AMERICAN PSYCHOLOGICAL ASSOCIATION

Journal of Experimental Psychology: General

ISSN: 0096-3445

2024, Vol. 153, No. 8, 2088–2099
https://doi.org/10.1037/xge0001611

Combining Forecasts From Advisors: The Impact of Advice Independence and Verbal Versus Numeric Format

Jeremy D. Strueder and Paul D. Windschitl
Department of Psychological and Brain Sciences, University of Iowa

Past research on advice-taking has suggested that people are often insensitive to the level of advice independence when combining forecasts from advisors. However, this has primarily been tested for cases in which people receive numeric forecasts. Recent work by Mislavsky and Gaertig (2022) shows that people sometimes employ different strategies when combining verbal versus numeric forecasts about the likelihood of future events. Specifically, likelihood judgments based on two verbal forecasts (e.g., "rather likely") are more often extreme (relative to the forecasts) than are likelihood judgments based on two numeric forecasts (e.g., "70% probability"). The goal of the present research was to investigate whether advice-takers' use of combination strategies can be sensitive to advice independence when differences in independence are highly salient and whether sensitivity to advice independence depends on the format in which advice is given. In two studies, we found that advice-takers became more extreme with their own likelihood estimate when combining forecasts from advisors who use separate evidence, as opposed to the same evidence. We also found that two verbal forecasts generally resulted in more extreme combined likelihood estimates than two numeric forecasts. However, the results did not suggest that sensitivity to advice independence depends on the format of advice.

Public Significance Statement
An important factor when combining forecasts from advisors is whether advisors are relying on the same or separate evidence to generate their forecasts. People have often been found to be insensitive to such differences in advisor independence. However, in the present work, we found that people can indeed use information about advice independence in a normative matter when differences in advice independence are sufficiently salient. Importantly, we also found that this sensitivity was not moderated by the format in which forecasts are presented. Participants in our studies attended to advice independence both when forecasts were in a verbal format (e.g., rather likely) and when both were in a numeric format (e.g., 70% probability). Replicating prior work (Mislavsky & Gaertig, 2022; Teigen et al., 2023), we also found evidence of a general advice format effect: When combining verbal as opposed to numeric forecasts, people's final likelihood estimates became more extreme (closer to certainty).

Keywords: forecasting, verbal probabilities, advice independence, predictions

Supplemental materials: https://doi.org/10.1037/xge0001611.supp

When trying to anticipate the future and make decisions under uncertainty, people often seek forecasts, perhaps from multiple sources. For instance, a bettor might search prediction websites before placing a significant bet, or a patient might consult with multiple specialists to determine the likelihood of a treatment being successful. In such cases, decision-makers are faced with an important question: How to combine the forecasts? As with most questions of this nature, the answer is dependent on many factors that can change across

Aysecan Boduroglu served as action editor.

Jeremy D. Strueder https://orcid.org/0009-0006-1155-8482

This work was supported by Grant SES-1851738 to Paul D. Windschitl and Andrew Smith from the National Science Foundation and a Graduate & Professional Student Government Research Grant to Jeremy D. Strueder. The authors have no conflicts of interest to disclose.

Data related to this work were presented at the 2022 Society for Judgment and Decision Making conference in San Diego, CA, and the Summer Institute on Bounded Rationality at the Max Planck Institute for Human Development in Berlin, Germany. Study 1 was preregistered at https://aspredicted.org/qt3p4.pdf. Study 2 was preregistered at https://aspredicted.org/b4p3y.pdf. Data, code, and materials are available at https://researchbox.org/1748.

Jeremy D. Strueder played a lead role in writing–original draft and an equal role in conceptualization, formal analysis, investigation, methodology, and writing–review and editing. Paul D. Windschitl played a supporting role in writing–original draft and an equal role in conceptualization, formal analysis, investigation, methodology, and writing–review and editing.

Correspondence concerning this article should be addressed to Jeremy D. Strueder or Paul D. Windschitl, Department of Psychological and Brain Sciences, University of Iowa, Iowa City, IA 52242, United States. Email: jeremy-strueder@uiowa.edu or paul-windschitl@uiowa.edu

contexts and environments (Winkler et al., 2019). The current research aims to examine two particular factors that are relevant to advice-takers: The level of independence between advisors and the format in which advice is presented.

Past research has shown that people have a strong tendency to average forecasts from others when making likelihood estimates (Budescu & Yu, 2006, 2007) and, in many cases, this can be the normative strategy to pursue (Hogarth, 1978). Studies have shown that averaging forecasts can lead to more accurate judgments and predictions about future events (e.g., Galton, 1907; Larrick & Soll, 2006; Wallsten et al., 1997; see Kerr & Tindale, 2011, for review), and it has generally been found that averaging the forecasts of three–six advisors can be enough to significantly reduce random error and even measurement bias (Clemen, 1989; Winkler & Poses, 1993). Empirical evidence also indicates that simple averages of advisor forecasts perform quite well and are oftentimes on par with more complex aggregation methods such as weighted averaging strategies (Makridakis, 1989). Therefore, it seems that rational agents would do well to adopt a simple averaging heuristic when combining forecasts from multiple advisors. However, this is not always the normatively correct strategy to pursue.

Advice Independence: Why It Should Affect Advice Combination Strategies

A key factor that can significantly impact the added value that comes from receiving an additional piece of advice is the level of independence that advisors and their sources have. When two advisors provide forecasts, those forecasts can be correlated to each other to varying degrees. Advisors can rely on the *same evidence* to provide their estimates, or they can use *separate evidence* as the basis for their forecasts. Generally, it can be assumed that the more independent a second forecast is from the first one, the more added value it provides (Mellers et al., 2014; Wallsten & Diederich, 2001). For example, for someone who is trying to estimate the likelihood that the Seattle Seahawks will have a winning record next season, looking at forecasts from an offensive scout and a defensive scout will be more informative than looking at forecasts from two offensive scouts. Importantly, advice combination strategies should therefore adapt to the level of advice independence. When a second forecast is based on new information that was not reflected by the first forecast—and the two forecasts point in the same direction—a person should be more certain about the target event than the average of the two forecasts would suggest (Mellers et al., 2014). This means that when advice independence is high, a person who uses a straight averaging strategy will arrive at an estimate that is too conservative; their estimate should be more extreme than the average of the advisor forecasts. The question now becomes are people attuned to differences in advice independence?

In prior work, Budescu and colleagues (Budescu et al., 2003; Budescu & Rantilla, 2000; Budescu & Yu, 2007) proposed a formal model of the determinants of confidence for decision-makers who combine multiple forecasts from advisors. Among these determinants, they list the total amount of information that is available to decision-makers, the level of consensus among advisors, and the level of intercue correlation between advisors (Budescu & Yu, 2007). Intercue correlation, in this case, refers to the degree to which the set of cues that one advisor is relying on is correlated with the set of cues that the other advisor is using; low intercue correlation is essentially high advice independence. The model predicts that decision-makers' confidence in advice aggregates should increase as the *intercue correlation decreases* (i.e., as the relatedness of the evidence that advisors are using decreases). However, in studies that tested the predictions of this model, people were shown to be largely insensitive to differences in intercue correlations (Budescu & Yu, 2007). These findings are in line with other research indicating that people often struggle to use information about correlations when forming opinions (Maines, 1996; Schulz-Hardt et al., 2022; Soll, 1999).

As noted by Budescu and Yu (2007), the *system neglect hypothesis* proposed by Massey and Wu (2005) offers some insight into why people may be insensitive to intercue correlations. The account posits that people overly focus on the external cues that are generated by systems and tend to neglect second-layer cues about the system itself. This is primarily because the external signals generated by a system are salient and easy to access, whereas second-layer cues, such as intercue correlations, are often not observable and thus need to be inferred. Within the context of advice-taking, the system neglect hypothesis predicts that overt factors such as the extremeness of advice, agreement between advisors, and advisor confidence take precedence over second-level parameters like advice independence. Interestingly, a system neglect account does not entirely rule out the possibility of advice independence factoring into decision-makers' advice aggregation strategies. It is possible that, in situations where intercue correlations are made sufficiently salient, decision-makers will incorporate the level of independence between advisors when combining multiple forecasts.

