbal (M = 3.50) and numeric (M = 3.37) conditions, t(400) = .96, *p* = .34.

Taken together, the results from our supplementary analyses do not provide strong evidence for one specific mechanism. It is possible, however, that people's combination strategies are multiply determined.

Further research and practical implications

In our studies, we tested how people combine forecasts from two advisors. However, it is reasonable to ask whether people continue to make even more extreme forecasts when they see more than two advisor forecasts in verbal format. We tested this in an additional study (Study S1 in Supplement 3) in which we provided participants with forecasts from five advisors and asked them to make their own forecast after each of the advisor's forecasts. In this study, we find a similar pattern of results, suggesting that our findings hold when there are more than two advisors. Nevertheless, we encourage future work to further investigate how an increasing number of advisors influences people's combination strategies.

We also focused on numerical probabilities in the form of percentages. However, numeric probabilities can also be represented as frequencies, such as "3 times out of 5" instead of 60%. Prior research suggests that people combine frequency forecasts differently than they combinepercentages. For example, when dealing with frequency data, people are more likely to take a Bayesian approach (Biswas, Zhao, & Lehmann, 2011; Gigerenzer & Hoffrage, 1995; Hoffrage, Lindsey, Hertwig, & Gigerenzer, 2000), although there is disagreement about why this is the case. Frequencies may be processed more intuitively than percentages are, making it easier for people to make normative inferences (Gigerenzer & Hoffrage, 1995). However, Biswas et al. (2011) find that certain frequencies may be more difficult to combine, causing people to view new frequencies as simply confirming or disconfirming their existing beliefs and adjusting their confidence accordingly. Future work could investigate how people's combination strategies for different forms of numeric probabilities compare to what we found in the present work.

Further, probability judgments are not the only type of values that can be expressed numerically or verbally. For example, a product's quality could be rated as "4 out of 5" or as "very good." As a result, consumers may combine two "4 out of 5" reviews differently than they combine two "very good" reviews. We do not expect these strategies to necessarily follow the same pattern observed here (i.e., counting verbal reviews) because, unlike with probability judgments, there does not seem to be a normative reason to do so. Examining this further could shed additional light on possible psychological mechanisms behind our effect.

We also focus on a relatively limited subset of verbal probabilities (e.g., somewhat or rather likely), primarily for internal consistency and to build on prior research (e.g., Windschitl & Weber, 1999). Of course, there is a wide variety of verbal probabilities that one can use (Beyth-Marom, 1982; Lichtenstein & Newman, 1967) that vary in their precision and direction. For instance, "more likely than not" has direction, but little precision, "nearly certain" has both precision and direction, and "possible" has neither. We would expect our effects to hold primarily when the verbal probabilities have a relatively unambiguous direction, although we cannot say this definitively given our results.

Finally, our studies exclusively examine individual decision making. However, it would be interesting to test whether these results hold in a group decision making context. For instance, the use of verbal rather than numeric probabilities in group discussions may exacerbate group polarization (e.g., Isenberg, 1986; Lord, Ross, & Lepper, 1979; Stoner, 1968).

Our results also have important practical implications. First, individuals and organizations that rely on probability forecasts from advisors to make decisions should be aware that the format of these forecasts can influence how decision makers combine these forecasts to arrive at their own judgments. Importantly, the current guidelines issued by the Director of National Intelligence (2015) and Intergovernmental Panel on Climate Change (Mastrandrea et al., 2011) do not account for this possibility and thus seem incomplete. Similarly, our research is also relevant to organizations that provide aggregated forecasts to others. Websites such as The Upshot[10], which typically presents multiple election predictions in a table with numeric (e.g., "65% Democrat") and verbal (e.g., "Lean Democrat") side-by-side, and Climendo[11], which provides weather forecasts from multiple sources with attached verbal "certainty" ratings, should consider whether their presentation choices cause readers to make more or less accurate judgments overall.

