Arithmetic Computation with Probability Words and Numbers

David R. Mandel[1], Mandeep K. Dhami[2], Serena Tran[1], and Daniel Irwin[3]

[1]Intelligence, Influence and Collaboration Section, Toronto Research Centre, Defence Research and Development Canada

[2]Department of Psychology, Middlesex University

[3]Department of National Defence

Correspondence: drmandel66@gmail.com

Acknowledgements: This research was supported by Canadian Safety and Security Program project CSSP-2018-TI-2394. We thank Robert Collins, Brenda Fraser, Tonya Hendriks, and Brooke Macleod for their research support. We also thank George Wright and three anonymous reviewers for their feedback on earlier drafts of this paper.

Word count (main body and notes, excluding refs, tables, bios and abstract): 9851

Abstract word count: 229

Abstract

Probability information is regularly communicated to experts who must fuse multiple estimates to support decision-making. Such information is often communicated verbally (e.g., "likely") rather than with precise numeric (point) values (e.g., ".75"), yet people are not taught to perform arithmetic on verbal probabilities. We hypothesized that the accuracy and logical coherence of averaging and multiplying probabilities will be poorer when individuals receive probability information in verbal rather than numerical point format. In four experiments (N = 213, 201, 26, and 343, respectively), we manipulated probability communication format between-subjects. Participants averaged and multiplied sets of four probabilities. Across experiments, arithmetic accuracy and coherence was significantly better with point than with verbal probabilities. These findings generalized between expert (intelligence analysts) and non-expert samples and when controlling for calculator use. Experiment 4 revealed an important qualification: whereas accuracy and coherence were better among participants presented with point probabilities than with verbal probabilities, *imprecise* numeric probability ranges (e.g., ".70 to .80") afforded no computational advantage over verbal probabilities. Experiment 4 also revealed that the advantage of the point over the verbal format is partially mediated by strategy use. Participants presented with point estimates are more likely to use mental computation than guesswork, and mental computation was found to be associated with better accuracy. Our findings suggest that where computation is important, probability information should be communicated to end users with precise numeric probabilities.

Keywords: verbal probability, numeric probability, arithmetic, accuracy, coherence, uncertainty communication

Experts are often called on to make probability judgments that support others' decision-making. For instance, physicians communicate the probability of treatment benefits and harms to patients (Wiles, Duffy, & Neill, 2019). Climate scientists estimate and communicate the probability of climate-change factors to policymakers and the public (Budescu, Broomell, & Por, 2009). And, intelligence analysts assess the probability of alternative futures to characterize uncertainty for policymakers and military decision-makers (Ho, Budescu, Dhami, & Mandel, 2015; Mandel & Barnes, 2018). In these and other areas (e.g., Morgan, 1998), experts typically assess and communicate probabilities with words such as "likely" rather than numeric quantifiers such as "75% chance". This is true even in stereotypically quantitative professions such as accounting (Kolesnika, Silska-Gembka, & Gierusz, 2019), and is consistent with the preference of communication senders who tend to favor the use of verbal over numeric probabilities (Erev & Cohen, 1990; Juanchich & Sirota, 2020; Olson & Budescu, 1997; Wallsten, Budescu, Zwick, & Kemp, 1993).

However, the fact that verbal probabilities often serve as inputs to others' judgments or decisions means that it is important to investigate how these probabilities are understood and used. In many cases, judgments and decisions that rely on earlier probability estimates require some form of arithmetic estimation or computation. For instance, a commander might receive multiple probability estimates from different intelligence sources or advisors and have to fuse them into an average estimate in order to decide on whether to undertake a high-risk operation, such as in the case of President Obama deciding on the military operation to capture or kill Osama bin Laden (Friedman & Zeckhauser, 2015). In such cases, the decision-maker must be able to estimate the average of the individual probability estimates received. As another example, a risk analyst might receive intelligence on the probabilities assigned to independent, necessary conditions that are judged to be jointly sufficient to yield a particular type of threat being monitored (e.g., a terrorist attack in a given location and timeframe). The analyst should ideally be able to multiply those values to determine the conjunctive probability of the threat. Conversely, the failure to correctly estimate the conjunctive probability of safety underlies many technological disasters (Perrow, 1984).

The requirement to perform arithmetic operations, such as averaging or multiplication, on verbal probabilities poses challenges because such probabilities have vague meanings that are poorly captured by precise numeric values. Although such vagueness can be represented by membership functions over the [0, 1] probability interval (Wallsten, Budescu, Rapoport, Zwick, & Forsyth, 1986; Zadeh, 1975), this does not offer a clear path to arithmetic computation in the real-world contexts described above. This is especially so because such representations involve detailed elicitations that are technically infeasible in many organizational contexts, such as national security intelligence. Furthermore, research has documented multiple violations of logical constraints on the integration of verbal probabilities (Budescu, Zwick, Wallsten, & Erev, 1990; Zwick, Budescu, & Wallsten, 1988). There is also large variability in the interpretation of verbal probabilities that are translated to numeric equivalents (e.g., Beyth-Marom, 1982; Budescu & Wallsten, 1990; Dhami & Wallsten, 2005; Lichtenstein & Newman, 1967). Such variability implies that even if verbal probabilities can be operated on arithmetically, the same probability information may be interpreted differently by different judges and even by the same judges across time, resulting in variability in the computed values as well. In collaborative and advisory decision-making situations, such variability translates into unreliability, which can obscure the

informational bases for sound judgment. For instance, Wiles et al. (2019) found that patients assigned significantly higher probability equivalents to verbal probability expressions used in the context of communicating about major peri-operative complications than did clinicians.

Research has also shown that participants rate verbal probabilities as conveying information about the specific *degree* of probability less clearly than numeric probabilities (Collins & Mandel, 2019). Probability information received that is unclear in terms of its degree may have deleterious consequences for receivers' subsequent computations. Perhaps this is why some studies have found numeric probabilities to yield more accurate and reliable judgments than verbal probabilities (Budescu, Weinberg, & Wallsten, 1988; Rapoport, Wallsten, Erev, & Cohen, 1990; but for studies showing comparable accuracy in diagnostic judgment using verbal and numeric probabilities, see Meder & Mayrhofer, 2017).

The Present Research

In the present research, we compared individuals' abilities to perform arithmetic operations (i.e., averaging and multiplication) on sets of probabilities that were either received verbally or numerically. This has not been the focus of earlier studies comparing probability formats on accuracy (e.g., Budescu et al., 1988; Rapoport et al., 1990). We chose these two operations because, as noted earlier, they are commonly required in a wide range of practical judgment and decision-making contexts.[1] To the best of our knowledge, no study has examined the effect of probability communication format on people's abilities to perform arithmetic calculations such as these. Children typically learn to perform arithmetic operations on numbers rather than on linguistic quantifiers, and this tendency continues in adulthood. Therefore, we expected that individuals would be less adequately prepared to arithmetically combine verbal compared to numeric probabilities. Specifically, we hypothesized that if individuals must compute averages or products from multiple probability estimates they would be less accurate when dealing with verbal probabilities than with numeric probabilities. We tested this *numeric superiority hypothesis* in four experiments.[2] In Experiments 1-3, we compared participants' ability to average and/or multiply verbal and precise numeric probabilities. In Experiment 4, we compared participants' ability to perform these computations with either verbal, precise numeric, or imprecise numeric-range probabilities. In Experiments 1, 2, and 4, we used large crowd-sourced samples, whereas in Experiment 3 we tested the numeric superiority hypothesis on a smaller sample of professional intelligence analysts, who are routinely required to work with probability estimates.