Advice Format: How It Could Relate to the Issue of Advice Independence

Past work on people's sensitivity to advice independence—as summarized above—has relied almost exclusively on situations in which the advice being received is numeric. However, there are two formats in which advice is typically given: Verbally (e.g., it is very likely that event × will occur) or numerically (e.g., there is an 80% probability that event × will occur). An extensive amount of research has been dedicated to identifying differences between the two formats across a broad range of domains (e.g., Beyth-Marom, 1982; Harris et al., 2013; Jenkins et al., 2018; Juanchich & Sirota, 2020; Lichtenstein & Newman, 1967; Teigen et al., 2014; Windschitl & Wells, 1996; see P. J. Collins & Hahn, 2018, for review), and mapping verbal likelihood statements to numeric ones (e.g., Hamm, 1991; Juanchich et al., 2013; Reagan et al., 1989; Wallsten et al., 1986; see Dhami & Mandel, 2022, for review).

Recently, Mislavsky and Gaertig (2022) revealed an intriguing finding about differences in how people use verbal and numeric likelihood advice. They found that advice that is presented verbally is often combined differently than advice that is presented numerically. Specifically, when people are given two verbal estimates (vs. two numeric estimates), they are more likely to combine these verbal estimates in a manner that was tentatively called counting—such that when two advisors both give estimates of "likely," the recipient of that advice gives a final estimate that is *more extreme* than "likely." "More extreme" refers to a final answer that is closer to certainty (i.e., closer to a scale endpoint and away from the scale midpoint, which reflects uncertainty). Given that the advisors both said "likely," a final response that reflects even more certainty that the event will

happen would be considered more extreme. If the advisors had both said "unlikely"—suggesting the event will not happen—a final response that reflects even more certainty that the event will *not* happen would be considered more extreme. This counting—or getting more extreme—may have to do with people's sensitivity to the directionality implied in verbal likelihood terms, as noted by Teigen et al. (2023). Terms such as "likely" not only suggest (vaguely) a probability value, but they reveal a positive focus—a focus on the event's possibility of happening (Teigen, 1988; Teigen & Brun, 1999). When people encounter these sorts of positive-directionality statements, the combined implications could lead to more extreme likelihood estimates. Conversely, combined estimates based on negative-directionality phrases tend to produce the opposite effect (Teigen et al., 2023). This overall pattern is termed the *reinforcement effect* by Teigen and colleagues. Critically, counting (or reinforcement) seems to rarely happen when the advice is given in numeric form. Instead, people predominantly average numeric advice from others (Budescu & Yu, 2006, 2007; Mislavsky & Gaertig, 2022)—although we note that the reinforcement hypothesis does not preclude numeric advice from producing a reinforcement effect in cases where numbers carry directionality, and evidence of numeric probabilities carrying some positive directionality has been observed by others (Teigen & Brun, 1995, 2003).

Our hypothesis for the current project was inspired by the findings from Mislavsky and Gaertig (2022). We saw their results as raising an intriguing possibility for the issues of whether and when people are sensitive to advice independence when receiving forecasts. Specifically, we reasoned that advice independence and format might interact in determining forecast-combination strategies. Whereas prior work indicates a general insensitivity to advice independence, it may be the case that this insensitivity stems from combination strategies that are associated with numeric advice—and does not generalize to instances where people combine verbal advice. We suspected that advice-takers who receive numeric forecasts may rigidly default to learned averaging strategies. As adults, people have a limited set of operations they can envision doing with two numbers—one of which is averaging. Imagine a situation in which a person receives two numeric probabilities from independent advisors. Even if the person has a vague intuition that says the two estimates are separately useful, the person probably does not have a trained and readily available algorithm for how to proceed in combining the two pieces of advice, but they do know how to average the numbers, and that might be an appealing approach by comparison. In contrast, imagine a situation in which that person was given two verbal estimates rather than two numbers. Here, the mathematical operation of averaging is not as clear and salient (Mandel et al., 2021). In this case, the person may be more willing to deviate from an averaging strategy (see Collsiöö et al., 2023, for similar reasoning regarding format dependence). The result is that advice independence might have more impact for combining verbal, as opposed to numeric, estimates. In this project, we directly examined the intersection of advice independence and format to address this possibility.

The Present Studies and Paradigm Overview

For the present pair of studies, we aimed to test whether people's combinations of advisor estimates were sensitive to advice independence and whether the impact of advice independence differed as a function of advice format (and whether format had a main effect). To create a strong test, we created an experimental paradigm that made the level of advice independence (interadvisor correlations) especially clear and salient. We preemptively note that this manipulation of advice independence, as described below, differs from past work on advice. Whereas other studies are often vague on how much additional information is contained in a second, independent forecast, we chose a paradigm in which participants could directly infer how much additional information was contained in a second independent versus dependent forecast. This will become relevant later on in our discussion of the present results.

Participants in our studies were asked to give likelihood estimates for hypothetical tennis doubles matches at a company retreat. They had to estimate the likelihood that a particular pair of teammates from a company, who were randomly drawn to create the team, would win an upcoming match against a random team from another company. This novel scenario, in which players were assigned to play together as teammates, allowed us to emphasize that players within a team were randomly placed together and could vary dramatically in skill. This meant that information regarding the skill of one player on the team was entirely uncorrelated with information about the skill of the other player on the team. This was a key component of our advice independence manipulation. Namely, on each trial, participants received forecasts from two tennis scouts who had either each looked at a different player within a doubles team (*separate evidence*) or each looked at both players within a team (*same evidence*). We also manipulated whether the two advisor estimates were numeric or verbal. By measuring how participants adjusted their likelihood estimates after receiving a second piece of advice, we could directly test whether differences in advice independence impacted combination strategies—and whether sensitivity to independence differed as a function of the advice format.

In Study 1, we tested advice combination strategies for situations in which the two advisors provided identical forecasts. For Study 2, we expanded the test to situations in which advisors provided forecasts that differed slightly from one another.

Transparency and Openness

All experiments were approved by the University of Iowa Institutional Review Board and all participants provided informed consent. All data exclusions, manipulations, and measures are reported in the article and Supplemental Materials. Preregistrations, data, materials, and code for Studies 1 and 2 are publicly available on ResearchBox (Strueder & Windschitl, 2024): https://researchbox.org/1748.

Study 1

In Study 1, we focused on instances in which the two advisors provided identical forecasts (e.g., "Rather Likely"). We used identical forecasts to avoid adjustment strategies being dominated by the level of agreement between advisors or the directionality of the second forecast relative to the first one. The key manipulations in the study (and in Study 2) were of advice independence and format. We predicted that participants' adjustment strategies would be sensitive to the level of independence between advisors—but that sensitivity to advice independence would be moderated by the format in which advice was presented. We also expected a format

main effect, broadly consistent with Mislavsky and Gaertig (2022). Study 1 was preregistered on AsPredicted at https://researchbox.org/1748.

Method

Participants and Design

We preregistered that we would aim to collect participants until a sample size of 240 was reached, which would give us at least 80% power to detect a small- to medium-sized effect for all relevant analyses. In total, 326 MTurk participants completed the study. Of those participants, 86 failed to pass at least one of two attention checks that were included in the study, resulting in a final sample size of 241 (131 male, 109 female, one nonbinary/third gender, M_{age} = 39.6). Participation was estimated to last around 7–8 min, and participants were paid $0.80 to complete the study.[1] We used a 2 (advice independence: same evidence vs. separate evidence) × 2 (advice format: verbal vs. numeric) × 2 (forecast side: both advisors' forecasts were either below or above the midpoint of the likelihood scale they were said to use) mixed-factor design. Each participant completed eight trials (four critical, four filler), which were split into a *same evidence* block and a *separate evidence* block (counterbalanced order).