In sum, our research suggests that people use different combination strategies when combining numeric versus verbal probability forecasts. Although people typically average numeric forecasts, they apply a counting strategy when combining verbal forecasts, making their own forecast more extreme than each advisor's forecast. Researchers and practitioners should be aware of these differences when investigating and using probability forecasts.

[10] http://www.nytimes.com/sections/upshot
[11] http://www.climendo.com

Appendix A. Links to preregistrations

Study 2: http://aspredicted.org/blind.php?x=dy2s5t
Study 3: http://aspredicted.org/blind.php?x=88hn67
Study 4: http://aspredicted.org/blind.php?x=yk6tr8
Study 5a: http://aspredicted.org/blind.php?x=je6r5p
Study 5b: http://aspredicted.org/blind.php?x=sv4b7n
Study 6: http://aspredicted.org/blind.php?x=3wn3se

Study S1: https://aspredicted.org/blind.php?x=js8q6m
Study S2: http://aspredicted.org/blind.php?x=3wn3se
Study S3: http://aspredicted.org/blind.php?x=v7mc8a
Study S4: http://aspredicted.org/blind.php?x=jx9mt9
Study S5: http://aspredicted.org/blind.php?x=8yq2nf

Appendix B. Study 4 Scenarios

Plane Ticket Scenario

You want to buy a plane ticket for a vacation you are taking. You found a ticket that fits your budget, but know prices can drop if you wait (although they can also go up or the flight could sell out). You're willing to wait up to two weeks. You check a price prediction website, which says the following:

"It is [somewhat/rather/55%/65%] likely that prices will drop in the next two weeks."

Would you buy the ticket or wait to see if the price goes down? (1 = Definitely buy, 7 = Definitely wait)

[page break]

Would you buy the ticket or wait to see if the price goes down?

Your previous answer: [participant's previous answer is displayed here]

You decide to get a second opinion and check a different site that also makes price predictions. The second site says:

"We think that it is [somewhat/rather/55%/65%] likely that prices will decrease within the next two weeks."

Would you buy the ticket or wait to see if the price goes down? (1 = Definitely buy, 7 = Definitely wait)

Study 4 Cell Phone Scenario

You want to buy a new cell phone. A phone you like is on sale, but you've heard rumors that a new model might come out soon. You prefer a newer model, so it could be worthwhile to wait a little bit, although you'd be upset if the rumors are wrong and the phone you like is no longer on sale. You look at a tech news website, which says the following:

"It is [somewhat/rather/55%/65%] likely that a new model will be released within the next month."

Would you buy the phone now or wait to see if a new model comes out? (1 = Definitely buy, 7 = Definitely wait)

[page break]

Would you buy the phone now or wait to see if a new model comes out?

Your previous answer: [participant's previous answer is displayed here]

You decide to get a second opinion and check a different tech news website. The second site says:

"It appears that it is [somewhat/rather/55%/65%] likely that the company will release a new version of the phone during the next month."

Would you buy the phone now or wait to see if a new model comes out? (1 = Definitely buy, 7 = Definitely wait)

References

Ariely, D., Tung Au, W., Bender, R. H., Budescu, D. V., Dietz, C. B., Gu, H., … Zauberman, G. (2000). The effects of averaging subjective probability estimates between and within judges. *Journal of Experimental Psychology: Applied*, *6*(2), 130–147. https://doi.org/10.1037/1076-898X.6.2.130

Ashton, A. H., & Ashton, R. H. (1985). Aggregating Subjective Forecasts: Some Empirical Results. *Management Science*, *31*(12), 1499–1508. https://doi.org/10.1287/mnsc.31.12.1499

Bagchi, R., & Ince, E. C. (2015). Is a 70% Forecast More Accurate Than a 30% Forecast? How Level of a Forecast Affects Inferences About Forecasts and Forecasters. *Journal of Marketing Research*, *53*(1), 31–45. https://doi.org/10.1509/jmr.12.0526