Furthermore, we examined whether there was a numeric superiority effect on the coherence of arithmetic responses. Wallsten, Budescu, and Zwick (1993) compared the additivity of probability judgments (i.e., the extent to which the sum of probabilities assigned to complementary events, x and $\neg x$, approach their logically constrained value of 1) and found little

[1]Although we anticipated that participants would be more accurate when averaging than when multiplying, this was not a focus of the research.
[2]Materials and data sets for all experiments are available at https://osf.io/5dwh8/.

difference between numeric and verbal probability formats. However, participants, on average, gave additive probability estimates in both formats. Therefore, their experiment was not sensitive enough to assess which format may be better in situations where incoherence is likely to flourish. In the present research, participants were judged to be coherent if they respected the following normative principles: For averaging, the average of a set of values cannot be greater than the highest value in the set and it cannot be less than the lowest set value:

$$\hat{p} \geq \leq \min(p_i) \cap \hat{p} \leq \max(p_i), \quad (1)$$

where $\hat{p}$ is the participant's response and p is the true response. For multiplication, the product of a set of probability values cannot be greater than the lowest set value:

$$\hat{p} \leq \min(p_i). \quad (2)$$

Violations of these coherence principles imply inaccuracy. However, inaccuracy does not necessarily imply such violations. For instance, judging the mean of .45 and .55 to be .53 is inaccurate but not in violation of the coherence criteria used here. As this example illustrates, the coherence violations examined in the present research not only imply inaccuracy, they imply an extreme degree of it. For instance, in the previous example, an incoherent participant would have to estimate a mean probability that was either less than .45 or greater than .55.

Coherence violations do not merely measure the magnitude of quantitative inaccuracy, they can also indicate that an individual lacks an appropriate mental representation of the task or schema for solving it (Mandel, 2008). Performing arithmetic operations on precise numeric values is something most individuals are taught to do at an early age. In contrast, even if individuals are periodically called on to arithmetically fuse verbal probability estimates to reach a decision, we hypothesize that their understanding of the process for doing so would be relatively aschematic, perhaps representing a full or partial breakdown in task construal (Clausner & Croft, 1999; Langacker, 1987). In other words, individuals may struggle or entirely fail to map verbal probabilities as inputs to a schema for performing arithmetic. If so, the asymmetry in schematicity for arithmetically processing verbal and numeric probabilities might be evident in the correlations between computational performance measures (i.e., accuracy and coherence), on the one hand, and individual differences in numeracy, on the other hand. Higher levels of numeracy (i.e., an individual's ability to perform basic mathematical operations required for statistical literacy in daily life) facilitate probability assessment and improve the interpretation of numerical data (Lipkus & Peters, 2009). We hypothesized that numeracy will correlate more strongly with accuracy and coherence among individuals presented with numeric probabilities than with those presented with verbal probability. We refer to this as the *differential schematicity hypothesis*.

Experiment 1

The principal aim of Experiment 1 was to test the numeric superiority hypothesis that participants will be more accurate and coherent in averaging and multiplying probabilities if the values they received were expressed with precise numeric probabilities rather than with verbal probabilities. We tested this hypothesis by manipulating probability format between subjects and

presenting participants with two sets of four probabilities. For each set, participants were asked to average and multiply the probabilities and we measured their accuracy and coherence.

In addition to numeracy, which was used to test the differential schematicity hypothesis, we also collected data on participants' verbal reasoning and actively open-minded thinking (AOT). Verbal reasoning skill, which measures abstract analogical reasoning abilities using language, correlates with IQ measures (Bilker, Wierzbicki, Brensinger, Gur, & Gur, 2014). AOT, which assesses an individual's willingness to evaluate evidence contrary to their beliefs as well as openness to alternative perspectives (Baron, Scott, Fincher, & Metz, 2015), is positively associated with accuracy in probabilistic judgment tasks (Haran, Ritov, & Mellers, 2013) and negatively associated with certain cognitive biases that violate coherence principles (Toplak, West, & Stanovich, 2017). We used these measures, along with numeracy, as statistical control variables in our primary analyses.

Method

Participants

Participants (58% male) between the ages of 18 and 60 ($M = 44.56$, $SD = 11.30$) were recruited using the online crowdsourcing service, Qualtrics Panels. They were required to have English as their first language and were sampled from Canada or the US. Participants were prohibited from completing the experiment on a smartphone. Participants who did not pass the instructional manipulation check, a one-item test designed to identify participants who are not attending carefully to instructions (Oppenheimer, Meyvis, & Davidenko, 2009), were also screened out of the experiment. The final sample was comprised of 213 participants.

Design

Experiment 1 used a 2 (Format: numeric, verbal) × 2 (Operation: averaging, multiplication) mixed factorial design with operation as a repeated measure. Participants in the verbal condition were given two sets of verbal probabilities. Participants in the numeric condition were given two sets of numeric probabilities that corresponded with the numeric interpretation of the verbal probabilities. Within each set, participants were asked to provide the arithmetic average and the product of the elements in a set.

The dependent variable for tests of accuracy was the standardized absolute error (SAE) between the participant's response, $\hat{p}$, and the true response, p:

$$\text{SAE} = |\hat{p} - p|/\max[p, (1 - p)].$$

In the analyses that follow, mean values of SAE are denoted as MSAE. We used MSAE rather than mean absolute error (MAE) to control for possible differences in the extremity of truth-values, p, due to the different bases for accessing these values (for similar use of standardized error scores, see Fan, Budescu, Mandel, & Himmelstein, 2019). Specifically, in the numeric condition, the correct response was based on the means and products of the numeric values in the

two sets. In the verbal condition, we used participants' own numeric equivalents of the verbal probabilities to compute the correct values.

The dependent variable for tests of coherence was the number of incoherent responses across the two sets and within arithmetic operation. Thus, the measure ranged from 0 (coherent on both Sets 1 and 2) to 2 (incoherent on both sets). We calculated this incoherence measure separately for responses to the averaging task and to the multiplication task using the two criteria described in Equations 1 and 2.

Procedure

Participants completed the experiment as part of three brief studies run online in randomized order. The other studies examined the effect of probability format on the perceived credibility of a hypothetical forecaster (which constitutes part of the data analyzed in Collins & Mandel, 2019) and the effect of probability format on the interpretation of probabilities in statements (which constitutes part of the data analyzed in Mandel & Irwin, 2020a).

The verbal condition of Experiment 1 was run first. Participants in this condition provided numeric-probability equivalents for the verbal terms in the two sets after completing the arithmetic task. We ran the verbal condition first because the median values of the numeric equivalents that participants provided, rounded to the nearest 5% interval, were used as the numeric-probability set values in the numeric condition that was subsequently run. The Qualtrics Panels sampling system was configured such that participants in the verbal condition were excluded from participating in the (subsequently run) numeric condition.

Participants performed arithmetic operations on two sets of four probabilities in succession. In the verbal condition, Sets 1 and 2 comprised the terms {*highly likely*, *unlikely*, *almost certain*, and *likely*} and {*unlikely*, *highly likely*, *remote chance*, and *highly unlikely*}, respectively. In the numeric condition, Sets 1 and 2 comprised the corresponding values {0.75, 0.25, 0.80, and 0.65} and {0.25, 0.75, 0.25, and 0.20}, respectively. For each set, participants were first asked what the arithmetic average (i.e., the mean value) of the values or terms were, followed by what their product was. Responses were provided on a slider scale ranging from 0 to 1 with .01 increments. The initial position of the slider was set to 0, and the value of the scale was visible as the slider was moved along the scale.

As mentioned earlier, participants in the verbal condition were also asked to indicate the best numeric-probability equivalent for each verbal probability in Set 1 using the same slider scale. Hereafter, for the sake of simplicity, we refer to the process of judging the equivalents of probabilities in one format to an alternative format as *translation*. The translation process was then repeated for Set 2.