Procedure, Materials, and Measures

Upon agreeing to participate, participants received basic information about the aforementioned hypothetical scenario in which a tech firm called "Booum" hosted a company retreat. As part of the company retreat, employees would practice tennis and compete against another firm in doubles matches. Critically, the instructions made it clear to participants that teammates for a given doubles team were randomly assigned and could vary in skill, age, and gender. Participants were then informed that their task would be to estimate the likelihood of each Booum team winning its matchup. To help them with their estimate, they were told that they would receive advice from two tennis scouts who watched the players train beforehand. Participants were also shown the scale that scouts used to provide their forecasts. In the numeric condition, the scale included 11 numeric labels (1 = 0% probability, 11 = 100% probability), whereas in the verbal condition, the scale included 11 verbal labels (1 = entirely unlikely, 11 = entirely likely). After seeing the scale, participants answered two comprehension questions about the numeric/verbal scale. Specifically, they were shown three labels that were all above the scale midpoint (Q1) or all below the midpoint (Q2) and asked to select the label that reflected the highest (Q1) or lowest (Q2) probability.[2]

Participants were then shown the first of eight hypothetical tennis matches, which were all structured the same way. For each match, participants first saw an image of the team that was playing in the tennis matchup and then sequentially received forecasts from two scouts regarding the doubles team's chances of winning its matchup (see Figure 1). Participants first provided an initial likelihood estimate regarding the doubles team's chances of winning its matchup based solely on the first scout's forecast. They then received the second scout's forecast and made a second likelihood estimate based on both forecasts. Both of these likelihood estimates were always made on the same scale. Specifically, the estimates were made on a 9-point scale with end labels (1 = *terrible chance*; 9 = *excellent chance*).

The eight tennis matches that participants saw were split into a same evidence block and a separate evidence block (counterbalanced order). For the same evidence block, participants were told that the scouts each watched both players in the team train beforehand and, therefore, were relying on the same evidence when providing their forecasts. For the separate evidence block, participants were told that each scout looked at a different player within the doubles team when estimating the likelihood of the team winning its matchup. Each block contained two critical trials—one above the midpoint and one below the midpoint—and two filler trials[3] (see Table 1 for a full list of forecast combinations that were used on critical trials). Which of the critical trials from Table 1 were assigned to the same evidence and separate evidence blocks was counterbalanced. Advisor forecasts were either all verbal or all numeric, depending on the condition. On a given trial—as a function of the forecast side manipulation—both advisors' forecasts were either above the scale midpoint or below.

After completing the eight trials, participants answered a series of exit questions regarding the number of advisors they were shown, the evidence that advisors were basing their forecasts on, and their assumptions about the players' skills. See the Supplemental Materials for reporting of these exit measures (Section A) and reporting of a preregistered exploratory measure that was included at the end of the experiment (Section B). Last, participants answered two demographic questions ("What is your age?"; "What is your gender?"). When reporting gender, participants could choose from three default options (male, female, nonbinary/third gender) or select to fill in a free-response box.

Results

Preliminary Rates (and Definition) of Extreme Adjustments

As preregistered, for some analyses, we dichotomized responses into those who became more extreme after receiving a second piece of advice and those who did not become more extreme after receiving a second piece of advice. This operationalization of "more extreme" is similar to Mislavsky and Gaertig (2022). For above-midpoint trials, a participant was classified as becoming more extreme, if their second estimate (based on both forecasts) was closer to the upper end of the scale than their first estimate (based only on one forecast). For below-midpoint trials, a participant was classified as becoming more extreme, if their second estimate was closer to the lower end of the scale than their first estimate. In general, participants adjusted their initial likelihood estimates after receiving the second advisor forecast on 42.7% of trials.[4] On trials

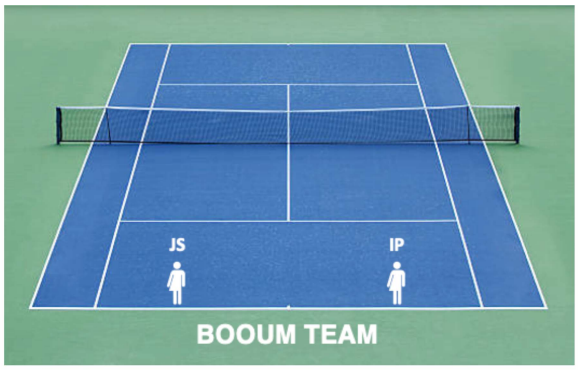
[1] The median completion times for Study 1 and Study 2 ended up being 6.92 and 7.23 minutes, respectively.

[2] Scale comprehension was generally high across the two advice format conditions, with 83% of participants in the verbal condition and 94% of participants in the numeric condition responding correctly to both comprehension questions.

[3] On filler trials, advisor forecasts differed from one another and could fall on opposite sides of the midpoint. This was done to avoid creating an impression that advisors were always in agreement. We randomized the order in which critical and filler trials appeared.

[4] See Supplemental Section D for the distribution of participants' likelihood estimates based on one and both advisor forecasts.

Figure 1
Example Trial From Study 1 (Numeric Condition)



Here is also the second scout's estimate:

	Looked at	Said (about this team's likelihood of winning)
Scout 1	**JS**	"70% probability"
Scout 2	**IP**	"70% probability"

Given the information from both scouts, how would you assess team JS&IP's chance of winning?

Terrible ○ ○ ○ ○ Even ○ ○ ○ ○ Excellent

Note. On each trial, participants were sequentially shown forecasts from two advisors. After seeing the first forecast and giving an initial estimate, participants were shown the second forecast and asked to give a second estimate based on both forecasts (depicted above). Original tennis court image by https://www.istockphoto.com/photo/empty-tennis-hard-court-gm182930701-14131451. See the online article for the color version of this figure.

where participants made likelihood adjustments, the majority of adjustments (81.3%) were toward the extreme ends of the scale (i.e., upward on above-midpoint trials and downward on below-midpoint trials). Only 18.7% of likelihood adjustments were in the less extreme direction.

When Did Participants Adjust Toward the Extreme?

See Figure 2 for a summary of the proportions of times—by condition—that participants became more extreme after receiving a second forecast. For the first of our preregistered analyses, we submitted the dichotomized extreme-or-not variable to a probit regression, with advice independence and advice format as key predictors. Standard errors were clustered at the participant level. Overall, when collapsing across above- and below-midpoint trials, participants were more likely to adjust toward the extreme when advisors were using separate evidence as opposed to the same evidence, which indicates that participants were indeed sensitive to differences in advice independence, $Z = 2.89$, $p = .004$. Participants were also generally more likely to make extreme adjustments for verbal advice compared to numeric advice, $Z = 3.18$, $p = .001$. This latter finding is consistent with the advice format effect observed by Mislavsky and Gaertig (2022). However, contrary to our prediction that advice format would moderate sensitivity to advice independence, we did not observe a significant interaction between advice independence and advice format, $Z = -0.22$, $p = .823$. In other words, whether or not participants took into consideration the level of independence between advisors when combining their estimates did not depend on the format in which advice was given.

