Management Science

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

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

Robert Mislavsky, Celia Gaertig

To cite this article:
Robert Mislavsky, Celia Gaertig (2021) Combining Probability Forecasts: 60% and 60% Is 60%, but Likely and Likely Is Very Likely. Management Science

Published online in Articles in Advance 25 Mar 2021

. https://doi.org/10.1287/mnsc.2020.3902

http://pubsonline.informs.org/journal/mnsc

MANAGEMENT SCIENCE
Articles in Advance, pp. 1–23
ISSN 0025-1909 (print), ISSN 1526-5501 (online)

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

Robert Mislavsky,[a] **Celia Gaertig**[b]

[a] Carey Business School, Johns Hopkins University, Baltimore, Maryland 21202; [b] Booth School of Business, University of Chicago, Chicago, Illinois 60637

Contact: mislavsky@jhu.edu, https://orcid.org/0000-0002-9620-3528 (RM); celia.gaertig@chicagobooth.edu, https://orcid.org/0000-0002-5444-4999 (CG)

Received: September 6, 2019
Revised: June 11, 2020; September 8, 2020
Accepted: October 1, 2020
Published Online in Articles in Advance: March 25, 2021

https://doi.org/10.1287/mnsc.2020.3902

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 eight studies ($n = 7{,}334$), we find that people are more likely to act as if they "count" verbal probabilities (i.e., they move closer to certainty than any individual advisor's forecast) than they are to "count" numeric probabilities. 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 discuss and rule out several candidate mechanisms for our effect.

History: Accepted by Yuval Rottenstreich, decision analysis.
Funding: Financial support for this research was provided by the Wharton Behavioral Lab, Johns Hopkins Carey Business School, and the Beatrice Foods Co. Faculty Research Fund at the University of Chicago Booth School of Business.
Supplemental Material: The data files and online appendix are available at https://doi.org/10.1287/mnsc.2020.3902.

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, or they may even consult one of a growing number of prediction websites.

Advisors, for their part, may choose to make their forecasts numerically (e.g., "There is a *60% chance* that prices will increase.") or verbally (e.g., "It is *likely* that prices will increase."), which can itself have consequences for how these forecasts are interpreted. A substantial body of research has shown that numeric and verbal probabilities are interpreted differently in isolation (e.g., Lichtenstein and Newman 1967, Beyth-Marom 1982, Windschitl and Wells 1996). However, given that people often prefer to get multiple opinions before acting (West and Broniarczyk 1998, Sarvary 2002, Schwartz et al. 2011, Sah and Loewenstein 2015) and that there are substantial benefits to aggregating multiple opinions (Yaniv 2004, Yaniv and 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 and Yu 2006) and optimal strategies for doing so (Ashton and Ashton 1985, Wallsten et al. 1997b, Ariely et al. 2000, Baron et al. 2014). However, although verbal probabilities are used much more often (Zimmer 1983, Budescu et al. 1988, Erev and Cohen 1990), there has been little research on how people combine multiple verbal probabilities.

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 people act as if they average numeric probabilities. However, we find that people act as if they "count" verbal probabilities, increasing the likelihood that their own forecast is *more extreme* than each of the advisors' forecasts.

To illustrate this, 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, your own forecast would be close to the average expert forecast (i.e., close to 60%). 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 own prediction and believing that a price increase is "very likely."

Combining Others' Probability Estimates

Although past research has significantly advanced our understanding of the differences between how numeric and verbal probabilities are interpreted (e.g., Budescu et al. 1988; Erev and Cohen 1990; Teigen and Brun 1995, 1999; Windschitl and Wells 1996; Windschitl and Weber 1999), surprisingly little is known about how people *combine* multiple probability estimates to form their own judgments, particularly for verbal probabilities.

Combining Numeric Probabilities

When we discuss numeric probabilities throughout this paper, we are primarily referring to percentages (e.g., "60%"). Consistent with the results from our studies, Budescu and Yu (2006, 2007) find that people typically average others' forecasts when these forecasts are provided to them as percentages. However, some researchers have also found that people sometimes 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 and Yu 2006) or when participants are making predictions about knowledge-based (i.e., epistemic) uncertainty (Wallsten et al. 1997a). Such a strategy may be normative if advisors have imperfectly correlated information (Wallsten et al. 1997b, Ariely et al. 2000, Baron et al. 2014), but there is little evidence that people take this into consideration when making their own judgments (Budescu and Yu 2007).

Combining Verbal Probabilities

In contrast to research on combining numeric probabilities, little research has examined how people combine others' verbal probability estimates for a future outcome. Budescu, Wallsten, and colleagues have conducted the only research that we are aware of on this question. In Budescu et al. (1990), participants saw two verbal probability estimates from hypothetical advisors that described how likely it was that a spinner would land on a certain color. Participants' own estimates of that probability were both more extreme and "less fuzzy" than the advisors' estimates. This provides preliminary evidence for a "counting" strategy for verbal probabilities. In this study, however, there is no equivalent numeric probability condition nor is there a condition in which participants see an estimate from only one advisor. As a result, it may be that participants made more extreme forecasts because they saw verbal forecasts, because they saw forecasts from two advisors, or simply because of the nature of the task itself.

In addition, Wallsten et al. (1997a) found that participants became more confident in their answers for knowledge questions when they saw multiple advisors indicate their confidence verbally (e.g., "It is certain that London is more populous than Paris.") compared with numerically (e.g., "There is a 90% chance that London is more populous than Paris."). In this case, there is a comparison between numeric and verbal probabilities, but there is no comparison of participant forecasts with the advisors' forecasts, only a comparison of overall confidence. In addition, similar to Budescu et al. (1990), there is no condition where participants only saw one advisor's estimate, so we cannot tell if this result is caused by people becoming more confident when they see verbal forecasts more generally, regardless of the number of advisors.

In the current research, we build on this existing work by directly contrasting participants' behavior both for numeric and verbal probabilities and for one or two advisors across a variety of future events (e.g., future stock prices, the outcome of baseball games).

Normative Combination Strategies

Given that we are chiefly interested in understanding how people combine others' probability forecasts, a natural question to ask is: "How *should* we combine probability forecasts?" The unsatisfying answer is, "it depends." Here, we discuss when and why it may be optimal to average probability estimates and when and why it may be optimal to "count" them, and we also provide examples of each. Finally, we discuss what the optimal strategies are in our studies (if there are any).

Although this paper focuses on differences in the use of averaging and "counting" strategies, these are not the only strategies that people could use to combine forecasts. For example, people may use a "hedging" strategy, where they agree with the general direction of an outcome (i.e., they are on the same side of 50% as the advisors) but are less certain than the advisors (i.e., they are closer to 50%). Alternatively, they may use a "contrarian" strategy by which they disagree with a consensus opinion if they believe there is a rational reason to do so (e.g., if they suspect irrational herding of advisors). However, these strategies are not common in our data, and more importantly, do not differ across formats in our studies. Therefore, we limit our investigation to averaging and "counting."

When Should People Average?

The benefits of averaging opinions are well documented. Galton (1907) first showed that the "error of

the average" (i.e., the difference between the average estimate and the truth) was substantially smaller than the "average error" (i.e., the average of the differences between individual estimates and the truth) among a group of people attempting to guess the weight of an ox. Although Galton did not speculate why this difference emerged, subsequent researchers have found the cause to be quite simple (e.g., Stroop 1932, Larrick and Soll 2006). Specifically, when estimating a value, people will have idiosyncratic biases or commit relatively random errors. Put simply, some people will overestimate and some people will underestimate the true answer. Averaging those estimates will therefore put you between these over- and underestimates and thus, closer to the truth. As a result, averaging the individual estimates will reduce the impact of these biases and errors and improve upon most individual estimates. Even if these errors are systematic (e.g., everyone overestimates), averaging estimates will still be at least as accurate as choosing a random individual estimate (see Larrick and Soll 2006 for a more detailed review).

Although the benefits of averaging are typically discussed in the context of unknown amounts (e.g., weights, sales projections), they often hold for unknown *probabilities*. Because individual advisors still have idiosyncratic biases and errors, averaging their probability estimates often improves accuracy (Ashton and Ashton 1985; Wallsten et al. 1997b, p. 199; Wallsten and Diederich 2001).

When Should People "Count?"

Although averaging often improves accuracy over individual estimates, it may not always be the optimal strategy. The underlying assumption behind the benefits of averaging is that the advisors have *similar information*. If everyone sees the same ox before guessing its weight, then the primary source of error between advisors is essentially random error or idiosyncratic bias. However, this is not always the case. Condorcet (1785) first theorized that as long as individual jurors have a greater than 50% chance of making a correct decision, adding an additional juror increases the likelihood that the jury, as a whole, will vote correctly. However, this only holds if the jurors' judgments are relatively uncorrelated (e.g., Berg 1993). The parable of the blind men and the elephant is illustrative here. If each individual man is only touching one part of the elephant, he cannot form a complete understanding of the whole without using information from the others. That is, when people have *different information*, a combination of their judgments gives a more complete picture of an event's likelihood (Wallsten and Diederich 2001, Baron et al. 2014).

Budescu and Yu (2007) provide an elegant demonstration of this concept, which we simplify slightly here. Imagine that you want to know how likely it is that a patient has a particular disease, and there are two different tests for the disease. After seeing the results of Test A, Dr. Jones believes that there is a 60% chance that the patient has the disease. After seeing the results of Test B, Dr. Smith also believes that there is a 60% chance. However, if you know *both* doctors' predictions, you now know that the results of both tests point to the patient likely having the disease. Therefore, you have "the right to much higher confidence" (Baron et al. 2014, p. 134) than each individual doctor who only saw one test, and you should estimate that the true probability is higher than 60%.[1]

What Is Normative in Our Studies?