As noted earlier, these translation values were used to establish the corresponding set values in the numeric condition. Note that Sets 1 and 2 in the verbal condition had two probability terms in common (namely, *highly likely* and *unlikely*). The mean absolute distance (MAD) between the

best numeric-probability equivalents of these terms (with MAD further averaged across the two terms) served as a measure of intra-individual reliability.[3] In addition to serving as measure of reliability, our calculation of MAD permitted a test of whether participants' accuracy in the verbal condition might have been influenced by the reliability of their verbal-to-numeric translations.

After the experimental tasks, participants completed measures of numeracy, verbal reasoning skill, and AOT. Numeracy was measured using eight questions from Lipkus, Samsa, and Rimer's (2001) numeracy scale and two questions from the Berlin Numeracy Test (Cokely, Galesic, Schulz, Ghazal, & Garcia-Retamero, 2012). The verbal-skills test comprised eight verbal-analogy questions drawn from the 29-item Penn Verbal Reasoning Test (PVRT) (Bilker et al., 2014). Finally, we used the eight-item AOT scale used in Baron et al. (2015). These measures were used as covariates in analyses of covariance and used to test the hypothesis that accuracy and coherence are more strongly related to these variables when participants are presented with numeric rather than verbal probabilities.

Results

Accuracy

We examined MSAE in a two-way (Format × Operation) mixed analysis of covariance (ANCOVA) with numeracy, PVRT, and AOT as covariates and operation as a repeated measure. As Table 1 shows, the main effects of format and operation were significant and the interaction between these factors was not. Figure 1 shows hybrid box-and-whisker and error plots for each condition in the two-way model. The box-and-whisker plots use sample data and the error plots use estimated marginal means from the ANCOVA model (this also applies to subsequent figures). Figure 1 shows a numeric superiority effect for both tasks; that is, MSAE was lower in the numeric condition than in the verbal condition. Also, MSAE was lower for the averaging task than for the multiplication task.

Next, we tested the differential schematicity hypothesis. Given that MSAE for averaging and multiplication tasks were highly correlated ($r[211] = .46$, $p < .001$), we averaged these measures. Numeracy was significantly correlated with MSAE in the numeric condition ($r[103] = –.41$, $p < .001$) but numeracy and MSAE were not significantly correlated in the verbal condition ($r[106] = –.13$, $p = .197$).[4] In support of the differential schematicity hypothesis, the difference between these correlations was significant, $z = 2.19$, one-tailed $p = .014$.

---

[3]In Experiments 1-4, there were also questions that asked participants to re-compute arithmetic results with the translated values. However, in hindsight, we did not judge these questions to have sufficient probative value and we do not report on them.

[4]This difference in correlational strength across format was also evident if the averaging and multiplication tasks are analyzed separately. In the numeric condition, rs = –.45 and –.32 and in the verbal condition, rs = –.10 and –.11 for averaging and multiplication tasks, respectively.

Finally, recall that to gauge intra-individual unreliability in the interpretation of verbal probabilities, we calculated MAD between the two terms (i.e., *unlikely* and *highly likely*) that appeared in both Sets 1 and 2. The mean of the two MAD measures was significantly greater than zero and yielded a large effect size, $M = 0.14$, $SD = 0.17$, one-sample $t(107) = 8.52$, $p < .001$, Cohen's $d = 0.82$. To examine whether inaccuracy among participants in the verbal condition was related to their reliability in mapping verbal probabilities onto numeric probabilities, we examined the correlation between MAD and MSAE. The correlation was small and nonsignificant, $r(106) = .13$, $p = .19$.

Coherence

We examined incoherence scores in a two-way (Format × Operation) ANCOVA with numeracy, PVRT, and AOT as covariates. As Table 2 shows, the main effects of format and operation and the interaction effect were significant. Demonstrating a numeric superiority effect, responses were less incoherent in the numeric condition ($M = 0.62$, $SE = 0.047$) than in the verbal condition ($M = 0.99$, $SE = 0.046$). Also, consistent with the analysis of accuracy, incoherence was lower for averaging ($M = 0.31$, $SE = 0.037$) than for multiplication ($M = 1.30$, $SE = 0.054$). The simple effect of format was significant at both levels of operation, but the effect was stronger for multiplication than for averaging. For multiplication, incoherence in the numeric condition ($M = 1.03$, $SE = 0.079$) was significantly less than in the verbal condition ($M = 1.56$, $SE = 0.077$), $F(1, 208) = 22.07$, $p < .001$, $\eta_p^2 = .096$. Likewise, for averaging, incoherence in the numeric condition ($M = 0.21$, $SE = 0.054$) was significantly less than in the verbal condition ($M = 0.42$, $SE = 0.053$), $F(1, 208) = 7.05$, $p < .009$, $\eta_p^2 = .033$.

Incoherence scores for averaging and multiplication tasks showed a significant but small correlation ($r[211] = .14$, $p = .037$), in contrast to the large correlation observed in the corresponding accuracy analysis. To be prudent (and under the assumption that averaging should not be undertaken for correlations less than a medium effect of $r = .3$), we examined these scores separately for averaging and multiplication tasks in order to test the differential schematicity hypothesis. For the averaging task, numeracy was significantly correlated with coherence in the numeric condition ($r[103] = –.39$, $p < .001$) and in the verbal condition ($r[106] = –.33$, $p = .001$). Contrary to the differential schematicity hypothesis, the difference between these correlations was not statistically significant, $z = 0.50$, one-tailed $p = .31$. For the multiplication task, numeracy was significantly correlated with coherence in the numeric condition ($r[103] = –.42$, $p < .001$) but not in the verbal condition ($r[106] = –.09$, $p = .35$). In support of the differential schematicity hypothesis, there was a significant difference between these correlations, $z = 2.57$, one-tailed $p = .005$.

Discussion

The findings of Experiment 1 support the numeric superiority hypothesis. Participants were more accurate at averaging and multiplying probabilities when they were presented numerically as point estimates rather than verbally. Experiment 1 further demonstrated that participants who were required to average and multiply probabilities were less likely to do so coherently when they received verbal probabilities than when they received numeric probabilities. There was also partial support for the differential schematicity hypothesis. As the hypothesis predicts, accuracy

was significantly more strongly related to numeracy when the information received was in the numeric rather than verbal format. Similarly, numeracy was significantly more strongly related to coherence in the numeric than verbal format. However, this difference was not statistically significant for the averaging task. Given that participants found averaging easier than multiplication, it may have provided a weak test of the differential schematicity hypothesis. Finally, we found that participants presented with verbal probabilities were highly unreliable in their mapping of probability terms to numeric equivalents, although such unreliability did not correlate with accuracy.

While the findings of Experiment 1 are informative, the experiment is also limited by the fact that the multiplication task always followed the averaging task. Therefore, it is unclear whether effects of operation are due to the arithmetic operation per se or to carryover effects. Accordingly, in Experiment 2, we manipulated the arithmetic operations in a between-subjects design and examined the replicability of the findings in Experiment 1.

Experiment 2

As noted, the aim of Experiment 2 was to replicate the findings that arithmetic accuracy and coherence were greater when participants receive precise numeric probabilities rather than verbal probabilities as inputs. Unlike Experiment 1, which presented averaging and multiplication tasks to participants in a fixed order, Experiment 2 independently manipulated both format and operation in a between-subjects design.

Method

Participants

Experiment 2 was administered to participants (47.8% male) between the ages of 18 and 60 (M = 40.64, SD = 10.84) using Qualtrics Panels. We used the same inclusion and exclusion criteria as in Experiment 1. The final sample was comprised of 201 participants.

Design

Experiment 2 used a 2 (Format: numeric, verbal) × 2 (Operation: averaging, multiplication) between-subjects design. The dependent variables were the same as in Experiment 1.