As preregistered, we also conducted the same analysis separately for above and below-midpoint trials. Looking just at above-midpoint trials, we again found a main effect of advice independence, $Z = 2.04$, $p = .042$. More participants adjusted toward the extreme (i.e., higher) when advisor forecasts were based on independent advice. Somewhat surprisingly, the advice format effect—while directional—was not significant when we restricted the analysis to above-midpoint

Table 1
Full List of Forecast Combinations That Were Used on Critical Trials in Studies 1 and 2

Study	Above/below midpoint	Critical trial	Condition: Verbal	Condition: Numeric
Study 1	Above midpoint	1	"Rather Likely"	"70% probability"
			"Rather Likely"	"70% probability"
		2	"Quite Likely"	"80% probability"
			"Quite Likely"	"80% probability"
	Below midpoint	3	"Rather Unlikely"	"30% probability"
			"Rather Unlikely"	"30% probability"
		4	"Quite Unlikely"	"20% probability"
			"Quite Unlikely"	"20% probability"
Study 2	Above midpoint	1	"Somewhat Likely"	"60% probability"
			"Rather Likely"	"70% probability"
		2	"Rather Likely"	"70% probability"
			"Quite Likely"	"80% probability"
		3	"Quite Likely"	"80% probability"
			"Rather Likely"	"70% probability"
		4	"Quite Likely"	"80% probability"
			"Very Likely"	"90% probability"
	Below midpoint	5	"Somewhat Unlikely"	"40% probability"
			"Rather Unlikely"	"30% probability"
		6	"Rather Unlikely"	"30% probability"
			"Quite Unlikely"	"20% probability"
		7	"Quite Unlikely"	"20% probability"
			"Rather Unlikely"	"30% probability"
		8	"Quite Unlikely"	"20% probability"
			"Very Unlikely"	"10% probability"

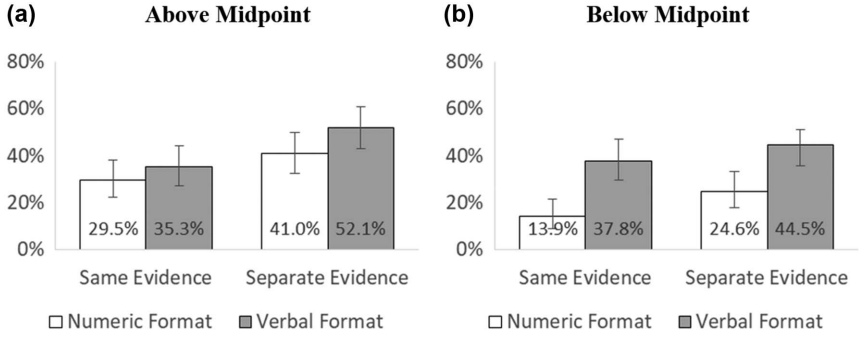
Note. In both studies, we fully counterbalanced which forecast combinations appeared on same versus separate evidence blocks. Moreover, in Study 2, the order in which forecasts were presented on a trial was counterbalanced as well.

trials, $Z = 0.96$, $p = .338$. As before, there was no interaction between the two factors, $Z = 0.58$, $p = .559$. For below-midpoint trials, more participants made extreme adjustments (i.e., downward) when advisor forecasts were based on separate evidence and in a verbal format; $Z = 2.36$, $p = .018$; and $Z = 4.21$, $p < .001$; respectively. The two factors did not interact, $Z = -1.01$, $p = .312$.

Analysis of Mean Likelihood Adjustments Across Conditions

We also tested for mean differences in participants' adjustments of their likelihood estimates. To do this, we created *likelihood-adjustment scores* by computing the extent to which participants' estimates based on both scouts differed from their estimates based solely on the first scout—with scores for adjustments made in the more extreme direction always given a positive sign and scores for adjustments made in a less extreme direction always given a negative sign. For instance, on above-midpoint trials, an adjustment score of +2 would indicate that a participant's second estimate was two scale points higher than their first estimate (more extreme), while a −1 would indicate that their second estimate was 1 point lower than their first estimate (less extreme). The opposite would apply for below-midpoint trials (i.e., positive numbers indicate becoming more extreme in the negative direction). We then submitted these

Figure 2
Study 1: Percentage of Participants in Each Condition That Adjusted Toward the Extreme After Receiving a Second Likelihood Forecast From the Second Advisor



Note. Error bars represent 95% confidence intervals.

likelihood-adjustment scores to mixed-factor analyses of variance (ANOVAs; as preregistered, once for all trials but also separated by whether the trials involved advice that was above or below the scale midpoint).

Looking first at trials collapsed across both sides of the scale midpoint—in a 2 (advice independence: same vs. separate) × 2 (advice format: verbal vs. numeric) × 2 (forecast side: below-midpoint vs. above-midpoint) mixed-factor ANOVA—we found that the magnitude of participants' likelihood adjustments differed as a function of advice independence and format. Participants made larger likelihood adjustments (toward the extreme) for separate evidence trials ($M = 0.38$, $SD = 1.05$) than for same evidence trials ($M = 0.22$, $SD = 0.79$), $F(1, 239) = 9.47$, $p = .002$, $adj\ \eta_p^2 = .034$.[5] They also made larger likelihood adjustments (toward the extreme) for verbal advice ($M = 0.45$, $SD = 0.83$) compared to numeric advice ($M = 0.15$, $SD = 1.00$), $F(1, 239) = 13.89$, $p < .001$, $adj\ \eta_p^2 = .051$. The effect of advice independence on likelihood adjustments did not differ across advice format, $F(1, 239) = 0.04$, $p = .848$, $adj\ \eta_p^2 = .000$. Last, the ANOVA also revealed a significant interaction between advice format and forecast side, $F(1, 239) = 6.92$, $p = .009$, $adj\ \eta_p^2 = .024$.

In an ANOVA involving just above-midpoint trials, we still find a significant main effect of advice independence, $F(1, 239) = 9.95$, $p = .002$, $adj\ \eta_p^2 = .036$. On average, likelihood adjustments were larger in the separate evidence condition ($M = 0.48$, $SD = 1.06$) than in the same evidence condition ($M = 0.25$, $SD = 0.75$). The effect of advice format was directional but not significant, with verbal advice ($M = 0.45$, $SD = 0.81$) resulting in similar—perhaps slightly larger—likelihood adjustments compared to numeric advice ($M = 0.28$, $SD = 1.02$), $F(1, 239) = 3.36$, $p = .068$, $adj\ \eta_p^2 = .010$. There was no significant interaction between advice independence and advice format, $F(1, 239) = 0.11$, $p = .745$, $adj\ \eta_p^2 = .000$.

For below-midpoint trials, the pattern of results was slightly different. While there was a main effect of advice format (verbal: $M = 0.87$, $SD = 1.05$; numeric: $M = 0.58$, $SD = 1.28$), $F(1, 239) = 20.70$, $p < .001$, $adj\ \eta_p^2 = .076$, the effect of advice independence was not significant for trials below the midpoint, $F(1, 239) = 2.13$, $p = .146$, $adj\ \eta_p^2 = .005$. Last, there was no significant interaction between the two factors, $F(1, 239) = 0.002$, $p = .968$, $adj\ \eta_p^2 = .000$.

Discussion

The results of Study 1 show that people can indeed be sensitive to the level of independence between advisors when intercue correlations are sufficiently salient. Participants showed a higher tendency to make extreme likelihood adjustments after they received a second forecast from an advisor who used entirely separate evidence from the first advisor, as opposed to the same evidence. Participants also exhibited a general tendency to make more extreme adjustments for verbal forecasts, rather than numeric forecasts, broadly replicating the format effect found by Mislavsky and Gaertig (2022). Another important finding is that the format in which advice was given did not significantly affect participants' sensitivity to the independence of advice when combining forecasts. While we suspected that numeric advice could trigger rigid averaging strategies, participants in the numeric advice condition were equally sensitive to differences in advice independence.