In our studies, the only difference between the advice that we provide to participants across conditions is whether it is presented numerically (e.g., "60%") or verbally (e.g., "rather likely"). With one exception (Study 5B), we do not manipulate or provide any information about whether advisors are using similar or dissimilar information in our studies. Given that we do not provide any information to participants about how similar or correlated the advisors' information is, participants should not be able to infer whether averaging or "counting" is the normative strategy. However, if the perceived precision of numeric forecasts causes participants to believe that two similar forecasts are from advisors who are using similar information, they may think it is appropriate to use the more normative averaging strategy. In contrast, they might not have this belief for similar verbal forecasts because verbal forecasts are typically perceived as more vague (Lichtenstein and Newman 1967, Zimmer 1983). As a result, participants may believe that a "counting" strategy is appropriate. In Studies 5A, 5B, 7, and S5, we test whether inferences about the similarity of the information on which the advisors base their forecasts influence participants' combination strategies. We find some suggestive evidence supporting this account in Study S5, but we do not replicate it in Studies 5A, 5B, or 7 (see Tables 1 and 2). Thus, although this mechanism may influence participants' combination strategies to some extent, it does not seem to be a particularly meaningful influence.

Comparing Combination Strategies

In this article, we examine how people combine multiple advisors' 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. We should be clear, however, that by using these terms, we mean that

Table 1. Exploratory Mechanisms Tested Across Studies: Measured Mechanisms in Studies 2, 5A, 7, S2, and S5

Measured mechanisms	Study	Numeric *M*	Verbal *M*	Main effect of verbal vs. numeric	Mediation analysis	
					Indirect effect	95% CI
Number of advisors agreeing of 100	2 (above midpoint)	67.27	72.73	$t(391) = 3.88$, $p < 0.001$	–0.063	(–0.19, 0.06)
	2 (below midpoint)	41.57	33.21	$t(390) = -3.60$, $p < 0.001$	0.001	(–0.04, 0.03)
Extent of advisors' agreement	2 (above midpoint)	5.80	6.11	$t(391) = 3.01$, $p = 0.003$	–0.015	(–0.08, 0.05)
	2 (below midpoint)	3.99	5.13	$t(390) = 5.78$, $p < 0.001$	–0.001	(–0.02, 0.02)
	5A	5.90	5.92	$t(791) = 0.38$, $p = 0.71$		—
	7	4.77	4.74	$t(389) = -0.01$, $p = 0.99$		—
Second advisor forecast provides new information	S5	2.37	2.92	$t(796) = 4.56$, $p < 0.001$	**0.052**	**(0.004, 0.106)**
Second advisor forecast is useful	S5	4.07	4.63	$t(797) = 4.75$, $p < 0.001$	0.063	(–0.005, 0.133)
Participant weighted second advisor forecast more	S5	3.95	3.98	$t(796) = 0.55$, $p = 0.58$		—
Advisor forecasts provide different (vs. same) information	S5	2.65	3.08	$t(796) = 4.46$, $p < 0.001$	**0.045**	**(0.007, 0.088)**
	5A	2.78	2.90	$t(791) = 0.95$, $p = 0.34$		—
	7	2.88	3.10	$t(390) = 1.77$, $p = 0.078$	–0.012	(–0.09, 0.05)
Advisor forecasts are based on knowledge (vs. luck)	7	4.95	4.76	$t(390) = -1.35$, $p = 0.18$		—
Advisor forecasts are based on inside view (vs. outside view)	7	4.83	4.80	$t(389) = -0.08$, $p = 0.94$		—
Advisor forecasts are based on opinions (vs. facts)	7	3.37	3.50	$t(389) = 0.96$, $p = 0.34$		—
Advisors' thoughtfulness	7	5.42	5.22	$t(389) = -1.73$, $p = 0.08$	–0.026	(–0.13, 0.07)
Advisors' strength of opinion	7	4.65	4.36	$t(389) = -2.05$, $p = 0.041$	–0.043	(–0.15, 0.05)
Participant's confidence in own forecast	S2	4.45	4.59	$t(392) = 0.90$, $p = 0.37$		—
Participant's use of intuition (vs. reason) for own forecast	S2	5.01	4.88	$t(392) = -0.98$, $p = 0.33$		—

Notes. All mechanisms were measured using seven-point scales, except for the measure that asked about the number of advisors agreeing of 100. Across studies, the wording of questions measuring the same mechanism differed slightly. We provide the exact wording of the mediator questions in each study's method section. We used OLS regression to regress each of the mediator questions on the advice format condition (1 = verbal, 0 = numeric), including fixed effects for stimulus (stock or scenario). For those measures for which we found a significant or marginally significant difference between the numeric and verbal conditions, we performed bootstrapped mediation analyses using 10,000 samples. Bold in the mediation analysis results indicates that the respective mechanism partially mediated the effect.

Table 2. Exploratory Mechanisms Tested Across Studies: Manipulated Mechanisms in Studies 5B and 6

Manipulated mechanisms	Study	Description of the results (see Studies 5B and 6 for details)
Advisor forecasts provide different (vs. same) information	5B	Showing that advisors are using similar vs. dissimilar information does not moderate effect. No significant interaction between information source and number of advisors, $p = 0.79$, nor between advice format, information source, and number of advisors, $p = 0.34$
Direction of forecast emphasized	6	Indicating to participants that a forecast is a "good" (vs. "bad") sign does not moderate effect. No significant interaction between the *numeric* and the *numeric with direction* conditions, $p = 0.126$

participants act in a way that is consistent with averaging and counting, although they may, in fact, be using a heuristic that does not rely on a specific mathematical operation.

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 1–5 indicate that the event is unlikely (i.e., below 50% chance) to occur and points 6–10 indicate that the event is likely (i.e., above 50% chance). Now imagine that the participant receives two advisors' 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 her own forecast in between the advisors' forecasts. Because the advisors' forecasts were 7 and 9, a participant who averages these forecasts would answer somewhere at or between the two advisors' forecasts (i.e., 7, 8, or 9). What exact forecast the participant will make depends, 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 act as if they are interpreting 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 advisors' 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 forecast is consistent with a counting strategy but not with an averaging strategy.

Overview of Studies

We conducted eight studies in which we asked participants to forecast the likelihood of an uncertain event (Studies 1–3, 5A, 5B, 6, and 7) 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 always qualitatively agreed on the outcome (i.e., the forecasts were in the same "direction"). For example, in Study 1, both advisors predicted that a stock price was more likely to increase than it was to decrease. 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 expert predicts that an event is "likely" to occur and another expert predicts that an event is "unlikely" to occur, then both counting and averaging strategies would cause a participant to say something like "neither likely nor unlikely" or "even chance." Focusing on forecasts that are in the same direction allows us to distinguish between counting and averaging strategies. Participants always made their own predictions on the same scale that the advisors 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 act as if they are counting, we would expect the proportion of participants making extreme forecasts to *increase* when they see the second

advisor's forecast. However, if participants act as if they are averaging, we predict that the proportion of participants making extreme forecasts will *decrease* when they see the second advisor's forecast. Therefore, if participants use different combination strategies across verbal and numeric formats, our key prediction would be that there is an interaction between *advice format* and the *number of advisors*.

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 with an increasing number of advisors because the amount of new information provided by, say, the 30th 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 in the online appendix 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 evidence that participants act as if they are averaging numeric forecasts and counting verbal ones, causing them to make extreme forecasts more often in the verbal condition. These results hold for probabilities above the scale midpoint (all studies) and below the scale midpoint (Study 2), for hypothetical scenarios (Studies 1, 2, and 5–7) and real events (Studies 3 and 4), and when presenting multiple probability forecasts simultaneously (Studies 1 and 3) and sequentially (Studies 2 and 4–7). Furthermore, Study 4 demonstrates that these different strategies meaningfully influence an incentivized decision. Studies 5–7 and S2–S4 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 3 in the online appendix 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://researchbox.org/62. Links to preregistrations are also listed in the appendix.

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 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 that were given either numerically (e.g., "60%–69%") 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 how likely it was that 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 10. 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* versus *verbal* advice format. Figure 1 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 and Weber 1999). The advisors' forecasts in both the numeric and the verbal conditions corresponded to the seventh point on participants' response scales, keeping the extremity of the advisors' forecasts constant across conditions.

Results

Main Analysis. We determine participants' combination strategies by assessing the proportion of participants who made *extreme* forecasts across the two conditions. An *extreme* forecast is a participant forecast that is closer to certainty than any of the advisors' forecasts. In this study, because both advisors always provided forecasts that corresponded to the seventh 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 probit regression 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. Unless stated otherwise, we use probit regressions in all studies where our dependent variable is a binary outcome (e.g., whether a participant made an extreme forecast). Our results are nearly identical when using binary logistic or ordinary least squares (OLS) regression. We report the additional regression results for all three models in Supplement 3 in the online appendix.

Figure 1. Sample Stimuli and Response Scale in Study 1

Example Stock Stimulus in Study 1

Information about the Stock:

Ticker Symbol	Company Name	Price
GIL	Gildan Activewear, Inc.	$30.00

Numeric Format Condition: Analyst predictions and response scale

How likely is it that GIL will close *above* $30.00 on July 11, 2017?

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

How likely do you think it is that GIL will close *above* $30.00 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 GIL will close *above* $30.00 on July 11, 2017?

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

How likely do you think it is that GIL will close *above* $30.00 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
○	○	○	○	○	○	○	○	○	○

Note. The advisors' forecasts always corresponded to the seventh point on the response scale.

More participants in the *verbal* condition made extreme forecasts (29.4%) than in the *numeric* condition (10.9%), $Z = 3.29$, $p = 0.001$. This result suggests that participants are more likely to "count" when combining verbal forecasts than when combining numeric forecasts.