Procedure

The procedure followed that of Experiment 1 except for three changes. First, because we used the same sets as in Experiment 1, there was no need to run the verbal condition first; that is, we used the same equivalence values established in Experiment 1 and randomly assigned participants to conditions. Second, participants gave only one type of arithmetic response

because operation was manipulated between-subjects. Third, the wording of the arithmetic questions was simplified (for precise changes, see https://osf.io/5dwh8/).[5]

Results

Accuracy

We examined MSAE in a two-way (Format × Operation) factorial ANCOVA controlling for numeracy, PVRT, and AOT as in Experiment 1. As Table 3 shows, the main effects of format and operation were significant, as was the interaction effect (see Figure 2). Although it is evident that the numeric superiority effect was stronger in the multiplication condition than in the averaging condition, simple-effect tests showed that the effect of format was significant in each operation condition: for averaging, $F(1, 100) = 11.14$, $p = .001$, $\eta_p^2 = .10$; for multiplication, $F(1, 91) = 20.12$, $p < .001$, $\eta_p^2 = .18$.

Next, we examined support for the differential schematicity hypothesis. Numeracy was significantly correlated with MSAE in the numeric condition ($r[91] = –.45$, $p < .001$) but not in the verbal condition ($r[106] = –.04$, $p = .65$). In support of the differential schematicity hypothesis and consistent with the findings of Experiment 1, these correlations were significantly different, $z = 3.10$, one-tailed $p = .001$.

Finally, we calculated MAD between the two pairs of common verbal probability terms that appeared in Sets 1 and 2 in the verbal condition. As in Experiment 1, MAD was significantly greater than zero, $M = 0.12$, $SD = 0.17$, one-sample $t(107) = 7.58$, $p < .001$, Cohen's $d = 0.71$. Also consistent with the findings of Experiment 1, the correlation between MAD and MSAE was small and non-significant, $r(106) = .12$, $p = .23$.

Coherence

We examined incoherence scores in a two-way (Format × Operation) factorial ANCOVA with numeracy, PVRT, and AOT as covariates. As Table 4 shows, the main effects of format and operation and the interaction effect were significant. Demonstrating a numeric superiority effect, responses were less incoherent in the numeric condition ($M = 0.50$, $SE = 0.065$) than in the verbal condition ($M = 1.17$, $SE = 0.060$). Also, consistent with our accuracy analysis and the results of Experiment 1, incoherence was lower for averaging ($M = 0.36$, $SE = 0.061$) than for multiplication ($M = 1.31$, $SE = 0.064$). The simple effect of format was significant at both levels of operation, but the effect was stronger for multiplication than for averaging. For multiplication, incoherence in the numeric condition ($M = 0.84$, $SE = 0.102$) was significantly less than in the verbal condition ($M = 1.78$, $SE = 0.088$), $F(1, 91) = 48.11$, $p < .001$, $\eta_p^2 = .346$. Likewise, for averaging, incoherence in the numeric condition ($M = 0.21$, $SE = 0.054$) was significantly less than in the verbal condition ($M = 0.42$, $SE = 0.053$), $F(1, 100) = 15.59$, $p < .001$, $\eta_p^2 = .135$.

---

[5]Responses to the arithmetic questions that followed the "translation" task were also altered, but as noted earlier, we do not analyze responses to those questions.

In support of the differential schematicity hypothesis, numeracy was significantly correlated with incoherence in the numeric condition ($r[91] = –.42$, $p < .001$) but numeracy and incoherence were not significantly correlated in the verbal condition ($r[106] = –.14$, $p = .161$). The difference between these correlations was statistically significant, $z = 2.14$, one-tailed $p = .016$.

Discussion

Experiment 2 replicated the key findings of Experiment 1. Specifically, participants were more accurate and coherent after receiving precise numeric probabilities than following receipt of verbal probabilities. These numeric superiority effects were evident for both averaging and multiplication tasks. Moreover, replicating Experiment 1, the accuracy of participants presented with verbal probabilities was not significantly related to the reliability with which participants assigned numeric probability equivalents to the two common verbal probability terms in Sets 1 and 2, although substantial unreliability was once again observed. Finally, the differential schematicity hypothesis was strongly supported in Experiment 2. As the hypothesis predicts, both accuracy and coherence were significantly more strongly related to numeracy when probabilities were presented numerically rather than verbally.

Experiment 3

Whereas Experiments 1 and 2 found support for the numeric superiority hypothesis using participants recruited online, Experiment 3 examined whether comparable effects are replicable among a sample of practicing intelligence analysts. In addition to the shift from a non-expert to expert sample, we also explicitly instructed participants not to use a calculator, which had not been explicitly requested in Experiments 1 and 2 (Experiment 4 deals with this issue further). We further asked participants to indicate whether they reached their answer by mental calculation or rough estimation. We hypothesized that mental calculation would be a more accurate strategy than rough estimation, especially because there were no time constraints placed on participants to complete the task. Moreover, we hypothesized that mental calculation would be more likely to be employed with the numeric format than with the verbal format, where, as we have noted, there is a lower likelihood of accessing an appropriate schema for performing the arithmetic operations.

Method

Participants

Experiment 3 was administered to 21 Canadian intelligence analysts during regular course time at the Canadian Forces School for Military Intelligence (CFSMI) at Canadian Forces Base Kingston in Kingston, Ontario, Canada. Five additional Canadian intelligence analysts participated remotely using a Qualtrics survey link distributed by their manager. Participants were informed that their participation was voluntary and that they would not be remunerated for their time. None of the CFSMI attendees refused to participate. The final sample ($N = 26$) was 92.3% male and aged 25 to 50 ($M = 36.35$, $SD = 6.72$).

Design

Like Experiment 2, Experiment 3 examined probability format (numeric, verbal) as a between-subjects factor. However, given the substantially lower power of this experiment compared to Experiment 2, only one arithmetic operation (averaging) was examined. The dependent variables were the same as in Experiment 2.

Procedure

The procedure followed that of Experiment 2 except for two changes. First, in an introductory paragraph modified from Experiment 2, participants received explicit instruction at the beginning of the experiment not to use a calculator. Second, after completing the experimental task, participants were asked (1) whether they arrived at their answers through mental calculation or rough estimation, and (2) whether or not they used a calculator for any of the questions. For precise changes, see https://osf.io/5dwh8/.

Participants completed the experiment as part of three brief studies run in randomized order. The other studies examined the effect of probability format on the perception of implicit recommendations from a hypothetical forecaster and on the interpretation of confidence statements in intelligence assessments. Following the core experimental tasks, all participants completed individual difference tests measuring numeracy[6] and AOT (PVRT was not included due to constraints on the overall time available for testing).

Results

Accuracy

MSAE was analyzed in a one-way (Format) ANCOVA with numeracy and AOT as covariates. Supporting the numeric superiority hypothesis, intelligence analysts who were asked to compute averages from precise numeric probabilities (M = 0.076 [0.033, 0.120]) were significantly more accurate than analysts who were asked to do so from verbal probabilities (M = 0.173 [.129, 216]), $F(1, 22) = 10.41$, $MSE = 0.06$, $p = .004$, $\eta_p^2 = .32$.

Next, we examined support for the differential schematicity hypothesis. The correlation between numeracy and MSAE was comparable in the numeric condition ($r[11] = -.28$, $p = .36$) and the verbal condition ($r[106] = -.27$, $p = .38$), $z = 0.02$, one-tailed $p = .49$. Therefore, we did not find support for this hypothesis in the analysis of accuracy scores.

A novel feature of Experiment 3 was the inclusion of a question asking participants whether they used mental calculation or rough estimation to answer the averaging question. Sixteen (61.5%) participants reported using mental calculation, whereas 10 (38.5%) reported using rough

[6]Where the numeracy test used in Experiments 1 and 2 elicited a combination of multiple choice and text-box inputs, here we used a purely multiple-choice version for ease of scoring. For both versions of the numeracy test, see https://osf.io/5dwh8/.

estimation. A greater proportion of participants in the numeric condition (53.8%) reported using mental calculation than in the verbal condition (23.1%). Although a test of non-independence did not reach a conventional significance level, there was a medium effect size detected, $\chi^2(1, N = 26) = 2.60$, $p = .107$, $\phi = .32$. Moreover, MSAE was marginally lower among participants who reported using mental calculation ($M = 0.082$, $SD = 0.085$) than among participants who reported using rough estimation ($M = 0.151$, $SD = 0.086$), $t(24) = -1.99$, $p = .059$, Hedges' $g = -0.81$.