Study 2

Study 1 focused exclusively on cases in which advisors provide identical forecasts. However, it is often the case that advisors, even expert ones, do not come to the exact same conclusions when generating forecasts. Therefore, for Study 2, we expanded our analysis to instances in which advisor forecasts differ from one another. While we still predicted that people would be sensitive to differences in advice independence and that the format of advice would affect combination strategies, we thought it was possible that advice order could dominate participants' likelihood adjustments—thereby canceling out any format or advice independence effects. One reason for this is that, unlike in Study 1 where any adjustment meant deviating from an averaging strategy, upward/downward adjustments could be consistent with an averaging strategy in Study 2. For example, someone who first receives a forecast of 70% probability and then a forecast of 80% probability would have to adjust upward to average the two forecasts. Relatedly, prior work on combining forecasts—which also investigated the role of advice format—found that the advice format effect was partially masked by an effect of presentation order (Teigen et al., 2023). This is perhaps because participants viewed differing forecasts as indicating a trend, which then dominated the likelihood-adjustment direction (Hohle & Teigen, 2015, 2018; Juanchich et al., 2010). In short, when the second advisor's forecast is slightly different from the first, the directions of adjustments could be highly shaped by averaging or trend perception, which might diminish our ability to detect robust effects of advice independence (or format). This might be especially true for analyses that consider only the direction of adjustment; for analyses that also involve the *magnitude* of adjustment, the impact of independence and format would presumably have a greater chance of being observed.

Study 2 was preregistered on AsPredicted at https://researchbox.org/1748.

Method

To reach our preregistered sample size of 240, we recruited 360 Mturkers. Of those, 238 (139 male, 97 female, three nonbinary/third gender; $M_{age} = 38.8$) passed the preregistered attention checks and were included in the Study, again giving us at least 80% power to detect a small- to medium-sized effect for all relevant tests. Participation was estimated to last around 7–8 min, and each participant was paid $0.80 to participate. We again employed a 2 (advice independence: same evidence vs. separate evidence) × 2 (advice format: verbal vs. numeric) × 2 (forecast side: below-midpoint vs. above-midpoint) mixed-factor design, with additional counterbalancing factors (see below). As with Study 1, the advice format factor was the only between-subjects factor. The only difference between Studies 1 and 2 was that, for Study 2, advisors provided nonidentical forecasts on critical trials. Specifically, each participant completed 12 trials (eight critical, four filler), which were split into a *same evidence* block and a *separate evidence* block (counterbalanced order). On critical trials, advisors' forecasts always differed from one another by one scale point (e.g., 70% probability and 80% probability/Rather Likely and Quite Likely)

[5] Throughout the article, we report the adjusted version of partial eta squared, as established in Mordkoff (2019), except in places where the adjustment would drop the value below zero.

and we counterbalanced the order in which the more extreme forecast and less extreme forecast were presented. Otherwise, the procedure and measures of Study 2 were identical to those of Study 1.[6]

Results

Preliminary Rates (and Definition) of Extreme Adjustments

Our operationalized definition of a more extreme response remained the same in Study 2 as in Study 1 (despite the fact that the two advisor forecasts were always the same in Study 1 but different in Study 2). For above-midpoint trials, a second estimate (based on both forecasts) that was closer to the upper end of the scale than first estimate was classified as "more extreme." For below-midpoint trials, a second estimate that was closer to the lower end of the scale than first estimates was classified as "more extreme." Participants adjusted their initial likelihood estimates after receiving the second advisor forecast on 62.6% of trials. Of those adjustments, 67.6% were toward the more extreme, while 32.4% were toward the less extreme. See Supplemental Section F for a full plot of the percentages of participants that adjusted toward the more extreme, less extreme, or made no adjustment; split by condition.

When Did Participants Adjust Toward the Extreme?

See Figure 3 for a summary of the proportions of times—by condition—that participants became more extreme after receiving a second forecast. As in Study 1, we submitted the dichotomized extreme-or-not variable to a probit regression with advice independence and advice format as key predictors. Standard errors were clustered at the participant level. In the overall analysis (i.e., collapsing across forecast side), participants were sensitive to advice independence. Upon receiving the second forecast, participants were more likely to adjust toward the extreme when advisors were using separate evidence for their forecasts, as opposed to the same evidence, $Z = 2.46$, $p = .014$. Unlike in Study 1, the format effect was not significant in this probit analysis, $Z = 0.74$, $p = .460$. As in Study 1, the interaction between advice independence and format was not significant, $Z = -0.74$, $p = .456$. Unsurprisingly, for Study 2, there was a large effect of advice order, $Z = -9.62$, $p < .001$. Participants were more likely to make an extreme adjustment when the second advisor's forecast was more extreme than the first advisor's forecast. Notably, the effect of order significantly interacted with advice format, $Z = 3.70$, $p < .001$. We discuss this interaction in more detail in the following two paragraphs.

Focusing only on above-midpoint trials, there was a similar pattern of results to the one observed for the collapsed analysis. A larger proportion of participants made extreme adjustments on trials where advisors used separate evidence, $Z = 2.18$, $p = .029$. There was no significant effect of format and no interaction between format and advice independence; $Z = 0.18$, $p = .861$; and $Z = -0.28$, $p = .782$; respectively. Advice order had a large impact on adjustment strategies, $Z = -7.14$, $p < .001$, and significantly interacted with advice format, $Z = 2.32$, $p = .020$. Specifically, the advice format effect was only present when participants received the more extreme advisor forecast first.[7] There was no interaction between advice order and advice independence, $Z = 0.49$, $p = .624$.

For below-midpoint trials, there was no effect of advice independence or advice format; $Z = 1.48$, $p = .139$; and $Z = -0.93$, $p = .355$; respectively. There was also no significant Independence × Format interaction, $Z = -0.84$, $p = .402$. Instead, we again found that advice order was the primary factor affecting response strategies, with a larger proportion of participants making extreme adjustments on trials where the extreme advice came second, $Z = -7.20$, $p < .001$. Advice order also again interacted with advice format, $Z = 3.23$, $p = .001$. The advice format effect was only significant when the more extreme advisor forecast came first. Finally, the interaction between order and advice independence was not significant, $Z = 1.79$, $p = .074$.

Analysis of Mean Likelihood Adjustments Across Conditions

For analyses that also were sensitive to the magnitude of adjustment—not just direction—we again start by examining effects collapsed across forecast side. Participants showed a general tendency to make larger extreme likelihood adjustments for separate evidence ($M = 0.36$, $SD = 1.17$) than for same evidence ($M = 0.22$, $SD = 1.11$) trials, $F(1, 236) = 7.36$, $p = .007$, $adj\ \eta_p^2 = .026$. Moreover, like in Study 1, there was a main effect of advice format, with participants making larger extreme adjustments for verbal advice ($M = 0.39$, $SD = 1.22$) than for numeric advice ($M = 0.19$, $SD = 1.06$), $F(1, 236) = 9.03$, $p = .003$, $adj\ \eta_p^2 = .033$. As with Study 1, there was no significant interaction between advice independence and format, $F(1, 236) = 1.36$, $p = .245$, $adj\ \eta_p^2 = .002$. There was no interaction between forecast side and format, $F(1, 236) = 0.00$, $p = .998$, $adj\ \eta_p^2 = .000$. In addition to the independence and format main effects, there was also a main effect of forecast side, $F(1, 236) = 8.22$, $p = .005$, $adj\ \eta_p^2 = .030$. The magnitude of extreme likelihood adjustments was generally larger for above-midpoint trials ($M = 0.36$, $SD = 1.08$) than for below-midpoint trials ($M = 0.22$, $SD = 1.20$). Last, there was a large main effect of advice order, $F(1, 236) = 162.62$, $p < .001$, $adj\ \eta_p^2 = .405$, and the effect of order significantly interacted with advice format, $F(1, 236) = 10.93$, $p = .001$, $adj\ \eta_p^2 = .040$. We will elaborate on this when looking separately at the above- and below-midpoint trials.