Does Seeing Verbal Forecasts Simply Make Participants' Forecasts Noisier? Our main analysis focused on comparing the proportion of participants making extreme predictions across conditions. However, one alternative account for our effect is that participants' own forecasts simply become *noisier* when they see verbal forecasts from two advisors than when they see numeric forecasts. For example, it may be that participants in the verbal condition are becoming not only more likely to make forecasts *above* the advisors' forecasts (i.e., above seven) but also more likely to make forecasts *below* the advisors' forecasts (i.e., below seven). However, we do not see this in the data. A similar proportion of participants provided forecasts *below* the advisors' forecasts in the *verbal* (28.4%) and *numeric* (25.7%) conditions, $Z = 0.37$, $p = 0.71$. This pattern holds in all of our studies, and we report the relevant comparisons in footnotes throughout the remainder of the article.

Discussion

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's forecast. Given that we presented participants in Study 1 with the forecasts from two advisors simultaneously, it is possible that participants' own forecasts in the verbal condition are already more extreme if they only see one advisor's forecast. In Study 2, we rule out this possibility.

Study 2: Sequential Evaluation (and Probabilities Below the Midpoint)

In Study 2, we show participants two forecasts sequentially and compare the proportion of participants who make extreme forecasts after seeing only one advisor's forecast with the proportion of participants who make extreme forecasts after seeing two advisors' forecasts. If people are more likely to act as if they are counting in the verbal condition, 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. We also 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 included 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 advisors' 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 the advisors' forecasts were on the seventh point on the response scale (i.e., "60%–69%" or "rather likely") or the fourth point on the scale (i.e., "30%–39%" or "rather unlikely").

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

After participants made their forecasts, we asked them two exploratory questions aimed at testing participants' perceptions of the advisors' agreement as a potential mechanism. The first question asked participants to indicate the extent to which they thought the advisors agree on their prediction ("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). The second question asked participants to indicate the number of advisors who would agree on the same forecast ("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).

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.

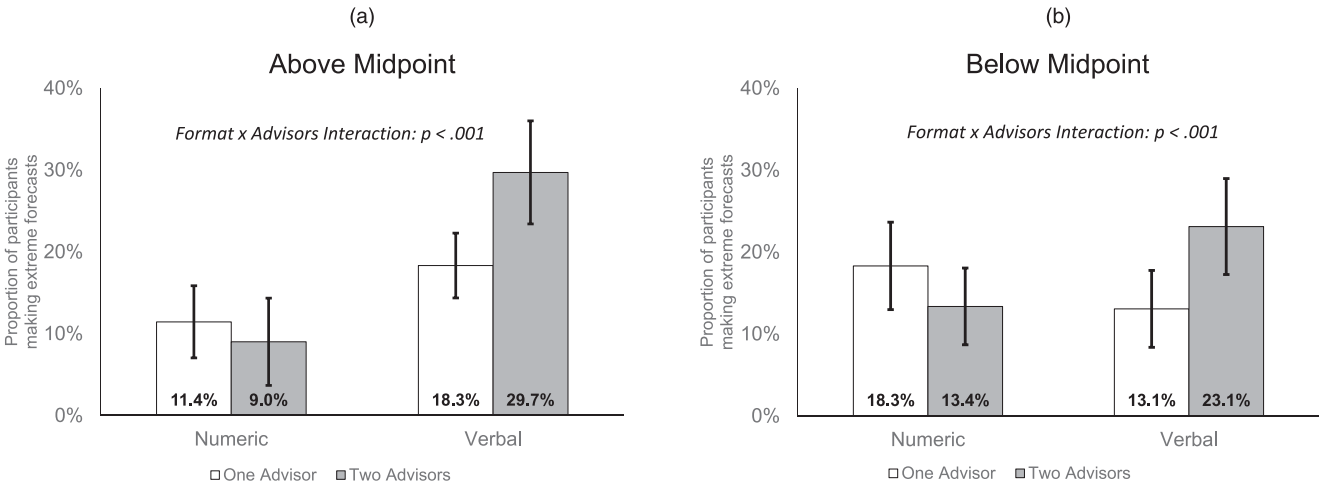
Figure 2 shows the main results for Study 2. For both the above and below midpoint conditions, participants became more likely to make extreme forecasts as they saw an additional advisor's forecast in the *verbal* condition, but they became less likely to do so in the *numeric* condition. For probabilities *above the midpoint*, there was a significant interaction between advice format and number of advisors, $Z = 3.62$, $p < 0.001$ (Figure 2(a)). When the advisors provided *verbal* probability forecasts, the proportion of participants making extreme forecasts increased from 18.3% after the first advisor to 29.7% after the second advisor, $Z = 4.13$, $p < 0.001$. In contrast, when the advisors made *numeric* probability forecasts, the proportion of extreme forecasts directionally decreased from 11.4% to 9.0%, $Z = -1.24$, $p = 0.21$.

This pattern also held for forecasts *below the midpoint* of the response scale (Figure 2(b)). There was again a significant interaction between advice format and number of advisors, $Z = 4.14$, $p < 0.001$. When the advisors provided *verbal* probability forecasts, the proportion of participants making extreme forecasts increased from 13.1% after the first advisor to 23.1% after the second advisor, $Z = 3.28$, $p = 0.001$. In contrast, when the advisors provided *numeric* probability forecasts, the proportion of extreme forecasts decreased from 18.3% to 13.4%, $Z = -2.46$, $p = 0.014$.[3]

Combining the *above* and *below* midpoint conditions into one analysis, the interaction between format and number of advisors was also significant, $Z = 3.99$, $p < 0.001$. There was no significant three-way interaction between advice format, number of advisors, and the above/below midpoint conditions, $Z = -0.41$, $p = 0.68$, indicating that the effect size does not substantially change when participants see forecasts on different sides of the midpoint.

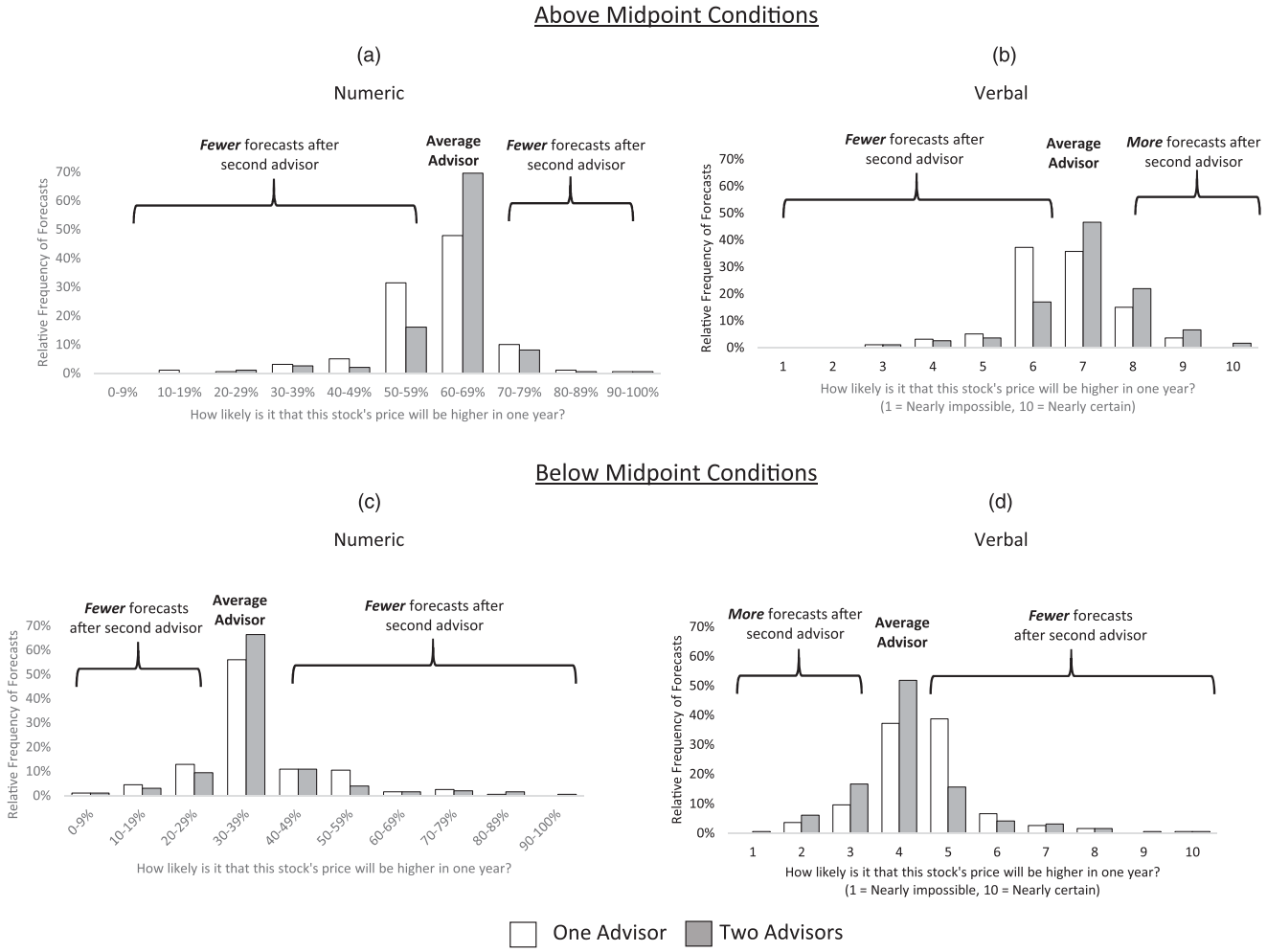
Figure 3 provides histograms that more precisely demonstrate the nature of our effect. Figure 3, (a) and (b) compares the numeric and verbal conditions when advisors provide forecasts above the midpoint. Figure 3, (c) and (d) compares the numeric and verbal conditions when advisors provide forecasts below the midpoint. The white bars show participants' forecasts after they see the first advisor's forecast, and the grey bars show participants' forecasts after they see the second advisor's forecast. In both cases, the histograms in the numeric condition (Figure 3, (a) and (c)) converge toward the "average" forecast as participants see the second advisor's forecast. That is, there are more estimates at "60%–69%" (above midpoint condition) or "30%–39%" (below midpoint condition) and fewer estimates at nearly every other point. Conversely, the histograms in the verbal condition (Figure 3, (b) and (d)) show a different pattern. Rather than only showing a spike at the average forecast (seven in the above midpoint condition and three in the below

Figure 2. Results from Study 2



Notes. Participants' forecasts become more extreme when they see a second verbal forecast but not when they see a second numeric forecast. Extreme forecasts are considered to be those closer to certainty than the forecasts of both advisors (i.e., higher than 7 of 10 in the *above midpoint* condition and lower than 4 of 10 in the *below midpoint* condition). Error bars represent 95% confidence intervals.