Baron, J., Mellers, B. A., Tetlock, P. E., Stone, E., & Ungar, L. H. (2014). Two Reasons to Make Aggregated Probability Forecasts More Extreme. *Decision Analysis*, *11*(2), 133–145. https://doi.org/10.1287/deca.2014.0293

Beyth-Marom, R. (1982). How probable is probable? A numerical translation of verbal probability expressions. *Journal of Forecasting*, *1*(3), 257–269. https://doi.org/10.1002/for.3980010305

Bilgin, B., & Brenner, L. (2013). Context affects the interpretation of low but not high numerical probabilities: A hypothesis testing account of subjective probability. *Organizational Behavior and Human Decision Processes*, *121*(1), 118–128. https://doi.org/10.1016/j.obhdp.2013.01.004

Biswas, D., Zhao, G., & Lehmann, D. R. (2011). The Impact of Sequential Data on Consumer Confidence in Relative Judgments. *Journal of Consumer Research*, *37*(5), 874–887. https://doi.org/10.1086/656061

Budescu, D. V., Broomell, S., & Por, H.-H. (2009). Improving Communication of Uncertainty in the Reports of the Intergovernmental Panel on Climate Change. *Psychological Science*, *20*(3), 299–308. https://doi.org/10.1111/j.1467-9280.2009.02284.x

Budescu, D. V., Por, H.-H., & Broomell, S. B. (2012). Effective communication of uncertainty in the IPCC reports. *Climatic Change*, *113*(2), 181–200. https://doi.org/10.1007/s10584-011-0330-3

Budescu, D. V., Por, H.-H., Broomell, S. B., & Smithson, M. (2014). The interpretation of IPCC probabilistic statements around the world. *Nature Climate Change*, *4*(6), 508–512. https://doi.org/10.1038/nclimate2194

Budescu, D. V., Weinberg, S., & Wallsten, T. S. (1988). Decisions based on numerically and verbally expressed uncertainties. *Journal of Experimental Psychology: Human Perception and Performance*, *14*(2), 281–294. https://doi.org/10.1037/0096-1523.14.2.281

Budescu, D. V., & Yu, H.-T. (2006). To Bayes or Not to Bayes? A Comparison of Two Classes of Models of Information Aggregation. *Decision Analysis*, *3*(3), 145–162. https://doi.org/10.1287/deca.1060.0074

Budescu, D. V., & Yu, H.-T. (2007). Aggregation of opinions based on correlated cues and advisors. *Journal of Behavioral Decision Making*, *20*(2), 153–177. https://doi.org/10.1002/bdm.547

Budescu, D. V., Zwick, R., Wallsten, T. S., & Erev, I. (1990). Integration of linguistic probabilities. *International Journal of Man-Machine Studies*, *33*(6), 657–676. https://doi.org/10.1016/S0020-7373(05)80068-9

Carlson, B. W., & Yates, J. F. (1989). Disjunction errors in qualitative likelihood judgment. *Organizational Behavior and Human Decision Processes*, *44*(3), 368–379. https://doi.org/10.1016/0749-5978(89)90014-9

Dunning, D. (2007). Prediction: The inside view. In *Social psychology: Handbook of basic principles, 2nd ed* (pp. 69–90). New York, NY, US: Guilford Press.

Erev, I., & Cohen, B. L. (1990). Verbal versus numerical probabilities: Efficiency, biases, and the preference paradox. *Organizational Behavior and Human Decision Processes*, *45*(1), 1–18. https://doi.org/10.1016/0749-5978(90)90002-Q

Gigerenzer, G., & Hoffrage, U. (1995). How to improve Bayesian reasoning without instruction: Frequency formats. *Psychological Review*, *102*(4), 684–704. https://doi.org/10.1037/0033-295X.102.4.684

González-Vallejo, C. C., Erev, I., & Wallsten, T. S. (1994). Do Decision Quality and Preference Order Depend on Whether Probabilities Are Verbal or Numerical? *The American Journal of Psychology*, *107*(2), 157–172. https://doi.org/10.2307/1423035