As in the previous experiments, MAD in the verbal condition was significantly greater than zero ($M = 0.012$, $SD = 0.020$, $t[12] = 2.21$, $p = .047$, Cohen's $d = 0.60$), indicating unreliability in the interpretation of verbal probabilities. Consistent with Experiments 1 and 2, MAD was not significantly correlated with MSAE, $r(11) = -.05$, $p = .87$. Thus, as in the previous experiments, accuracy in the verbal condition does not appear to depend on the reliability of participants' mapping of verbal to numeric probabilities.

Coherence

All of the participants in Experiment 3 provided coherent averages in either the numeric or verbal condition. Therefore, tests of the differential schematicity hypothesis such as those conducted in Experiments 1 and 2 could not be performed.

Discussion

Using an expert sample of intelligence analysts, Experiment 3 generalized the key finding of Experiments 1 and 2; namely, that arithmetic computation (i.e., averaging) involving probabilities is more accurate if the probability information is received as precise numeric probabilities rather than as verbal probabilities. However, the correlation between accuracy and numeracy was virtually identical in the two format conditions. This apparently contradicts the differential schematicity hypothesis. However, the difference in results may be explained in terms of the differential skill level of the samples. Accuracy on the averaging task was better in the expert sample of Experiment 3 than in the crowd-sourced samples of Experiments 1 and 2. Perhaps even more important, in the earlier experiments the correlational analysis included data from participants who completed the more difficult multiplication task, whereas in Experiment 3 participants were not asked to compute products. A high degree of mean accuracy suggests that few participants in Experiment 3 were at a loss for recruiting a relevant schema for the task. This is reaffirmed by the fact that no participant was incoherent in either the numeric or verbal condition. Therefore, Experiment 3 may be incapable of providing an adequately sensitive test of the differential schematicity hypothesis.

Experiment 3 nevertheless provided tentative support, based on self-reported strategy use, for the hypothesis that mental calculation (as opposed to guesswork) is more likely to be used if the probabilities received are in numeric rather than verbal format (where the effect size was medium). Also, mental calculation rather than guessing was found to be associated with greater accuracy (where the effect size was large). However, due to the low statistical power associated with Experiment 3's small sample size, these effects fell short of a conventional significance level. We replicated these hypothesis tests in Experiment 4 using a much larger sample.

Experiment 4

As stated above, Experiment 4 was administered online to replicate the key results of Experiment 3. To expand on previous findings, we also examined participants' ability to perform arithmetic operations with imprecise numeric probabilities (i.e., ranges). The inclusion of an imprecise numeric condition permitted an examination of whether the numeric superiority effect reflects the benefits of precision more than quantification. Given that numeric ranges are unambiguous but imprecise, we hypothesized that computational accuracy and coherence would be greatest with the precise numeric format, followed by the imprecise numeric format, and then with the verbal format. As in Experiment 3, we explicitly instructed participants not to use a calculator and elicited their computational strategy following the core experimental tasks. We expected to replicate the findings of Experiment 3 indicating that mental calculation is more likely to be used by participants operating on numeric probabilities than those operating on verbal probabilities, and that mental calculation is more strongly associated with arithmetic accuracy than rough guessing. We also expected to find support for the differential schematicity hypothesis. As with the numeric superiority hypothesis, we expected results obtained in the numeric range condition to fall between those obtained in the precise and verbal conditions. Thus, we expected the strongest correlation with numeracy with the precise format and the weakest correlation with numeracy with the verbal format. As before, we aimed to test this hypothesis on both accuracy and coherence measures.

Method

Participants

Experiment 4 was administered to participants (50.8% male) between the ages of 18 and 60 (M = 44.20, SD = 10.77) using Qualtrics Panels. We used the same inclusion and exclusion criteria as in Experiments 1 and 2. In spite of our instructions, 31 (8.3%) participants reported using a calculator during the core experimental tasks. These cases were removed and the final sample included 343 participants.

Design

Experiment 4 used a 3 (Format: precise numeric [henceforth, point], imprecise numeric [henceforth, range], verbal) × 2 (Operation: averaging, multiplication) between-subjects design. The dependent variables were the same as in Experiment 3. In the range condition, accuracy and coherence were scored as in the point condition, given that ranges can be converted into point estimates corresponding to the midpoint of the range with margins of error corresponding to half the range (Moore, Kearfott, & Cloud, 2009).[7]

[7]We regard this measure as a suitably conservative test of performance in the range condition. An alternative, such as calculating the best score possible given the range of values would undermine the comparability of performance across conditions given that no such allowances are made in the point and verbal conditions.

Procedure

The procedure followed that of Experiment 3, but the sample was large enough for us to again examine both averaging and multiplication. Participants in the point and verbal conditions were shown the same sets used in Experiments 1-3. In the range condition, Sets 1 and 2 comprised the corresponding values {0.70 to 0.80, 0.20 to 0.30, 0.75 to 0.85, 0.60 to 0.70} and {0.20 to 0.30, 0.70 to 0.80, 0.20 to 0.30, 0.15 to 0.25}, respectively. These values were established by taking the corresponding values in the point condition and adding a 5% margin of error above and below the value. Following the core experimental tasks, participants completed measures of numeracy, PVRT, and AOT. The numeracy measures had multiple-choice options as in Experiment, whereas PVRT and AOT were the same as in Experiments 1 and 2.

Results

Accuracy

We examined MSAE in a two-way (Format × Operation) factorial ANCOVA controlling for numeracy, PVRT, and AOT. As Table 5 shows, the main effects of format and operation were significant, but the two-way interaction was not significant. Figure 3 illustrates that the main effect of operation was due to the lower error in the averaging condition than in the multiplication condition. Pairwise testing using Fisher's Least Significant Difference test showed that error was significantly lower in the point condition than in the verbal ($p = .002$) and range ($p = .015$) conditions. The latter two conditions did not differ significantly ($p > .52$).

In support of the differential schematicity hypothesis, numeracy was significantly correlated with MSAE in the point condition ($r[107] = –.41$, $p < .001$) and the range condition ($r[112] = –.29$, $p = .002$) but these variables were not significantly correlated in the verbal condition ($r[118] = –.01$, $p = .89$). Given the virtually nil correlation in the verbal condition, it is evident that both the point and range conditions yield significantly stronger correlations, although the correlations in the point and range condition did not differ significantly from each other, $z = 1.01$, one-tailed $p = .16$.

As shown in Table 6, a majority of participants reported using rough estimation to solve the arithmetic task. Table 6 also shows the percentage of participants using each strategy as a function of format. To examine whether strategy use varied systematically by format, we treated verbal, range, and point conditions as an ordered set ranging from qualitative-imprecise to quantitative-precise, respectively. We hypothesized that mental calculation use would increase along this scale. Supporting this hypothesis, a Somers' D test of ordinal association treating format as the independent variable was equal to –.096, $SE = .043$, $t = –2.23$, $p = .025$. Consistent with Experiment 3, MSAE was significantly lower among participants who reported using mental calculation ($M = 0.253$, $SD = 0.247$) than among those who reported using rough estimation ($M = 0.398$, $SD = 0.273$), $t(341) = –5.02$, $p < .001$, Hedges' $g = –0.55$.