Figure 3. Histograms for Study 2



Notes. The distribution of participants' forecasts after seeing the first advisor's forecast is represented by white bars, and the distribution after seeing the second advisor's forecast is represented by grey bars. In the numeric conditions (panels (a) and (c)), participants' forecasts move toward the average advisor's forecast when the advisors provide forecasts both above (panel (a)) and below 50% (panel (c)). However, in the verbal conditions (panels (b) and (d)), participants' forecasts shift rightward when the advisors provide forecasts above 50% ("rather likely" panel (b)) and leftward when the advisors provide forecasts below 50% ("rather unlikely," panel (d)).

midpoint condition), there is a general shift in the distribution. Specifically, there are increases in the number of participants making a forecast at every point in the distribution that is more extreme than the average and decreases at nearly every point that is less extreme than the average.

Finally, Table 1 shows the results for the exploratory mechanism questions that we included in this study. In both the above and below midpoint conditions, participants perceived more advisor agreement in the *verbal* condition than in the *numeric* condition (extent to which advisors agreed and number of advisors agreeing; $p \leq 0.003$). However, this did not mediate our effect (i.e., for each of the mediation analyses, the confidence interval (CI) included zero) (Table 1).

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 use different combination strategies for verbal versus numeric probabilities, resulting in them making extreme forecasts more often when combining multiple verbal probabilities than when combining multiple numeric probabilities.

Study 3: Real Experts

In the first two studies, 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 (a) natural variation across advisors' forecasts, meaning that the advisors did not provide identical forecasts, and (b) more granular predictions (on a 0%–100% scale). In addition, in this study, we manipulated the number of advisors (one versus two) 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 versus verbal) × 2 (number of advisors: one versus two) design. All participants were shown information about 10 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 the 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 at least two of three expert websites). Additionally, they saw either one or two forecasts of how likely it was that the favorite would win the game, randomly selected from the websites 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, the advisors' 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 0–100 slider scale with numeric or verbal labels, depending on the condition. Participants saw advice from either 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 the forecasts (1 = completely distrust; 7 = completely trust). In addition, we asked participants what their favorite team was.

Results

Participants made forecasts for 10 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 those games in our analyses 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 included 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 footnotes).

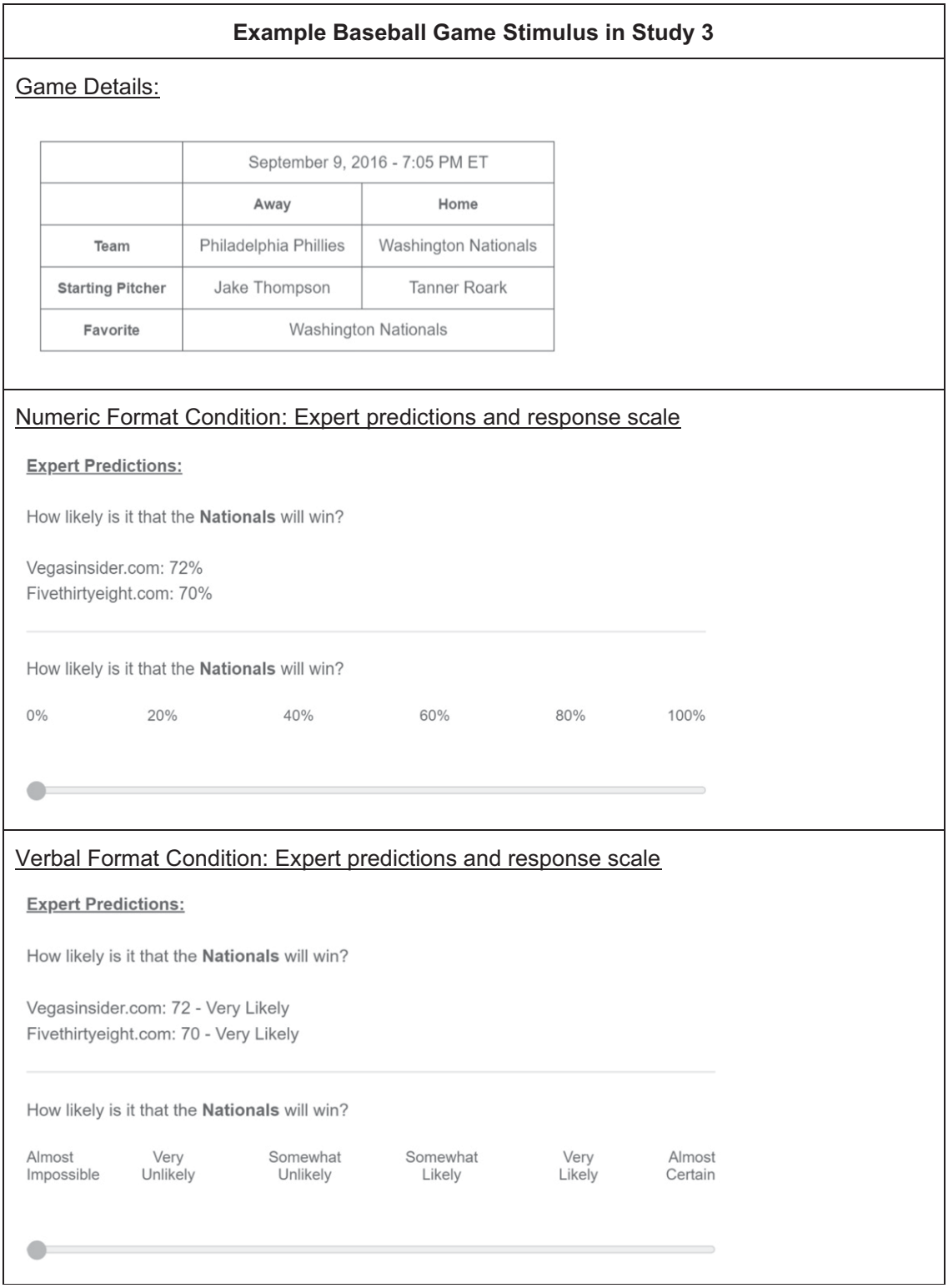
As in Studies 1 and 2, we classified participant forecasts as "extreme" if they were closer to certainty than each advisor's forecast for each game (i.e., closer to 100 than each advisor).[4] We used probit regression to regress whether participants made an extreme forecast (1 = yes; 0 = no) on (1) the 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 the average advisor's forecast as control variables in the regression. Our results do not change if these controls are not included (see Supplement 3 in the online appendix).

When participants saw only one advisor's forecast, there were no differences between the proportion of extreme forecasts in the *numeric* (50.0%) and *verbal* (55.5%) conditions, $Z = 1.39$, $p = 0.17$. However, participants who saw two advisors' forecasts were more likely to make more extreme forecasts in the *verbal* condition (46.6%) than in the *numeric* condition (29.8%), $Z = 5.11$, $p < 0.001$. The interaction between advice format and number of advisors was significant, $Z = 2.93$, $p = 0.003$. The interaction is also significant when we exclude the preregistered control variables, $Z = 2.56$, $p = 0.011$.[5]

Discussion

Study 3 shows that 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 versus verbal probability forecasts. Note that, unlike in Study 2, the number of extreme forecasts decreased in both conditions when participants saw two advisors' forecasts. We believe that this is because of the granularity of the scale in this study. Because there were

Figure 4. Sample Stimuli and Response Scales in Study 3



Notes. Stimuli shown correspond to the *Two-Advisors* condition. Participants in the *One-Advisor* condition saw nearly identical stimuli, except with only one expert prediction listed. Expert predictions were randomly selected from a list of three different websites. In this game, the Nationals won on a two-out walk-off home run.

101 possible predictions participants could have made, perhaps those that wanted to make an "average" forecast could have been satisfied to be in the general vicinity of the advisors' average. Indeed, only 16.1% of forecasts were exactly "average" when participants only saw one advisor's forecast in this study, compared with 44.2% in Study 2, where there were only 10 possible forecasts. It is also possible that participants had stronger prior beliefs in this study, leading to more initially extreme estimates.

Study 4: Incentivized 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 act as if they are "counting" 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, becoming more certain of an event's outcome. Thus, in Study 4, we test whether our findings extend to participants' decision making informed by their beliefs about the event's likelihood. This decision also included a meaningful incentive for accuracy. Specifically, we gave participants the opportunity to bet money on the outcome of a football game.

Method

We recruited 1,024 participants on MTurk, of which 811 (40.3% female, M_{age} = 36.8 years) passed both an attention check and a comprehension check and continued to the rest of the survey, each paid $0.50. All participants were truthfully told that 80 of the approximately 800 participants recruited would receive a $5 bonus and that, if they were selected to receive the bonus, the amount of this bonus could increase or decrease based off their decisions in the study.