González-Vallejo, C. C., & Wallsten, T. S. (1992). Effects of probability mode on preference reversal. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *18*(4), 855–864. https://doi.org/10.1037/0278-7393.18.4.855

Hoffrage, U., Lindsey, S., Hertwig, R., & Gigerenzer, G. (2000). Communicating Statistical Information. *Science*, *290*(5500), 2261–2262. https://doi.org/10.1126/science.290.5500.2261

Hsee, C. K. (1996). The Evaluability Hypothesis: An Explanation for Preference Reversals between Joint and Separate Evaluations of Alternatives. *Organizational Behavior and Human Decision Processes*, *67*(3), 247–257. https://doi.org/10.1006/obhd.1996.0077

Isenberg, D. J. (1986). Group polarization: A critical review and meta-analysis. *Journal of Personality and Social Psychology*, *50*(6), 1141–1151. https://doi.org/10.1037/0022-3514.50.6.1141

Kahneman, D., & Lovallo, D. (1993). Timid Choices and Bold Forecasts: A Cognitive Perspective on Risk Taking. *Management Science*, *39*(1), 17–31. https://doi.org/10.1287/mnsc.39.1.17

Lichtenstein, S., & Newman, J. R. (1967). Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychonomic Science*, *9*(10), 563–564. https://doi.org/10.3758/BF03327890

Lord, C. G., Ross, L., & Lepper, M. R. (1979). Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *Journal of Personality and Social Psychology*, *37*(11), 2098–2109. https://doi.org/10.1037/0022-3514.37.11.2098

Mastrandrea, M. D., Mach, K. J., Plattner, G.-K., Edenhofer, O., Stocker, T. F., Field, C. B., … Matschoss, P. R. (2011). The IPCC AR5 guidance note on consistent treatment of uncertainties: a common approach across the working groups. *Climatic Change*, *108*(4), 675. https://doi.org/10.1007/s10584-011-0178-6

Mellers, B., Stone, E., Murray, T., Minster, A., Rohrbaugh, N., Bishop, M., … Tetlock, P. (2015). Identifying and Cultivating Superforecasters as a Method of Improving Probabilistic Predictions: *Perspectives on Psychological Science*. https://doi.org/10.1177/1745691615577794

Office of the Director of National Intelligence. (2015). *Intelligence Community Directive 203: Analytic Standards*. Washington, DC. Retrieved from https://www.dni.gov/files/documents/ICD/ICD%20203%20Analytic%20Standards.pdf

Rapoport, A., Wallsten, T. S., Erev, I., & Cohen, B. L. (1990). Revision of opinion with verbally and numerically expressed uncertainties. *Acta Psychologica*, *74*(1), 61–79. https://doi.org/10.1016/0001-6918(90)90035-E

Sah, S., & Loewenstein, G. (2015). Conflicted advice and second opinions: Benefits, but unintended consequences. *Organizational Behavior and Human Decision Processes*, *130*, 89–107. https://doi.org/10.1016/j.obhdp.2015.06.005

Sarvary, M. (2002). Temporal Differentiation and the Market for Second Opinions. *Journal of Marketing Research*, *39*(1), 129–136. https://doi.org/10.1509/jmkr.39.1.129.18933

Schwartz, J., Luce, M. F., & Ariely, D. (2011). Are Consumers Too Trusting? The Effects of Relationships with Expert Advisers. *Journal of Marketing Research*, *48*(SPL), S163–S174. https://doi.org/10.1509/jmkr.48.SPL.S163

Stoner, J. A. F. (1968). Risky and cautious shifts in group decisions: The influence of widely held values. *Journal of Experimental Social Psychology*, *4*(4), 442–459. https://doi.org/10.1016/0022-1031(68)90069-3