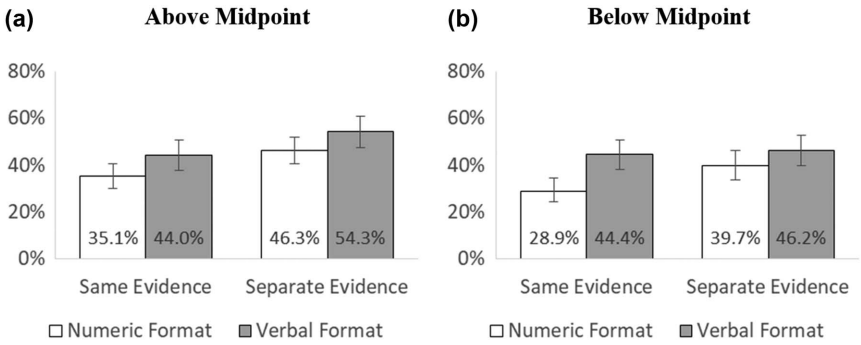
For above-midpoint trials (i.e., trials where both advisors gave forecasts that were above the scale midpoint), we found that the level of advice independence significantly affected likelihood adjustments, $F(1, 236) = 11.80$, $p < .001$, $adj\ \eta_p^2 = .044$. Participants adjusted further toward the extreme when advisors used separate evidence ($M = 0.47$, $SD = 1.07$) as opposed to the same evidence ($M = 0.25$, $SD = 1.08$). Again, there was no interaction between advice independence and advice format, $F(1, 236) = 0.96$, $p = .328$, $adj\ \eta_p^2 = .001$. Moreover, there was no significant interaction between advice independence and the order in which participants received the advisors' forecasts, $F(1, 236) = 0.02$, $p = .898$, $adj\ \eta_p^2 = .000$. Aside from the effect of advice independence, we also observed a significant main effect of advice format, $F(1, 236) = 7.01$, $p = .009$, $adj\ \eta_p^2 = .025$. Interestingly, however, the advice format effect significantly interacted with advice order, $F(1, 236) = 5.41$, $p = .021$, $adj\ \eta_p^2 = .018$. When receiving the more extreme

[6] See Supplemental Section E for reporting of exit questions for Study 2, as with Study 1.

[7] See Supplemental Materials Section G for subplots of Figure 3, which additionally split the results by advice order (i.e., trials where the more extreme forecast was given first vs. second).

Figure 3

Study 2: Percentage of Participants in Each Condition That Adjusted Toward the Extreme After Receiving a Second Likelihood Forecast From the Second Advisor



Note. Error bars represent 95% confidence intervals.

forecast first, there was a significant difference between likelihood adjustments in the verbal ($M = 0.20$, $SD = 1.11$) and numeric ($M = -0.16$, $SD = 0.98$) condition, $F(1, 236) = 11.33$, $p < .001$, *adj* $\eta_p^2 = .042$. When receiving the more extreme forecast second, the difference between the verbal ($M = 0.74$, $SD = 1.07$) and numeric ($M = 0.67$, $SD = 0.90$) condition was directional but not significant, $F(1, 236) = 0.41$, $p = .520$, *adj* $\eta_p^2 = .000$.

For below-midpoint trials, there was no significant effect of advice independence, $F(1, 236) = 0.75$, $p = .386$, *adj* $\eta_p^2 = .000$. Whether advisors relied on the same evidence ($M = 0.19$, $SD = 0.05$) or separate evidence ($M = 0.25$, $SD = 0.06$) did not impact participants' likelihood adjustments. In contrast, the advice format effect was again significant, with participants making more extreme adjustments for verbal advice than for numeric advice, $F(1, 236) = 5.14$, $p = .024$, *adj* $\eta_p^2 = .017$. There was no interaction between advice format and advice independence, $F(1, 236) = 0.65$, $p = .420$, *adj* $\eta_p^2 = .000$. However, as with above-midpoint trials, the Format × Order interaction was significant, $F(1, 236) = 6.93$, $p = .009$, *adj* $\eta_p^2 = .024$. The advice format effect was present when participants received the more extreme forecast first (verbal: $M = 0.13$, $SD = 1.25$; numeric: $M = -0.25$, $SD = 0.98$), $F(1, 236) = 11.33$, $p < .001$, *adj* $\eta_p^2 = .042$, but not when they received the extreme forecast second (verbal: $M = 0.51$, $SD = 1.33$; numeric: $M = 0.48$, $SD = 1.04$), $F(1, 236) = 0.09$, $p = .768$, *adj* $\eta_p^2 = .000$.

Discussion

Study 2 extends the findings from Study 1 to instances where advisors give nonidentical forecasts. The overall effect of advice independence was significant in the analyses of both direction and magnitude of adjustments. The overall effect of advice format was significant in the latter type of analyses. The fact that it was not significant in the analysis of only the direction of adjustment is not particularly surprising; as discussed earlier, the basic direction of adjustment can be heavily determined by the order in which forecasts are presented. Indeed, the effect of forecast order was significant. Finally, in both the tests of the direction and magnitude of adjustments, the effect of advice independence was not significantly affected by format.

General Discussion

The goal of the present research was to investigate whether advice-takers' strategies for combining forecasts can be sensitive to advice independence when differences in independence are highly salient and whether sensitivity to advice independence depends on the format in which advice is given. Regarding the first issue, we found that participants in both Studies 1 and 2 were sensitive to the level of independence between advisors. Specifically, participants were more prone to make extreme likelihood adjustments (i.e., become more certain in their predictions) after they received a second piece of advice that was based on different information as opposed to the same information. This was true when advisors provided identical forecasts (Study 1) but also when advisors provided differing forecasts (Study 2). In closely related analyses on the magnitude of adjustments, the conclusions were similar: Participants made larger adjustments when advisors used separate as opposed to the same evidence—though this pattern was somewhat attenuated when advisor forecasts were below the scale midpoint.

Participants' tendencies to become more extreme when combining independent forecasts and to average when combining dependent forecasts are consistent with normative theories for how to combine forecasts (Mellers et al., 2014; Wallsten & Diederich, 2001). As mentioned in our introduction, multiple forecasts that were based on different information provide a more complete picture than forecasts that are based on the same information and thus should result in higher confidence and combined likelihood estimates that are closer to certainty. Interestingly, previous research has primarily found that people are relatively insensitive to the level of independence between advisors and their forecasts (e.g., Budescu & Yu, 2007). We suspect that the discrepancy between our findings and those from prior work is due to the highly salient manipulation of advice independence that we employed in our studies. Participants in our studies were told that players were randomly paired, which allowed participants to easily recognize that information coming from scouts who looked at different players was entirely uncorrelated. Moreover, several other factors such as advisor confidence, the validity of advisors' cues, and the number of advisors, were fixed in our studies (compared to prior work), which likely made it easier for participants to focus on second-layer cues like advisor independence. It is quite possible

that sensitivity to advice independence diminishes, as the number of external—more salient—cues relevant to combining forecasts increases. Regardless, our participants' sensitivity to advice independence is an important finding, because it suggests that people may be capable of combining advice in a Bayesian-rational manner when they have a sufficiently clear understanding of the independence of advice.

How Does Advice Format Factor in?

Another main question of ours was whether sensitivity to advice independence could potentially differ as a function of the format in which the advice is presented. Many studies have documented systematic differences between how numeric and verbal probabilities are interpreted (Teigen & Brun, 1995, 1999; Windschitl & Wells, 1996), and one of our a priori predictions was that people's insensitivity to advice independence—as observed in other work (e.g., Budescu & Yu, 2007)—could be due to the numeric format in which advice is typically presented. However, we did not find any evidence of advice format moderating the degree to which people attend to advice independence. Participants' likelihood adjustments differed as a function of advice independence both when advisors gave numeric forecasts and when advisors gave verbal forecasts.

The finding that people adjust to differences in advice independence in a similar manner when combining verbal and numeric probability forecasts offers additional insights into how people combine advice. For example, it rules out the possibility that insensitivity to intercue correlations is due to people defaulting to rigid averaging strategies when combining numeric probabilities. The finding is also in line with prior work showing that verbal and numeric probability formats can lead to judgments of similar quality (e.g., Budescu & Wallsten, 1995; Wallsten et al., 1993). While we did find differences in combination strategies between formats (see below), participants were generally able to incorporate information about the independence of advisors in a normative manner both when advisors gave numeric and verbal forecasts. This is noteworthy, given that verbal probability forecasts are often criticized for their vagueness and inherent directionality (e.g., Collins et al., 2024; Dhami & Mandel, 2022). In the present work, we do not find evidence of these properties interfering with sensitivity to advice independence.