Participants then learned that they would have the opportunity to bet their bonus on the outcome of an upcoming National Football League (NFL) game. Specifically, participants learned that they could bet any amount of the $5 bonus on the favored team (defined as being the consensus favorite of the two websites). If the favored team won, the amount they bet would be added to the original bonus of $5, and if the favored team lost, it would be subtracted from the bonus. For example, if the participant chose to bet $4 on the Patriots and was selected to receive the bonus, she would receive $9 if the Patriots won ($5 + $4) and $1 if the Patriots lost ($5 − $4). Participants were asked to answer a comprehension check about these instructions and were only able to continue with the study after they answered the comprehension check correctly.

After completing a comprehension check, participants read that they would see expert predictions about an NFL game provided on a 10-point scale (similar to those used in Studies 1 and 2). They were also shown a chart describing how to translate the verbal probabilities into their equivalent numeric probabilities and vice versa. That is, they saw a chart that said "0%–9% = nearly impossible," "10%–19% = extremely unlikely," and so on. Participants then saw information about an NFL game (randomly selected from a list of five games) along with the expert predictions, taken from either Fivethirtyeight.com or Vegasinsider.com. The expert predictions were converted to a common scale, corresponding to the chart that they previously saw.

Participants were randomly assigned to one of two between-subjects conditions: *numeric* versus *verbal* advice format. If the expert website predicted that the favorite had between a 50% and 59% chance of winning, participants were shown a prediction of "somewhat likely" or "50%–59%," depending on whether they were in the *verbal* or *numeric* condition. We originally also included two games in this study (Bears versus Chiefs, Vikings versus Packers) for which the expert websites predicted that the favorite had between a 60% and 69% chance of winning. Unfortunately, after running the study, we discovered that there was a coding error in our survey, so that participants saw a prediction of "60%–69%" in the numeric condition, but in the verbal condition, they saw "somewhat likely" (which corresponds to the "50%–59%" numeric condition) instead of "rather likely" (which corresponds to the "60%–69%" numeric condition). Our main analysis below includes only the three games that did not contain this error. However, including all five games in the analysis does not change the results.

After seeing the expert predictions, participants could then bet any portion of their bonus on the team identified by the website as the favorite (i.e., the team with a greater than 50% chance of winning). We asked participants, "If you are chosen to receive the bonus, how much of that bonus would you like to bet on the [favored team]?" Participants responded using a slider scale that ranged from $0 to $5 in increments of $0.10. After making their first bet, participants then saw a second expert prediction either in *verbal* or *numeric* format depending on the condition they were assigned to (from whichever website they did not see

a prediction from previously) and could adjust their bet. The second website's prediction was the same as the first website's prediction.

At the end of the study, we asked participants to answer five NFL knowledge questions.

Results

Our preregistered dependent variable is the proportion of participants who *increased* their bets after seeing the second expert opinion. We predicted that participants in the verbal condition would be more likely to *increase* their bets (which would be consistent with a counting strategy) than participants in the numeric condition.

We used OLS regression to regress whether participants increased their bets (1 = yes; 0 = no) on the advice format (1 = verbal; 0 = numeric), including fixed effects for game. As preregistered, we also included the total number of knowledge questions answered correctly as a control variable. More participants increased their bets in the *verbal* condition (25.4%) than in the *numeric* condition (11.6%), $t(478) = 4.21$, $p < 0.001$, $\eta^2 = 0.036$. These results remain the same when we do not add game fixed effects and the football knowledge control to the regression, $t(484) = 3.98$, $p < 0.001$, $\eta^2 = 0.032$.[6] The results also hold when we include all games in the analysis: that is, more participants increased their bets in the *verbal* condition (20.5%) than in the *numeric* condition (13.3%), including controls, $t(800) = 2.98$, $p = 0.003$, $\eta^2 = 0.011$, and excluding controls, $t(809) = 2.74$, $p = 0.006$, $\eta^2 = 0.009$.[7]

Discussion

Study 4 shows that the differences in people's strategies for combining verbal versus numeric probabilities can affect subsequent decision making that is informed by participants' beliefs about an event's likelihood. In the remainder of the studies, we test several potential mechanisms for this finding.

Studies 5A and 5B: Are People Being Bayesian?

As discussed in the introduction, although it is typically better to average others' forecasts if forecasters are using correlated information (e.g., Ashton and Ashton 1985), making more extreme estimates is the normatively correct strategy if the forecasters are using uncorrelated or poorly correlated information (Wallsten and Diederich 2001, Baron et al. 2014). If participants perceive less information correlation between two advisors who provide similar verbal forecasts than they do between advisors who provide similar numeric forecasts, it would be normatively appropriate for participants to use a "counting" strategy more often when combining verbal forecasts (Wallsten et al. 1997a, p. 51).

In Study S5 (see Supplement 2 in the online appendix), we first tested whether participants' inferences about information correlation between advisors influences their combination strategies and found suggestive evidence that this is indeed the case. Specifically, in Study S5, compared with those in the numeric condition, participants in the verbal condition felt that a second forecast is more likely to provide new information, that a second forecast is more likely to be useful, and that the two forecasts are more likely to provide different information ($p < 0.001$) (see Table 1). This suggests that participants may perceive similar verbal forecasts to contain less correlated information than similar numeric forecasts do. If participants make such inference, then using a "counting" strategy would be the more normative strategy in the verbal condition. Indeed, two of the three significant mediators in Study S5 ("To what extent did the second site provide you with new information?" and "Do you think the two sites used the same or different information to make their predictions?") partially mediated the effect of forecast format on participants' decisions (see Table 1).

However, despite the significant mediation shown in Table 1, the evidence supporting this account in Study S5 is not particularly strong, as the two mediators only account for between 11% and 16% of our effect. In addition and importantly, the task that participants completed in Study S5 (i.e., a consumption decision based on a likelihood forecast; somewhat similar to the task in Study 4) was qualitatively different from those in Studies 1–3 (i.e., making predictions about stock prices or sporting events). We therefore conducted Studies 5A and 5B, using a scenario similar to that in Study 2, to test this potential mediator more completely. In Study 5A, we measured the potential mediator, and in Study 5B, we directly manipulated whether 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 1,017 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 seventh 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 (format: numeric

versus verbal; between subjects) × 2 (number of advisors: one versus two; within subjects) mixed design.

We included two exploratory 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. This question was similar to one of the mechanism questions in Study S5 and read, "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). In addition, we added another question designed to test participants' perceptions of the advisors' agreement, similar to the exploratory mechanism question that we included in Study 1. We asked participants to indicate the extent to which both advisors agreed on the forecast ("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).

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, $Z = 2.60$, $p = 0.009$. The proportion of extreme forecasts increased from 24.1% to 33.5% when participants saw the second advisor's forecast in the *verbal* condition but did not substantially change when participants saw the second advisor's forecast in the *numeric* condition (14.2% versus 14.7%). Thus, we replicated our main finding that participants tend to average numeric forecasts but tend to act as if they are "counting" verbal forecasts.[8]

Analysis of Mediator Questions. We present the results for the exploratory mechanism questions in Table 1. Unlike in Study S5, which suggested perceived correlation as a promising mechanism, we did not find a significant difference in the perceived correlation of advisors' information across conditions ($p = 0.34$). There was also no significant difference in perceived advisor agreement ($p = 0.71$). Given that we do not find a difference in the perceived correlation of the information that the advisors based their forecast on across conditions, this is initial evidence against the idea that participants' combination strategies are driven by their beliefs about the correlation of the advisors' information. However, 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 versus two; within subjects) × 2 (advice format: numeric versus verbal; between subjects) × 2 (information source: different versus 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[9]). 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 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, $p < 0.001$.

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, $p = 0.003$.[10] That is, we again find that participants in the verbal format condition become more likely to make an extreme forecast after seeing the second advisor's forecast than those in the numeric format condition do. However, neither the interaction between information source and number of advisors, Z = 0.27, $p = 0.79$, nor the three-way interaction between format, number of advisors, and information source, Z = –0.95, $p = 0.34$, are significant. Figure 5 illustrates this: the results are nearly identical in both the *similar* (Figure 5(a)) and *different* (Figure 5(b)) information conditions, indicating that manipulating the information source does not impact how participants are combining verbal and numeric forecasts.

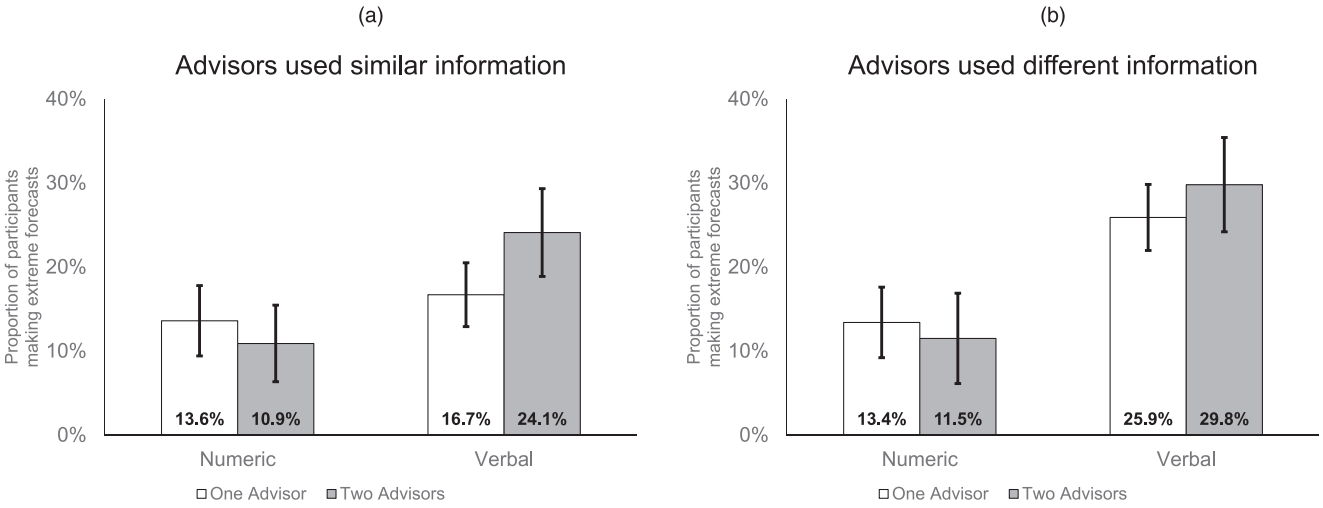
Discussion

Taken together, the results of Studies 5A and 5B suggest that, unlike in Study S5, 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. Because of the conflicting results between Study S5 and Studies 5A and 5B, we do not want to definitively rule out the idea that participants are using a Bayesian strategy in the verbal condition, but the combined results of all of the studies suggest that, if they are, it does not account for a meaningful proportion of our effect.