Taken together, the preceding findings suggest that the accuracy advantage of point numeric probabilities over verbal probabilities observed in Experiments 1-4 may be mediated by strategy

use. That is, receivers of numeric point probabilities may favour calculation over guesswork more than receivers of verbal probabilities, and mental calculation, in turn, is associated with better accuracy than guesswork. To test this mediation hypothesis, we excluded the range condition and treated format as a predictor variable, strategy as a mediator variable, and MSAE as the dependent variable. Figure 5 shows the results, which suggest that strategy use partially mediates the relation between format and accuracy. This is further supported by the result of a Sobel test, which confirms that the attenuation of the predictor's influence is statistically significant when the mediator is included in the model, $z = -2.03$, $p = .042$.

Finally, as in Experiments 1-3, MAD in the verbal condition was significantly greater than zero ($M = 0.128$, $SD = 0.165$, $t[119] = 8.49$, $p < .001$, Cohen's $d = 0.78$), indicating unreliability in the interpretation of verbal probabilities. As in the prior experiments, MAD calculated in the verbal condition was also not significantly correlated with MSAE, $r(118) = .01$, $p = .90$.

Coherence

We examined incoherence scores in a two-way (Format × Operation) factorial ANCOVA with numeracy, PVRT, and AOT as covariates. As Table 7 shows, the main effects of format and operation and the interaction effect were significant. As in Experiments 1 and 2, incoherence was more pronounced for the multiplication task than for the averaging task. Figure 4 plots the interaction effect. Simple-effect tests (once again controlling for numeracy, PVRT, and AOT) show that, for averaging, incoherence was significantly greater in the range condition than in the point and verbal conditions, $p = .002$ and .047, respectively; $F(2, 172) = 5.08$, $p = .007$, $\eta_p^2 = .056$. The point and verbal conditions did not differ significantly. For multiplication, incoherence was significantly greater in the verbal condition than in the point condition ($p < .001$), and neither condition significantly differed from the range condition; $F(2, 159) = 7.34$, $p = .001$, $\eta_p^2 = .085$.

Next, we tested support for the differential schematicity hypothesis. The correlations between numeracy and incoherence followed a similar pattern as observed with the comparable accuracy analysis. The correlations in the point, range, and verbal conditions were –.31 ($df = 107$, $p = .001$), –.16 ($df = 112$, $p = .084$), and –.05 ($df = 118$, $p = .611$), respectively. Supporting the differential schematicity hypothesis, the correlation in the point condition was significantly greater than in the verbal condition ($z = 2.02$, one-tailed $p = .022$). The range condition was not significantly different from either the point condition ($z = 1.17$, one-tailed $p = .121$) or the verbal condition ($z = 0.84$, one-tailed $p = .20$).

Finally, we examined whether coherence differed as a function of reported strategy use. Consistent with the results for accuracy, mean incoherence was significantly greater among the subsample who reported using rough estimation ($M = 1.04$, $SD = 0.94$) than among the subsample who reported using mental calculation ($M = 0.61$, $SD = 0.88$), $t(341) = 4.31$, $p < .001$, Hedges' $g = 0.47$.

Discussion

Experiment 4 replicated key findings of our earlier experiments. Specifically, we found that accuracy was better with the point numeric format than with the verbal format and participants were more coherent with the point format than with the verbal format when computing products, where the greatest level of incoherence was evident. This replication is important because unlike Experiments 1 and 2, in Experiment 4, participants were explicitly instructed not to use calculators and those who reported that they did were excluded from the analyses. In this regard, the findings of Experiment 4 reinforce those of Experiment 3, which also prohibited participants from using calculators, but which relied on a small sample and did not examine multiplication.

Experiment 4 also replicated the findings of Experiment 3 showing that use of a mental calculation strategy as opposed to rough estimation is more likely to be adopted when presented with point numeric probabilities rather than verbal probabilities. Interestingly, we found that the range format fell between the numeric and verbal formats in terms of the proportion of participants reporting to use mental calculation. Moreover, we replicated the finding of Experiment 3 that self-reported mental-calculation users were more accurate than those who reported using rough estimation, and this effect also was generalized to coherence violations. In both cases, the observed effect sizes were medium by conventional standards. These results further enabled us to test a mediator model and we found that strategy use partially mediated the effect of format (i.e., point numeric vs. verbal) on accuracy.

Experiment 4 also replicated support for the differential schematicity hypothesis, finding that the correlation between numeracy and accuracy was significantly smaller with the verbal format than with the precise and range formats. Similarly, the correlation between numeracy and coherence was significantly smaller with the verbal format than with the precise format.

Finally, Experiment 4 revealed an important qualification to the numeric superiority hypothesis: whereas accuracy and coherence tended to be better among participants presented with point probabilities than with verbal probabilities, participants presented with numeric ranges showed no appreciable computational advantage over participants presented with verbal probabilities. In fact, on at least one performance criterion i.e., coherence on the averaging task, participants presented with numeric ranges were significantly less coherent than participants presented with either point numeric or verbal probabilities. In the General Discussion, we examine the implications of this finding for the numeric superiority hypothesis.

General Discussion

As noted earlier, decision-makers rely on experts' probability judgments across a wide range of situations. Often decision-makers have to fuse multiple judgments in order to support their decision-making and planning objectives. Sometimes it is useful, if not necessary, to combine multiple probability estimates into an average or a product, such as when estimates from multiple advisors require aggregating or when threat probabilities can be estimated from the conjunctive probability of their necessary and jointly sufficient preconditions. We investigated how accurately and coherently individuals could compute such results from probabilities presented verbally or numerically, given that probability judgments may be communicated verbally in many consequential domains.

Across the four experiments we observed a high degree of consistency in the results. Invariably, participants presented with precise numeric probabilities were more accurate in their arithmetic computations than participants presented with verbal probabilities. With the exception of Experiment 3 in which all participants (intelligence analysts) responded coherently, participants presented with precise numeric probabilities also tended to exhibit greater coherence than participants presented with verbal probabilities. These differences cannot be explained by differential rates of calculator use or by success in using calculators because Experiments 3 and 4, which prohibited calculator use and excluded self-reported calculator users, yielded findings comparable to Experiments 1 and 2, which did not instruct participants to avoid using calculators. Nor does the numeric superiority effect appear to be due to sample characteristics given that the effect was evident not only in multiple crowd-sourced samples but also in a sample of professional intelligence analysts.

The results of Experiments 3 and 4, in particular, shed light on why point probabilities supported computation more effectively. First, we observed in both of these experiments that participants who received point numeric probabilities were more likely than participants who received verbal probabilities to report using a mental calculation strategy rather than relying on a rough estimate. We also showed that participants who used mental calculation had better accuracy. Furthermore, in Experiment 4, which had sufficient statistical power to test a mediation model, we confirmed that the effect of format on accuracy was significantly, albeit partially, mediated by strategy use. Therefore, it appears that part of the causal basis for the numeric superiority effect is that receiving point numeric probabilities as opposed to verbal probabilities makes it more likely that those who must compute with the probabilities will use an explicit computational approach as opposed to an implicit one consistent with guesswork, and the explicit approach tends to yield better accuracy.

The findings of Experiment 4 also shed light on the generalizability of the numeric superiority effect. The fact that participants who received numeric probability ranges were not significantly more accurate than participants who received verbal probabilities indicates that quantification of probability information *per se* may be less important than whether such information is expressed in precise or imprecise terms. The findings of Experiment 4 suggest that when communicated numeric probabilities are imprecise, people are less inclined to use calculation and more inclined to use guesswork to estimate arithmetic values than when the communicated probabilities are precise. In hindsight, we find this result unsurprising because, as with verbal probabilities, few people have experience computing arithmetic operations on ranges and doing so is unlikely to be part of one's formal education. Moreover, the response mode for the arithmetic tasks, which called for a point value to be selected, is incongruent with processing ranges defined by lower- and upper-bound quantities. Therefore, it would be useful in future research to examine the accuracy of computations with imprecise numeric probabilities when participants are required to provide lower and upper bounds on their best estimates. Perhaps the bounds would be more easily calculable than the best estimates given that they are congruent with the information provided in the range condition. Another option would be to require participants in each of the three format conditions to provide upper and lower bounds as well as a best estimate. Participants presented with the verbal format could also be asked to provide their upper and lower bounds when giving their numeric probability equivalents for the relevant probability terms.