Notably, although sensitivity to advice independence did not differ as a function of advice format, we did find evidence of a general advice format effect. As previously shown by Mislavsky and Gaertig (2022), participants in Study 1 had a higher tendency to "count" verbal—as opposed to numeric—forecasts from advisors. In other words, participants more frequently gave combined likelihood estimates that were higher than the average of advisor forecasts, when they were presented in a verbal format rather than a numeric format. Additionally, in both Studies 1 and 2, the magnitude of extreme adjustments was, on average, larger for verbal than for numeric advice.

As a secondary note, we did find exceptions to the pattern found by Mislavsky and Gaertig (2022). For one, when advisors relied on separate evidence (and in some cases even the same evidence), a significant proportion of participants—in particular on above-midpoint trials—became more extreme with their own likelihood estimates even when forecasts were presented numerically. Cases of nonaveraging of numeric forecasts were also observed by Teigen et al. (2023), who attributed this in part to the fact that even numeric probabilities contain directional properties. Second, when advisors gave nonidentical forecasts (Study 2), our analyses that involved dichotomizations (of whether a participant did or did not become more extreme in the estimates) did not show a main format effect, even though our analysis of the magnitude of adjustment did show a format effect. We suspect that this was because, when receiving a second, more extreme forecast, participants adjusted toward the extreme for both formats—which was the normatively correct thing to do—thereby overshadowing a potential format effect (cf. Teigen et al., 2023). This is supported by the fact that—when dichotomizing responses into extreme versus nonextreme—the format effect was significant when participants received the more extreme forecast first, but not significant when participants received the more extreme forecast second.

In general, our findings are also compatible with Teigen et al.'s (2023) hypothesis that the counting of verbal forecasts is due to the directionality—not magnitude—of specific verbal likelihood phrases and that two verbal forecasts primarily reinforce each other when they have a similar directionality. In our studies, the directionality of advice was always positive for above-midpoint trials (e.g., quite likely) and negative for below-midpoint trials (e.g., quite unlikely). As such, our finding that participants tended to make more/larger upward adjustments on above-midpoint trials and downward adjustments on below-midpoint trials could be attributed to the directionality of verbal forecasts having a reinforcing effect. That said, we cannot conclusively determine that the advice format effect in our studies was solely due to the directionality and not the magnitude of verbal forecasts. This is because in our studies, we did not test instances where forecasts were low (high) in probability magnitude but positive (negative) in directionality. Thus, we cannot discriminate between the two causal explanations. However, Teigen et al. (2023) showed in their work that two verbal forecasts that convey low probabilities can actually lead to combined likelihood estimates that are less extreme when they have a positive directionality (e.g., a chance), which suggests that it is directionality (not probability magnitude) that drives the format effect.

Constraints on Generality and Future Directions

All of our studies were conducted with online samples of U.S. participants. It is possible that some of the observed effects—in particular, the advice format effect—differ across other cultures and languages. Others have found cultural differences in the interpretation of verbal probability expressions (e.g., Davidson & Chrisman, 1993, 1994; Doupnik & Richter, 2004), and these differences might extend to advice combination strategies. For instance, the degree to which forecasts reinforce each other could depend on how directional the verbal probabilities are perceived to be. If these perceptions differ across cultures, as has been suggested by prior work (Doupnik & Richter, 2004), then the same verbal probabilities could be combined differently depending on a person's cultural background.

Future research is also needed to test whether the effect of advice independence is substantially different across contexts or numbers of advisors. Our study involved a tennis context. Tennis is not particularly unusual nor do we expect that advice usage in a tennis context would be atypical. Nevertheless, future work should explore additional contexts to identify those in which advice independence could play an elevated or diminished role. For example, it is possible that domain familiarity could affect the extent to which people are

sensitive to advice independence. Our participants may have had enough familiarity with tennis to be able to imagine the independence of information concretely (i.e., receiving separate advice based on two players that were randomly put on a team), but understanding this independence might be more difficult in other settings, especially more unfamiliar ones. Moreover, participants in our studies also always received advice from two advisors. However, there are many instances in which people have to combine more than two forecasts. It is not clear yet whether people continue to be sensitive to differences in advice independence when the number of forecasts that they are combining increases beyond two. We suspect that advice independence could play a diminished role as the number of forecasts that are being combined increases because other factors (e.g., level of advisor agreement) move to the forefront of advice-takers' attention. It is also possible that sensitivity to advice independence starts to differ as a function of format when the number of forecasts increases. For example, combining forecasts from one format might be more cognitively demanding than combining forecasts from the other. If so, as the number of forecasts increases, the additional cognitive load from combining more forecasts could detract from people's ability to attend to advice independence (depending on the format).

Conclusions

In the present research, we demonstrated that people's likelihood estimates based on forecasts from multiple advisors can be sensitive to the level of independence between the advisors. Whereas past research has generally concluded that advice independence does not factor into the advice combination process to a large extent, we found strong evidence that people can adjust to the level of advice independence when the level of independence is clear and salient. We also found that people tend to combine verbal forecasts in a different manner than numeric forecasts, replicating recent work by Mislavsky and Gaertig (2022) and Teigen et al. (2023). Finally, these two main effects were not accompanied by the predicted interaction: The format in which advice is given did not appear to impact people's sensitivity to advice independence when combining forecasts from advisors.

References

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

Budescu, D. V., & Rantilla, A. K. (2000). Confidence in aggregation of expert opinions. *Acta Psychologica*, *104*(3), 371–398. https://doi.org/10.1016/S0001-6918(00)00037-8

Budescu, D. V., Rantilla, A. K., Yu, H. T., & Karelitz, T. M. (2003). The effects of asymmetry among advisors on the aggregation of their opinions. *Organizational Behavior and Human Decision Processes*, *90*(1), 178–194. https://doi.org/10.1016/S0749-5978(02)00516-2

Budescu, D. V., & Wallsten, T. S. (1995). Processing linguistic probabilities: General principles and empirical evidence. *Psychology of Learning and Motivation*, *32*, 275–318. https://doi.org/10.1016/S0079-7421(08)60313-8

Budescu, D. V., & Yu, H. T. (2006). To Bayes or not to Bayes? A comparison of two classes of models of information aggregation. *Decision Analysis*, *3*(3), 145–162. https://doi.org/10.1287/deca.1060.0074

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

Clemen, R. T. (1989). Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, *5*(4), 559–583. https://doi.org/10.1016/0169-2070(89)90012-5

Collins, P. J., & Hahn, U. (2018). Communicating and reasoning with verbal probability expressions. *Psychology of Learning and Motivation*, *69*, 67–105. https://doi.org/10.1016/bs.plm.2018.10.003

Collins, R. N., Mandel, D. R., & MacLeod, B. A. (2024). Verbal and numeric probabilities differentially shape decisions. *Thinking & Reasoning*, *30*(1), 235–257. https://doi.org/10.1080/13546783.2023.2220971

Collsiöö, A., Juslin, P., & Winman, A. (2023). Is numerical information always beneficial? Verbal and numerical cue-integration in additive and non-additive tasks. *Cognition*, *240*, Article 105584. https://doi.org/10.1016/j.cognition.2023.105584

Davidson, R. A., & Chrisman, H. H. (1993). Interlinguistic comparison of international accounting standards: The case of uncertainty expressions. *The International Journal of Accounting*, *28*(1), 1–16.