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 caused by participants' assumptions that forecasters are more likely to be using uncorrelated information when providing verbal forecasts. In Study 6, we test another explanation for our effect. It may be easier to infer the *direction* of the forecast from

Figure 5. Study 5B Results



Notes. The overall pattern of results is consistent with prior studies and does not change according to whether we told participants that the advisors were using similar information (panel (a)) or different information (panel (b)), indicating that this does not influence participants' combination strategies. Error bars represent 95% confidence intervals.

verbal probabilities than from numeric probabilities. That is, when participants see verbal forecasts, they may have a better intuition about whether a forecast is a "good sign" or a "bad sign" that an event will occur (Teigen and 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 (Windschitl and Wells 1998, Teigen 2001). In contrast, a candidate that is "likely" to win will be considered the favorite, regardless of the number of other candidates.

If seeing verbal forecasts shifts participants' focus to interpreting forecasts as simply good or bad signs, a counting strategy may feel more intuitive. We designed Study 6 to test this account by adding information about the direction of the forecasts to numeric probabilities and assessing whether this increases people's tendency to act as if they are counting numeric forecasts.[11]

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, 5A, and 5B. 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 versus 5.22), $t(799) = -4.03$, $p < 0.001$, and that adding direction to numeric predictions increases participants' belief that the prediction is a good sign (5.22 without direction versus 5.66 with direction), $t(799) = -6.76$, $p < 0.001$.

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 act as if they are counting *verbal* forecasts than they are *numeric* forecasts. The interaction between advice format (*verbal* versus *numeric*) and the number of advisors (*one* versus *two*) was again significant, $Z = 5.60$, $p < 0.001$.[12] In the *verbal* condition, the proportion of extreme forecasts increased from 19.9% to 31.0%, $Z = 7.04$, $p < 0.001$, after participants saw the second advisor's forecast. In the *numeric* condition, the proportion of extreme forecasts directionally decreased from 12.9% to 11.6%, $Z = -1.27$, $p = 0.205$.

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 act as if they are counting, then we would expect that participants in the *numeric with direction* condition also become more likely to "count." 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* versus *numeric with direction* advice format and the number of advisors (*one* versus *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%, $Z = 0.93$, $p = 0.35$, as participants saw a second advisor's forecast, and the interaction between format and number of advisors is not significant, $Z = 1.53$, $p = 0.126$, providing no evidence that making numeric forecasts more evaluable by telling participants that the forecast is a positive sign increases people's tendency to make more extreme forecasts in this condition.

Discussion

The results from Study 6 suggest that it is unlikely that the perceived evaluability of the advisors' forecasts is driving our effect, as making numeric forecasts more evaluable did not increase participants' tendency to

make more extreme forecasts in this condition. In Study 7, we measure and rule out several additional exploratory mechanisms.

Study 7: Additional Mechanisms Tested

Although we tested what we thought were the most plausible mechanisms in Studies 5 and 6, we tested several additional candidate mechanisms in Study 7. In this study, we also included an additional condition that shows the advisors' forecasts in both numeric and verbal format (e.g., "60%–69%–rather likely"). This condition allows us to test whether one of the formats "dominates." If, for example, averaging is the "default" and people switch to counting when they see verbal forecasts, we would expect the *numeric with verbal* condition to look more like the verbal condition and vice versa.

Method

We recruited 604 participants on MTurk (51.8% female, M_{age} = 36.5 years). The design of Study 7 was nearly identical to that of Study 2, except that we added a new condition and new exploratory mechanism questions.

We randomly assigned participants to one of three between-subjects conditions: *numeric* versus *verbal* versus *numeric with verbal* advice format. All participants saw a stock and its most recent closing price and were asked to predict whether the stock's price would be higher in one year. Participants saw predictions from two advisors, one at a time, and made their own forecasts after seeing each advisor's forecast. Both advisors said either that there was a "60%–69%" chance that the stock's price would be higher (*numeric* condition) or that it was "rather likely" (*verbal* condition). In the third, new *numeric with verbal* condition, we combined the two formats (i.e., the advisors' forecasts were "60%–69%–rather likely").

After participants made their forecasts, we asked them a series of exploratory mechanism questions. Questions 1 and 2 are similar to what we tested in previous studies, and questions 3–7 are additional exploratory mechanisms. The wording of the questions was as follows:

1. *Advisors' agreement.* If these analysts gave you predictions about another different stock, how likely do you think they are to agree with each other? (1 = they definitely would not; 7 = they definitely would)
2. *Same versus different information.* Do you think Analyst A and Analyst B were using the same information or different information to make their predictions for this question? (1 = definitely the same; 7 = definitely different)
3. *Knowledge versus luck.* Do you think that accurately predicting if a stock will increase over the next year requires more luck or more knowledge? (1 = definitely more luck; 7 = definitely more knowledge)
4. *Inside versus outside view.* When making their predictions, do you think the analysts thought more about how stocks will perform on average or how this specific stock will perform? (1 = definitely about how stocks will perform on average; 7 = definitely about this specific stock)
5. *Opinions versus facts.* Do you think the analysts based their predictions more on facts or more on their own opinions? (1 = entirely on facts; 7 = entirely on opinions)
6. *Advisors' thoughtfulness.* How much thought do you think each analyst put into deciding whether the stock would increase? (1 = not a lot; 7 = a lot)
7. *Advisors' strength of opinion.* Do you think each analyst has a strong opinion about whether this stock will increase? (1 = definitely not; 7 = definitely)

We discuss the relevance of each of these potential mechanisms below, along with presenting the results from the analyses.

Results

Main Analysis. We used probit regressions to regress whether participants made an extreme forecast (1 = yes; 0 = no) on (1) the advice format condition (using the *numeric* condition as the reference and indicators for both the *verbal* and *numeric with verbal* conditions), (2) the number of advisors, and (3) their interactions. These analyses included fixed effect for stock and clustered standard errors by participant.

Table 3 shows the results for Study 7. In line with the results from our prior studies, when comparing

Table 3. Proportion of Extreme Participant Forecasts by Condition in Study 7

Advice format condition	One advisor	Two advisors	One vs. two advisors
Numeric	13.9%	11.9%	$Z = -0.82$, $p = 0.42$
Verbal	17.8%	35.1%	$Z = 5.35$, $p < 0.001$
Numeric with verbal	18.9%	23.4%	$Z = 1.75$, $p = 0.08$

Notes. Interaction between advice format (numeric vs. verbal) and number of advisors: $Z = 4.54$, $p < 0.001$. Interaction between advice format (numeric vs. numeric with verbal) and number of advisors: $Z = 1.91$, $p = 0.056$.

the numeric and verbal conditions only, we again found a significant interaction between advice format and number of advisors, Z = 4.54, $p < 0.001$.[13] When the advisors provided *verbal* probability forecasts, the proportion of participants making extreme forecasts increased from 17.8% after the first advisor to 35.1% after the second advisor, Z = 5.35, $p < 0.001$. In contrast, when the advisors provided *numeric* probability forecasts, the proportion of participants making extreme forecasts directionally decreased from 13.9% to 11.9%, Z = –0.82, $p = 0.42$.

In addition, when comparing the *numeric* and *numeric with verbal* conditions, the interaction between advice format and number of advisors was marginally significant, Z = 1.91, $p = 0.056$. Specifically, although the proportion of participants making extreme forecasts in the *numeric* condition did not change after seeing the second advisor's forecast, the proportion of extreme forecasts in the *numeric with verbal* condition increased from 18.9% to 23.4% when participants saw the second advisor's forecast, Z = 1.75, $p = 0.08$. That is, it seems that some participants in the numeric with verbal condition used a combination strategy that was similar to those in the verbal condition. However, rerunning the analysis with the *verbal* condition as a reference, the interaction between the *numeric with verbal* condition and the number of advisors was negative and significant, Z = –2.83, $p = 0.005$, indicating that although some participants in the *numeric with verbal* condition used what looked like a counting strategy, this strategy was still more prevalent in the *verbal* condition.

Analysis of Potential Mechanisms. Table 1 presents the results for the exploratory mechanisms. For simplicity, we only compare results from the *numeric* condition and the *verbal* condition because those are the conditions of interest for our primary effect.

Advisors' Agreement. Consistent with the results from Study 5A, we did not find a significant difference between the *verbal* (M = 4.74) and *numeric* (M = 4.77) conditions in the perceived extent to which the advisors agree, t(389) = –0.01, $p = 0.99$.