Our findings suggest that people have difficulty thinking about verbal probabilities as inputs for arithmetic computations. Not only was accuracy and coherence impeded when participants received verbal probabilities, there was also substantial support for the differential schematicity hypothesis in the present research. In the experiments in which the more difficult multiplication task was administered, accuracy and coherence were each significantly more strongly correlated with numeracy in the point numeric condition than in the verbal condition. For the easier averaging task, the results were less consistent, but where there were significant differences in correlation strength detected (i.e., in Experiments 1, 2, and 4 for accuracy, and Experiments 2 and 4 for coherence), the differences were in the direction predicted by the differential schematicity hypothesis. The relations observed in the point numeric condition are as one might expect: more numerate individuals are less likely to violate coherence principles (e.g., Stanovich & West, 2000). The attenuation or near-elimination of these relations in the verbal condition across three experiments is therefore noteworthy as it suggests that at least for many individuals, regardless of their numeracy, the task of arithmetically computing with verbal probabilities is sufficiently difficult, and this is likely because they lack an adequate schema for computing with verbal probabilities. This might be why participants in the verbal condition are more likely to rely on guesswork than mental computation. Whatever affordances verbal probabilities may provide, the present research indicates that computability is not one of them. Nevertheless, computability is vital in many areas of expert judgment and decision making in which probability information is routinely communicated verbally (Dhami & Mandel, 2020; Mandel & Irwin, 2020b).

Our findings pertaining to the coherence of participants' responses contrast with those of Wallsten et al. (1993) who observed no advantage of a point numeric probability format over a verbal probability format in promoting additivity. However, Wallsten et al. (1993) measured additivity violations for binary complements that were averaged across multiple items. As Mandel (2005) noted, studies that have found additivity of binary complements tend to average probability estimates over multiple items, just as Wallsten et al. (1993) did. Accordingly, their study seems to have constituted a weaker test of the numeric superiority effect than that provided in the present research. Indeed, in our own experiments, the advantage of receiving point numeric probabilities over verbal probabilities for coherent responding was significantly greater in the more challenging multiplication task than in the easier averaging task. It would be useful to test the difference in violation of the additivity property using tasks that avoid binary complements and that are known to induce subadditivity-producing "unpacking effects" (e.g., Mandel, 2005, 2008; Rottenstreich & Tversky, 1997; Tversky & Koehler, 1994) or that have successfully yielded additivity violations using binary complements where probabilities were presented in the verbal format (Karvetski & Mandel, 2020).

Despite the aforementioned comparison between the present research and Wallsten et al. (1993), we propose that the tasks we used nevertheless constitute conservative tests of the numeric superiority hypothesis because we started with common, verbal probability phrases and then found average numeric equivalents for those terms. Alternatively, we could easily have devised experiments that started with numeric probabilities that would almost surely pose great difficulty for arithmetic computation. For instance, we could have asked participants to multiply a 45/1,000 chance by a 1/1,000,000 chance and then given them a log-linear scale that offers order-of-

magnitude granularity between probabilities of 0 and .01 and between .99 and 1. However, given that there are no words to adequately convey variation in probability across those ranges, conducting such an experiment seems unnecessary–the verbal equivalents (e.g., *extremely unlikely*) used to establish the corresponding inputs in the verbal condition would not back-translate well and would yield highly inaccurate results for arithmetic calculations. Yet, events characterized by extremely low probabilities, high timing unpredictability, and high consequence severities are precisely the sort that decision-makers in many consequential domains must prepare for (Makridakis & Taleb, 2009). It is for the same reason that we did not seek to ensure that response modes were matched to probability formats. We could, for instance, have required participants in the verbal condition to provide their averages and products in terms of a verbal-probability response. For instance, we might have asked them to provide a verbal probability that best captures their answer, and then we could have asked participants for a numeric probability equivalent of that term, which would be scored for accuracy. Although such an experiment would be informative, we do not believe it is necessary to examine the effect of probability format on arithmetic computation because the response "modes" do not simply differ in modality—they also differ in their potential for granularity and coverage over the possible range of computed values. Whereas numeric response modes cover the full possibility space, verbal response modes cover opaquely-defined patches of that space.

Future research on the present topic might focus on any of the following issues. First, to better understand the causal bases for the numeric superiority effect, it would be useful to conduct a more detailed process-tracing study to elucidate how arithmetic strategy use might differ when computing with numeric versus verbal probability inputs. For instance, perhaps individuals presented with verbal probabilities have difficulty translating them into numeric equivalents, which then need to be operated on while being held in working memory. If this task is too challenging, it might prompt individuals to simply guess. Second, it would be instructive to examine how individuals combine probability estimates that are communicated in mixed formats. For instance, if a decision-maker is presented with three numeric probability estimates and three verbal probability estimates and the decision-maker wishes to average them, would the numeric probabilities (which are easier to compute) be given disproportionate weight in the average? If so, would the differential weighting be due to primarily to differences in computability (e.g., perhaps the numeric estimates would be weighted more strongly because they are easier to work with) or perhaps to users' inferences about the underlying sources of uncertainty they convey (e.g., with verbal probabilities being suggestive of epistemic uncertainty and numeric probabilities being suggestive of aleatory uncertainty; Juanchich & Sirota, 2020)? Third, the numeric superiority hypothesis could be tested on other arithmetic operations, such as adding probabilities to compute the disjunctive probability of two or more mutually exclusive events or adjusting probabilities to reflect a proportional increase or decrease in the present value (e.g., new intelligence reveals that a particular threat probability just increased by one-third of its currently recorded level). Finally, given that the interpretation of verbal probabilities is context dependent in several respects (e.g., Brun & Teigen, 1988; Harris & Corner, 2011; Mandel, 2015; Wallsten, Fillenbaum, & Cox, 1986; Weber & Hilton, 1990), consistent with their linguistic function as relative adjectives whose meaning depends on their specific use (Clark, 1990), experiments could examine arithmetic ability in tasks that vary contextual factors such as event base-rate, severity and valence that have been shown to influence the interpretation of verbal probabilities. In fact, context-rich tasks could be used to test the reliability of arithmetic

computation across content domains. Given that translations of verbal to numeric probabilities are highly variable across content domains, especially for those less skilled in probabilistic judgment (Mellers, Baker, Chen, Mandel, & Tetlock, 2017), we might expect that less numerate individuals would show especially poor cross-domain reliability. More generally, given the effects of multiple contextual factors on the interpretation of verbal probabilities, we predict that verbal probabilities would yield less reliable computations than numeric probabilities.

In summary, our research revealed that simple arithmetic computations of the kind often required in several expert judgment and decision-making domains were more accurate and coherent if probability information was provided in the form of point numeric estimates rather than verbal estimates. Moreover, in our last experiment, we showed that computation using point estimates also outperformed computation based on numeric range information. The results are notable because organizations that generate probability estimates for consumption by other experts, decision-makers, or the general public typically provide such estimates in the form of verbal probabilities. These verbal probabilities are sometimes further anchored using imprecise numeric ranges, as in the Intergovernmental Panel on Climate Change standard or various intelligence community standards for communicating probabilities in intelligence estimates (e.g., Ho et al., 2015). Although coarse and fuzzy probability estimates may suffice or even be preferable in some communication contexts (Wallsten & Budescu, 1995; Zimmer, 1984), communicating with verbal probabilities may be woefully inadequate in others. Nor does our research (see Experiment 4) offer optimism for organizational remedies for the vagueness of verbal probabilities that call for translating a lexicon of verbal probability terms into numeric ranges (e.g., Dhami, 2018; Ho et al., 2015) and embedding numeric-range equivalents where such terms appear in text (Budescu et al., 2009; Budescu, Por, Broomell, & Smithson, 2014; Mandel & Irwin, 2020a; Wintle, Fraser, Wills, Nicholson, & Fidler, 2019). Rather, our findings suggest that when communicated probabilities serve as inputs for judgments or decisions that require mathematical computation of those values, they should be communicated as numeric point probabilities. As noted earlier, point probabilities are rated as conveying probability information more clearly than verbal probabilities (Collins & Mandel, 2019). Using point probabilities would also enable more granular assessments to be made and communicated to others, and this has been shown to yield substantial accuracy gains in forecasting (Friedman, Baker, Mellers, Tetlock, & Zeckhauser, 2018).