Tannenbaum, D., Fox, C. R., & Ülkümen, G. (2016). Judgment Extremity and Accuracy Under Epistemic vs. Aleatory Uncertainty. *Management Science*, *63*(2), 497–518. https://doi.org/10.1287/mnsc.2015.2344

Teigen, K. H. (2001). When Equal Chances = Good Chances: Verbal Probabilities and the Equiprobability Effect. *Organizational Behavior and Human Decision Processes*, *85*(1), 77–108. https://doi.org/10.1006/obhd.2000.2933

Teigen, K. H., & Brun, W. (1995). Yes, but it is uncertain: Direction and communicative intention of verbal probabilistic terms. *Acta Psychologica*, *88*(3), 233–258. https://doi.org/10.1016/0001-6918(93)E0071-9

Teigen, K. H., & Brun, W. (1999). The Directionality of Verbal Probability Expressions: Effects on Decisions, Predictions, and Probabilistic Reasoning. *Organizational Behavior and Human Decision Processes*, *80*(2), 155–190. https://doi.org/10.1006/obhd.1999.2857

Teigen, K. H., & Brun, W. (2003). Verbal probabilities: a question of frame? *Journal of Behavioral Decision Making*, *16*(1), 53–72. https://doi.org/10.1002/bdm.432

Wallsten, T. S., Budescu, D. V., Erev, I., & Diederich, A. (1997). Evaluating and Combining Subjective Probability Estimates. *Journal of Behavioral Decision Making*, *10*(3), 243–268. https://doi.org/10.1002/(SICI)1099-0771(199709)10:3<243::AID-BDM268>3.0.CO;2-M

Wallsten, T. S., Budescu, D. V., & Tsao, C. J. (1997). Combining linguistic probabilities. *Psychologische Beitrage*, *39*(1–2), 27–55.

Wallsten, T. S., & Diederich, A. (2001). Understanding pooled subjective probability estimates. *Mathematical Social Sciences*, *41*(1), 1–18. https://doi.org/10.1016/S0165-4896(00)00053-6

West, P. M., & Broniarczyk, S. M. (1998). Integrating Multiple Opinions: The Role of Aspiration Level on Consumer Response to Critic Consensus. *Journal of Consumer Research*, *25*(1), 38–51. https://doi.org/10.1086/209525

Windschitl, P. D., & Weber, E. U. (1999). The interpretation of “likely” depends on the context, but “70%” is 70%--right? The influence of associative processes on perceived certainty.

Journal of Experimental Psychology. Learning, Memory, and Cognition, *25*(6), 1514–1533.

Windschitl, P. D., & Wells, G. L. (1996). Measuring psychological uncertainty: Verbal versus numeric methods. *Journal of Experimental Psychology: Applied*, *2*(4), 343–364. https://doi.org/10.1037/1076-898X.2.4.343

Windschitl, P. D., & Wells, G. L. (1998). The alternative-outcomes effect. *Journal of Personality and Social Psychology*, *75*(6), 1411–1423. https://doi.org/10.1037/0022-3514.75.6.1411

Yaniv, I. (2004). The Benefit of Additional Opinions. *Current Directions in Psychological Science*, *13*(2), 75–78.

Yaniv, I., & Milyavsky, M. (2007). Using advice from multiple sources to revise and improve judgments. *Organizational Behavior and Human Decision Processes*, *103*(1), 104–120. https://doi.org/10.1016/j.obhdp.2006.05.006

Yates, J. F., & Carlson, B. W. (1986). Conjunction errors: Evidence for multiple judgment procedures, including “signed summation.” *Organizational Behavior and Human Decision Processes*, *37*(2), 230–253. https://doi.org/10.1016/0749-5978(86)90053-1

Zimmer, A. C. (1983). Verbal Vs. Numerical Processing of Subjective Probabilities. In R. W. Scholz (Ed.), *Advances in Psychology* (Vol. 16, pp. 159–182). North-Holland. https://doi.org/10.1016/S0166-4115(08)62198-6