Same Vs. Different Information. Participants were marginally more likely to think that the advisors were using different information in the *verbal* condition (M = 3.10) than in the *numeric* condition (M = 2.88), t(390) = 1.77, $p = 0.078$. However, this difference did not mediate our effect, $b = –0.01$, 95% CI (–0.09, 0.05).

Knowledge Vs. Luck. Tannenbaum et al. (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/luck). 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, participants showed no difference in how they rated whether the analysts were relying on knowledge or luck between the *verbal* (M = 4.76) and *numeric* (M = 4.95) conditions, t(390) = –1.35, $p = 0.18$, making it unlikely that this was a conscious factor in participants' combination strategies.

Inside Vs. Outside View. To the extent that the use of percentages causes participants to think about base rates, it may be that seeing numeric probabilities prompts participants to take the outside view (i.e., they think about an average stock or the stock market as a whole), whereas 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 and Lovallo 1993), this might be driving participants to make more extreme forecasts in the *verbal* conditions. However, 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(389) = –0.08, $p = 0.94$, making it unlikely that this is driving our effect.

Opinions Vs. Facts. Numeric forecasts might seem to represent a more objective truth about the stock, whereas 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. However, there was no difference in the extent to which participants thought the advisors' forecasts where based on opinions versus facts across the *verbal* (M = 3.50) and *numeric* (M = 3.37) conditions, t(389) = 0.96, $p = 0.34$.

Advisors' Thoughtfulness. Perhaps participants are more likely to select the exact same forecast as the advisors gave in the numeric condition because they think that the advisors put more thought into their forecasts. Alternatively, if they perceive verbal advice as more thoughtful, this may allow them to become more confident in the direction of the forecasts. Although participants seemed to believe that advisors in the *numeric* condition (M = 5.42) put slightly more thought into their forecasts than those in the *verbal*

condition (M = 5.22), t(389) = –1.73, $p = 0.08$, this did not mediate our effect, $b = –0.03$, 95% CI (–0.13, 0.07).

Advisors' Strength of Opinion. Because of the inherent direction of verbal probabilities (Teigen and Brun 1995, 1999), participants may conclude that advisors who provide verbal probabilities have more strongly held beliefs. This, in turn, may increase participants' confidence in their own forecasts. However, given that numeric probabilities are more precise than verbal probabilities (Lichtenstein and Newman 1967), it is also possible that numeric forecasts signal stronger beliefs. Indeed, participants thought advisors had slightly stronger opinions in the *numeric* condition (M = 4.65) than in the *verbal* condition (M = 4.36), t(389) = –2.05, $p = 0.041$. However, this did not mediate our effect, $b = –0.04$, 95% CI (–0.15, 0.05). In an additional study (Study S2), we found no difference in participants' self-reported confidence between the *verbal* (M = 4.59) and *numeric* (M = 4.45) conditions, t(392) = 0.90, $p = 0.37$ (see Supplement 2 in the online appendix).

Discussion

Taken together, the results from these additional exploratory analyses do not provide strong evidence for one specific mechanism. In Study S2 (see Supplement 2 in the online appendix), we also tested the extent to which participants reported using intuition versus reason when making their own forecast. Whereas verbal probabilities may lead people to make quicker and more intuitive Bayesian judgments (Gigerenzer and Hoffrage 1995; although cf. Biswas et al. 2011), numeric probabilities tend to trigger rule-based thinking. This could lead participants to move in the direction of making deliberate mathematical calculations, such as averaging, in the numeric condition. 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(392) = –0.98, $p = 0.33$.

General Discussion

In eight 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 seem as if they are averaging numeric probability forecasts and are more likely to act as if they are "counting" verbal probability forecasts. As a result, participants' own forecasts move closer to the average of the advisors' forecasts as they see additional numeric forecasts, but they become more extreme (i.e., they move closer to certainty) as they see additional verbal forecasts from advisors. We show that this effect occurs for both probabilities above and below 50%, for hypothetical scenarios and real events, and when presenting the advisors' forecasts simultaneously or sequentially. Importantly, the differences in these strategies affect internal representations of uncertainty enough to influence consequential decisions based on this uncertainty, as shown in Study 4.

Our findings are broadly consistent with those of Budescu, Wallsten, and colleagues, who found that participants make more extreme and "less fuzzy" forecasts when combining identical verbal probability estimates for chance events (Budescu et al. 1990) and that participants are generally more confident about knowledge questions when they see multiple others express their confidence verbally, compared with numerically (Wallsten et al. 1997a). By directly contrasting participants' behavior both for numeric and verbal probabilities and for one or two advisors, we can more definitively say that the differences found in these prior studies were driven by differences in how participants combine probabilities of different formats. In addition, we show that this effect holds for a variety of contexts, is not exclusive to combining identical forecasts, and influences consequential downstream decision making.

We tested several potential mechanisms for our effect. The results from Studies 5A and 5B indicate that it is unlikely that the differences in combination strategies are caused by participants believing that a verbal forecast from an additional advisor provides more new information than a numeric forecast from an additional advisor. In addition, the results from Study 6 indicate that it is unlikely that the inherent direction of verbal probabilities is driving the effect. We tested several additional candidate mechanisms in Studies 7, S2, and S5, but we did not find strong evidence for any of them. Although we did not find strong evidence for any one of these mechanisms specifically, it is possible that: (a) the difference in how people combine numeric versus verbal probabilities is multiply determined, with each possible mechanism only accounting for a small percentage of the overall effect; (b) our mediator questions were not fully able to measure the processes we tried to study; or (c) participants are using a nonconscious process that we cannot measure. Indeed, it may be a combination of all three of these, where the impacts of multiple small, potentially immeasurable differences in how participants view verbal and numeric probabilities combine or otherwise interact with each other to affect participants' combination strategies. Of course, there is the final possibility that there is a potential mechanism we have not yet considered.

Further Research and Practical Implications

In the studies we present in the manuscript, we tested how people combine forecasts from two advisors. But why only two? It is reasonable to ask whether our

effect persists when participants see even more advisors. We tested this in an additional study (Study S1 in Supplement 2 in the online appendix) 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 "three times out of five" rather than 60%. Prior research suggests that people combine frequency forecasts differently than they combine percentages. For example, when dealing with frequency data, people are more likely to take a Bayesian approach (Gigerenzer and Hoffrage 1995, Hoffrage et al. 2000, Biswas et al. 2011), 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 and 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 with 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 "four out of five" or as "very good." As a result, consumers may combine two "four out of five" 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, we do not see any intuitive reason why a person, for example, would see two "good" ratings and conclude that a product is "very good." 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 and Weber 1999). Of course, there is a wide variety of verbal probabilities that one can use (Lichtenstein and Newman 1967, Beyth-Marom 1982) 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., Stoner 1968, Lord et al. 1979, Isenberg 1986).

Our results 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. For example, current guidelines issued by the Director of National Intelligence (Clapper 2015) and the Intergovernmental Panel on Climate Change (Mastrandrea et al. 2011) provide decision makers with "conversion tables" to help them translate verbal probabilities into numeric ones. However, our findings suggest that these conversion tables are of limited use if readers see multiple forecasts for the same event. Similarly, our research is also relevant to organizations that provide aggregated forecasts to others. Websites such as The Upshot,[14] which typically presents multiple election predictions in a table with numeric (e.g., "65% Democrat") and verbal (e.g., "lean Democrat") probabilities side by side, 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 act as if they are "counting" 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.

This paper contains an online appendix. The contents of the online appendix are listed in Table 4.

Table 4. Index of Supplementary Materials (Available from https://researchbox.org/62)

Section	Pages
Supplement 1. Study 5B stimuli	3
Supplement 2. Supplementary studies not included in main manuscript	4–12
Supplement 3. Full regression results and additional preregistered analyses for studies in manuscript	13–35

Acknowledgments

The authors thank Uri Simonsohn, Joe Simmons, Deborah Small, Barbara Mellers, and the review team for their thoughtful and constructive comments. The authors also thank Soaham Bharti, Beidi Hu, Jordyn Schor, and Sara Sermarini for valuable research assistance.

Appendix. Links to Preregistrations

Study 2: https://aspredicted.org/nb3mq.pdf
Study 3: https://aspredicted.org/er3uv.pdf
Study 4: https://aspredicted.org/ke7hj.pdf
Study 5A: https://aspredicted.org/bb75p.pdf
Study 5B: https://aspredicted.org/s7wb4.pdf
Study 6: https://aspredicted.org/8py85.pdf
Study 7: https://aspredicted.org/tj5g5.pdf
Study S1: https://aspredicted.org/83fa5.pdf
Study S2: https://aspredicted.org/38te7.pdf
Study S3: https://aspredicted.org/xn7zv.pdf
Study S4: https://aspredicted.org/wq5z9.pdf
Study S5: https://aspredicted.org/u8gx6.pdf

Endnotes

[1] In addition, a counting strategy may be optimal when combining forecasts from many advisors. This is because individual forecasts are often too conservative, particularly for hard to predict events (Wallsten et al. 1997b, Ariely et al. 2000, Baron et al. 2014). However, this is not relevant to our studies because participants in our studies typically only see forecasts from two advisors.

[2] Study 4 uses a behavioral measure on a different scale, making this comparison impossible.

[3] To check whether participants' forecasts simply became noisier in the verbal conditions, we also tested whether participants became more likely to make forecasts below the expert forecasts in the above midpoint condition (or above the expert forecasts in the below midpoint condition). We found no significant interaction between format and number of advisors for this dependent measure in the above midpoint conditions, $Z = -0.52$, $p = 0.60$. There was a significant interaction in the below midpoint condition, $Z = -3.86$, $p < 0.001$, and when collapsing the above and below midpoint conditions, $Z = -3.00$, $p = 0.003$. However, this is because participants' forecasts in the below midpoint/verbal condition become less noisy (i.e., the interaction is significant in the opposite direction of what would be expected by an "increased noise" explanation). In the below midpoint/verbal condition, the proportion of participants who make a forecast above the expert forecasts decreased from 49.7% to 25.1%. In the below midpoint/numeric condition, these numbers were 25.7% and 20.3%, respectively.