Davidson, R. A., & Chrisman, H. H. (1994). Translations of uncertainty expressions in Canadian accounting and auditing standards. *Journal of International Accounting, Auditing & Taxation*, *3*(2), 187–203. https://doi.org/10.1016/1061-9518(94)90016-7

Dhami, M. K., & Mandel, D. R. (2022). Communicating uncertainty using words and numbers. *Trends in Cognitive Sciences*, *26*(6), 514–526. https://doi.org/10.1016/j.tics.2022.03.002

Doupnik, T. S., & Richter, M. (2004). The impact of culture on the interpretation of "in context" verbal probability expressions. *Journal of International Accounting Research*, *3*(1), 1–20. https://doi.org/10.2308/jiar.2004.3.1.1

Galton, F. (1907). The ballot-box. *Nature*, *75*(1952), 509–510. https://doi.org/10.1038/075509f0

Hamm, R. M. (1991). Selection of verbal probabilities: A solution for some problems of verbal probability expression. *Organizational Behavior and Human Decision Processes*, *48*(2), 193–223. https://doi.org/10.1016/0749-5978(91)90012-I

Harris, A. J., Corner, A., Xu, J., & Du, X. (2013). Lost in translation? Interpretations of the probability phrases used by the Intergovernmental Panel on Climate Change in China and the UK. *Climatic Change*, *121*(2), 415–425. https://doi.org/10.1007/s10584-013-0975-1

Hogarth, R. M. (1978). A note on aggregating opinions. *Organizational Behavior & Human Performance*, *21*(1), 40–46. https://doi.org/10.1016/0030-5073(78)90037-5

Hohle, S. M., & Teigen, K. H. (2015). Forecasting forecasts: The trend effect. *Judgment and Decision Making*, *10*(5), 416–428. https://doi.org/10.1017/S1930297500005568

Hohle, S. M., & Teigen, K. H. (2018). When probabilities change: Perceptions and implications of trends in uncertain climate forecasts. *Journal of Risk Research*, *22*(5), 555–569. https://doi.org/10.1080/13669877.2018.1459801

Jenkins, S. C., Harris, A. J., & Lark, R. M. (2018). Understanding 'unlikely (20% likelihood)'or '20% likelihood (unlikely)'outcomes: The robustness of the extremity effect. *Journal of Behavioral Decision Making*, *31*(4), 572–586. https://doi.org/10.1002/bdm.2072

Juanchich, M., & Sirota, M. (2020). Do people really prefer verbal probabilities? *Psychological Research*, *84*(8), 2325–2338. https://doi.org/10.1007/s00426-019-01207-0

Juanchich, M., Teigen, K. H., & Gourdon, A. (2013). Top scores are possible, bottom scores are certain (and middle scores are not worth mentioning): A pragmatic view of verbal probabilities. *Judgment and Decision Making*, *8*(3), 345–364. https://doi.org/10.1017/S193029750000601X

Juanchich, M., Teigen, K. H., & Villejoubert, G. (2010). Is guilt 'likely'or 'not certain'?: Contrast with previous probabilities determines choice of

verbal terms. *Acta Psychologica*, *135*(3), 267–277. https://doi.org/10.1016/j.actpsy.2010.04.016

Kerr, N. L., & Tindale, R. S. (2011). Group-based forecasting? A social psychological analysis. *International Journal of Forecasting*, *27*(1), 14–40. https://doi.org/10.1016/j.ijforecast.2010.02.001

Larrick, R. P., & Soll, J. B. (2006). Intuitions about combining opinions: Misappreciation of the averaging principle. *Management Science*, *52*(1), 111–127. https://doi.org/10.1287/mnsc.1050.0459

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

Maines, L. A. (1996). An experimental examination of subjective forecast combination. *International Journal of Forecasting*, *12*(2), 223–233. https://doi.org/10.1016/0169-2070(95)00623-0

Makridakis, S. (1989). Why combining works? *International Journal of Forecasting*, *5*(4), 601–603. https://doi.org/10.1016/0169-2070(89)90017-4

Mandel, D. R., Dhami, M. K., Tran, S., & Irwin, D. (2021). Arithmetic computation with probability words and numbers. *Journal of Behavioral Decision Making*, *34*(4), 593–608. https://doi.org/10.1002/bdm.2232

Massey, C., & Wu, G. (2005). Detecting regime shifts: The causes of under- and overreaction. *Management Science*, *51*(6), 932–947. https://doi.org/10.1287/mnsc.1050.0386

Mellers, B., Ungar, L., Baron, J., Ramos, J., Gurcay, B., Fincher, K., Scott, S. E., Moore, D., Atanasov, P., Swift, S. A., Murray, T., Stone, E., & Tetlock, P. E. (2014). Psychological strategies for winning a geopolitical forecasting tournament. *Psychological Science*, *25*(5), 1106–1115. https://doi.org/10.1177/0956797614524255

Mislavsky, R., & Gaertig, C. (2022). Combining probability forecasts: 60% and 60% is 60%, but likely and likely is very likely. *Management Science*, *68*(1), 541–563. https://doi.org/10.1287/mnsc.2020.3902

Mordkoff, J. T. (2019). A simple method for removing bias from a popular measure of standardized effect size: Adjusted partial eta squared. *Advances in Methods and Practices in Psychological Science*, *2*(3), 228–232. https://doi.org/10.1177/2515245919855053

Reagan, R. T., Mosteller, F., & Youtz, C. (1989). Quantitative meanings of verbal probability expressions. *Journal of Applied Psychology*, *74*(3), 433–442. https://doi.org/10.1037/0021-9010.74.3.433

Schulz-Hardt, S., Wanzel, S. K., Rollwage, J., Treffenstädt, C., & Schultze, T. (2022). Do judges prefer advisors with dependent or independent errors? Investigating judges' advice selection and advice weighting. *Journal of Experimental Psychology: General*, *151*(7), 1636–1654. https://doi.org/10.1037/xge0001153

Soll, J. B. (1999). Intuitive theories of information: Beliefs about the value of redundancy. *Cognitive Psychology*, *38*(2), 317–346. https://doi.org/10.1006/cogp.1998.0699

Strueder, J. D., & Windschitl, P. D. (2024). *Combining forecasts from advisors*. https://researchbox.org/1748

Teigen, K. H., Juanchich, M., & Filkuková, P. (2014). Verbal probabilities: An alternative approach. *The Quarterly Journal of Experimental Psychology*, *67*(1), 124–146. https://doi.org/10.1080/17470218.2013.793731

Teigen, K. H. (1988). The language of uncertainty. *Acta Psychologica*, *68*(1–3), 27–38. https://doi.org/10.1016/0001-6918(88)90043-1

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

Teigen, K. H., & Brun, W. (1999). The directionality of verbal probability expressions: Effects on decisions, predictions, and probabilistic reasoning. *Organizational Behavior and Human Decision Processes*, *80*(2), 155–190. https://doi.org/10.1006/obhd.1999.2857

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

Teigen, K. H., Juanchich, M., & Løhre, E. (2023). Combining verbal forecasts: The role of directionality and the reinforcement effect. *Journal of Behavioral Decision Making*, *36*(2), Article e2298. https://doi.org/10.1002/bdm.2298

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

Wallsten, T. S., Budescu, D. V., Rapoport, A., Zwick, R., & Forsyth, B. (1986). Measuring the vague meanings of probability terms. *Journal of Experimental Psychology: General*, *115*(4), 348–365. https://doi.org/10.1037/0096-3445.115.4.348

Wallsten, T. S., Budescu, D. V., & Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science*, *39*(2), 176–190. https://doi.org/10.1287/mnsc.39.2.176

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

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

Winkler, R. L., Grushka-Cockayne, Y., Lichtendahl, K. C., Jr., & Jose, V. R. R. (2019). Probability forecasts and their combination: A research perspective. *Decision Analysis*, *16*(4), 239–260. https://doi.org/10.1287/deca.2019.0391

Winkler, R. L., & Poses, R. M. (1993). Evaluating and combining physicians' probabilities of survival in an intensive care unit. *Management Science*, *39*(12), 1526–1543. https://doi.org/10.1287/mnsc.39.12.1526

Received October 5, 2023
Revision received April 4, 2024
Accepted April 11, 2024 ■