[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 are significant, with controls ($Z = 2.72$, $p = 0.007$) or without controls ($Z = 2.44$, $p = 0.015$).

[5] Including all games (i.e., including games where the experts did *not* agree on the outcome), the interaction is still significant, with controls ($Z = 3.23$, $p = 0.001$) or without controls ($Z = 2.80$, $p = 0.005$). We also preregistered that we would run our analyses excluding games involving participants' favorite teams. The interaction is still significant with controls ($Z = 2.86$, $p = 0.004$) or without controls ($Z = 2.51$, $p = 0.012$). When comparing the proportion of participants giving a forecast that is below the lowest expert forecast, the interaction is not significant with controls ($Z = -0.29$, $p = 0.77$) or without controls ($Z = -0.31$, $p = 0.76$), again indicating that our effect is not caused by participants making noisier forecasts in the verbal condition.

[6] There is no difference in the proportion of participants who decrease their bets (verbal: 9.02%, numeric: 10.3%), with controls ($t(478) = -0.45$, $p = 0.65$, $\eta^2 = 0.0004$) or without controls ($t(484) = -0.49$, $p = 0.63$, $\eta^2 = 0.0005$).

[7] There is also no difference in the proportion of participants that decrease their bets (verbal: 11.4%, numeric: 9.9%), with controls ($t(800) = 0.81$, $p = 0.42$, $\eta^2 = 0.0008$) or without controls ($t(809) = 0.70$, $p = 0.49$, $\eta^2 = 0.0006$).

[8] The interaction between advice format and number of advisors is not significant when comparing the proportion of participants making a forecast below the advisors' forecasts, $Z = 0.90$, $p = 0.37$, demonstrating again that participants' forecasts do not simply become "noisier."

[9] To account for any effects caused by the information source itself (i.e., a statistical model versus 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 ($M = 2.70$), $t(505) = 0.73$, $p = 0.47$.

[10] When comparing the proportion of participants who give a forecast below the advisors' forecasts, this interaction is not significant, $Z = 0.72$, $p = 0.47$, demonstrating again that participants' forecasts do not simply become "noisier."

[11] Prior to running this study, we ran three nearly identical studies (Studies S2–S4 in Supplement 3 in the online appendix) 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.

[12] The interaction when predicting the likelihood of making a prediction below the advisors' forecasts is not significant, $Z = -0.58$, $p = 0.56$.

[13] This interaction is not significant when comparing the proportion of participants making a forecast below the advisors' forecasts, $Z = 0.47$, $p = 0.64$, demonstrating again that participants' forecasts do not simply become "noisier."

[14] See https://www.nytimes.com/section/upshot.

References

Ariely D, Tung Au W, Bender RH, Budescu DV, Dietz CB, Gu H, Wallsten TS, Zauberman G (2000) The effects of averaging subjective probability estimates between and within judges. *J. Experiment. Psych. Appl.* 6(2):130–147.

Ashton AH, Ashton RH (1985) Aggregating subjective forecasts: Some empirical results. *Management Sci.* 31(12):1499–1508.

Baron J, Mellers BA, Tetlock PE, Stone E, Ungar LH (2014) Two reasons to make aggregated probability forecasts more extreme. *Decision Anal.* 11(2):133–145.

Berg S (1993) Condorcet's jury theorem, dependency among jurors. *Soc. Choice Welfare* 10(1):87–95.

Beyth-Marom R (1982) How probable is probable? A numerical translation of verbal probability expressions. *J. Forecasting* 1(3): 257–269.

Biswas D, Zhao G, Lehmann DR (2011) The impact of sequential data on consumer confidence in relative judgments. *J. Consumer Res.* 37(5):874–887.

Budescu DV, Yu H-T (2006) To Bayes or not to Bayes? A comparison of two classes of models of information aggregation. *Decision Anal.* 3(3):145–162.

Budescu DV, Yu H-T (2007) Aggregation of opinions based on correlated cues and advisors. *J. Behav. Decision Making* 20(2): 153–177.

Budescu DV, Weinberg S, Wallsten TS (1988) Decisions based on numerically and verbally expressed uncertainties. *J. Experiment. Psych. Human Perception Performance* 14(2):281–294.

Budescu DV, Zwick R, Wallsten TS, Erev I (1990) Integration of linguistic probabilities. *Internat. J. Man-Machine Stud.* 33(6): 657–676.

Clapper J (2015) Intelligence Community Directive 203: Analytic Standards. Accessed May 22, 2020, https://www.dni.gov/files/documents/ICD/ICD 203 Analytic Standards.pdf.

Condorcet M (1785) *Essay on the Application of Analysis to the Probability of Majority Decisions* (Imprimerie Royale, Paris).

Dunning D (2007) Prediction: The inside view. Kruglanski AW, Higgins ET, eds. *Social Psychology: Handbook of Basic Principles*, 2nd ed. (Guilford Press, New York), 69–90.

Erev I, Cohen BL (1990) Verbal vs. numerical probabilities: Efficiency, biases, and the preference paradox. *Organ. Behav. Human Decision Processes* 45(1):1–18.

Galton F (1907) Vox populi. *Nature* 75(1907):450–451.

Gigerenzer G, Hoffrage U (1995) How to improve Bayesian reasoning without instruction: Frequency formats. *Psych. Rev.* 102(4): 684–704.

Hoffrage U, Lindsey S, Hertwig R, Gigerenzer G (2000) Communicating statistical information. *Science* 290(5500):2261–2262.

Isenberg DJ (1986) Group polarization: A critical review and meta-analysis. *J. Personality Soc. Psych.* 50(6):1141–1151.

Kahneman D, Lovallo D (1993) Timid choices and bold forecasts: A cognitive perspective on risk taking. *Management Sci.* 39(1):17–31.

Larrick RP, Soll JB (2006) Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Sci.* 52(1): 111–127.

Lichtenstein S, Newman JR (1967) Empirical scaling of common verbal phrases associated with numerical probabilities. *Psychonomic Sci.* 9(10):563–564.

Lord CG, Ross L, Lepper MR (1979) Biased assimilation and attitude polarization: The effects of prior theories on subsequently considered evidence. *J. Personality Soc. Psych.* 37(11): 2098–2109.

Mastrandrea MD, Mach KJ, Plattner G-K, Edenhofer O, Stocker TF, Field CB, Ebi KL, Matschoss PR (2011) The IPCC AR5 guidance note on consistent treatment of uncertainties: A common approach across the working groups. *Climatic Change* 108(4):675.

Sah S, Loewenstein G (2015) Conflicted advice and second opinions: Benefits, but unintended consequences. *Organ. Behav. Human Decision Processes* 130:89–107.

Sarvary M (2002) Temporal differentiation and the market for second opinions. *J. Marketing Res.* 39(1):129–136.

Schwartz J, Luce MF, Ariely D (2011) Are consumers too trusting? The effects of relationships with expert advisers. *J. Marketing Res.* 48(SPL):S163–S174.

Stoner JAF (1968) Risky and cautious shifts in group decisions: The influence of widely held values. *J. Experiment. Soc. Psych.* 4(4): 442–459.

Stroop JR (1932) Is the judgment of the group better than that of the average member of the group? *J. Experiment. Psych. General* 15(5):550–562.

Tannenbaum D, Fox CR, Ülkümen G (2016) Judgment extremity and accuracy under epistemic vs. aleatory uncertainty. *Management Sci.* 63(2):497–518.

Teigen KH (2001) When equal chances = good chances: Verbal probabilities and the equiprobability effect. *Organ. Behav. Human Decision Processes* 85(1):77–108.

Teigen KH, Brun W (1995) Yes, but it is uncertain: Direction and communicative intention of verbal probabilistic terms. *Acta Psych.* 88(3):233–258.

Teigen KH, Brun W (1999) The directionality of verbal probability expressions: Effects on decisions, predictions, and probabilistic reasoning. *Organ. Behav. Human Decision Processes* 80(2): 155–190.

Wallsten TS, Diederich A (2001) Understanding pooled subjective probability estimates. *Math. Soc. Sci.* 41(1):1–18.

Wallsten TS, Budescu DV, Tsao CJ (1997a) Combining linguistic probabilities. *Psychologische Beiträge* 39(1–2):27–55.

Wallsten TS, Budescu DV, Erev I, Diederich A (1997b) Evaluating and combining subjective probability estimates. *J. Behav. Decision Making* 10(3):243–268.

West PM, Broniarczyk SM (1998) Integrating multiple opinions: The role of aspiration level on consumer response to critic consensus. *J. Consumer Res.* 25(1):38–51.

Windschitl PD, Weber EU (1999) The interpretation of "likely" depends on the context, but "70%" is 70%–right? The influence of associative processes on perceived certainty. *J. Experiment. Psych. Learn, Memory Cognition* 25(6):1514–1533.

Windschitl PD, Wells GL (1996) Measuring psychological uncertainty: Verbal vs. numeric methods. *J. Experiment. Psych. Appl.* 2(4):343–364.

Windschitl PD, Wells GL (1998) The alternative-outcomes effect. *J. Personality Soc. Psych.* 75(6):1411–1423.

Yaniv I (2004) The benefit of additional opinions. *Current Directions Psych. Sci.* 13(2):75–78.

Yaniv I, Milyavsky M (2007) Using advice from multiple sources to revise and improve judgments. *Organ. Behav. Human Decision Processes* 103(1):104–120.

Zimmer AC (1983) Verbal vs. numerical processing of subjective probabilities. *Adv. Psych.* 16(1983):159–182.