However, a valid concern with using point probabilities is the risk of communicating more precision in the estimate than is warranted. In fact, recent proposals to use numeric probabilities instead of verbal probabilities in communicating to end-users (e.g., Dhami & Mandel, 2020; Mandel & Irwin, 2020b; Mandel, Wallsten, & Budescu, 2020) have noted that numbers can be as imprecise as required since estimates can be expressed as ranges. Therefore, it would be useful to examine in future research how well point estimates accompanied by margins of error fare as a basis for arithmetic computation. Perhaps expressing imprecise numeric probabilities in that format (e.g., 70% ± 10%) supports computation more effectively than numerically equivalent ranges (e.g., 60% to 80%) given that the central estimate is explicit in the former case. What is clear, however, is that, whereas numerical probability ranges can be converted into point estimates with margins of error (Moore et al., 2009), verbal probabilities cannot be objectively translated from imprecise to precise representations. Additionally, verbal probabilities are vague as well as imprecise. Our research makes that abundantly clear: in each experiment, participants

(both expert analysts and non-experts) were substantially unreliable in their interpretations of verbal probabilities, even within a short timespan and across conceptually equivalent tasks. Under optimal conditions, this vagueness does not preclude the possibility of arithmetic computation: if one were to elicit, say, a membership function for each probability term used in a computational context, interval calculus could be applied to those functions (e.g., Dubois & Prade, 1978; Kosiński, Prokopowicz, & Ślęzak, 2003; Zadeh, 1975). However, most contexts in which verbal probabilities are communicated do not have this feature. The vagueness is neither quantified through elicitation of membership functions nor, in the absence of such transformations, amenable to fuzzy computational processes.

References

Baron, J., Scott, S., Fincher, K., & Metz, S. E. (2015). Why does the cognitive reflection test (sometimes) predict utilitarian moral judgment (and other things)? *Journal of Applied Research in Memory and Cognition, 4*(3), 265-284. doi:10.1016/j.jarmac.2014.09.003

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

Bilker, W. B., Wierzbicki, M. R., Brensinger, C. M., Gur, R. E., & Gur, R. C. (2014). Development of abbreviated eight-item form of the Penn Verbal Reasoning Test. *Assessment, 21*(6), 669-678. doi:10.1177/1073191114524270

Brun, W., & Teigen, K. H. (1988). Verbal probabilities: Ambiguous, context-dependent, or both? *Organizational Behavior and Human Decision Processes, 41*(3), 390-404. doi:10.1016/0749-5978(88)90036-2

Budescu, D. V., Broomell, S., & Por, H. (2009). Improving communication of uncertainty in the reports of the intergovernmental panel on climate change. *Psychological Science, 20*(3), 299-308. doi:10.1111/j.1467-9280.2009.02284.x

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

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

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

Budescu, D. V., & Wallsten, T. S. (1990). Dyadic decisions with numerical and verbal probabilities. *Organizational Behavior and Human Decision Processes, 46*(2), 240-263. doi:10.1016/0749-5978(90)90031-4

Clark, H. H. (1990). Quantifying probabilistic expressions: Comment. *Statistical Science, 5*(1), 12-16. doi:10.1214/ss/1177012243

Clausner, T. C., & Croft, W. (1999). Domains and image schemas. *Cognitive Linguistics, 10*(1), 1-31.

Collins, R. N. & Mandel, D. R. (2019). Cultivating credibility with probability words and numbers. *Judgment and Decision Making, 14*(6), 683-695.

Cokely, E. T., Galesic, M., Schulz, E., Ghazal, S., & Garcia-Retamero, R. (2012). Measuring risk literacy: The Berlin Numeracy Test. *Judgment and Decision Making, 7*(1), 25-47.

Dhami, M. K. (2018). Towards an evidence-based approach to communicating uncertainty in intelligence analysis. *Intelligence and National Security*, *33,* 257-272. doi:10.1080/02684527.2017.1394252

Dhami, M. K., & Mandel, D. R. (2020). Words or numbers? Communicating probability in intelligence analysis. *American Psychologist*. Advance online publication. doi:10.1037/amp0000637

Dhami, M. K., & Wallsten, T. S. (2005). Interpersonal comparison of subjective probabilities: Toward translating linguistic probabilities. *Memory & Cognition*, *33*(6), 1057-1068. doi:10.3758/BF03193213

Dubois, D., & Prade, H. (1978). Operations on fuzzy numbers. *International Journal of Systems Science*, *9*(6), 613-626. doi:10.1080/00207727808941724

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

Fan, Y., Budescu, D. V., Mandel, D., Himmelstein, M. (2019). Improving accuracy by coherence weighting of direct and ratio probability judgments. *Decision Analysis, 16*(3)*,* 197-217. doi:10.1287/deca.2018.0388

Friedman, J. A., Baker, J. D., Mellers, B. A., Tetlock, P. E., & Zeckhauser, R. (2018). The value of precision in probability assessment: Evidence from a large-scale geopolitical forecasting tournament. *International Studies Quarterly, 62*(2), 410-422. doi:10.1093/isq/sqx078

Friedman, J. A., & Zeckhauser, R. (2015). Handling and mishandling estimative probability: Likelihood, confidence, and the search for Bin Laden. *Intelligence and National Security, 30*(1), 77-99. doi:10.1080/02684527.2014.885202

Harris, A. J. L., & Corner, A. (2011). Communicating environmental risks: Clarifying the severity effect in interpretations of verbal probability expressions. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(6), 1571-1578. doi:10.1037/a0024195

Haran, U., Ritov, I., & Mellers, B. A. (2013). The role of actively open-minded thinking in information acquisition, accuracy, and calibration. *Judgment and Decision Making, 8*(3), 188-201.

Ho, E. Budescu, D. V., Dhami. M. K., & Mandel, D. R. (2015). Improving the communication of uncertainty in climate science and intelligence analysis. *Behavioral Science & Policy, 1*(2), 43-55.

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

Karvetski, C. W., & Mandel, D. R. (2020). Coherence of probability judgments from uncertain evidence: Does ACH help? *Judgment and Decision Making*, *15*(6), 939-958.

Kolesnik, K., Silska-Gembka, S., & Gierusz, J. (2019). The interpretation of the verbal probability expressions used in the IFRS – The differences observed between Polish and British accounting professionals. *Journal of Accounting and Management Information Systems, 18*(1),25-49. doi: 10.24818/jamis.2019.01002

Kosiński, W., Prokopowicz, P., & Ślęzak, D. (2003). On algebraic operations on fuzzy numbers. In: Kłopotek M.A., Wierzchoń S.T., & Trojanowski K. (Eds.), *Intelligent Information Processing and Web Mining. Advances in Soft Computing*, vol. 22. Berlin, Germany: Springer. doi:10.1007/978-3-540-36562-4_37

Langacker, R. (1987). *Foundations of Cognitive Grammar. Volume 1: Theoretical Prerequisites*. Palo Alto, CA: Stanford University Press.

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

Lipkus, I. M., Samsa, G., & Rimer, B. K. (2001). General performance on a numeracy scale among highly educated samples. *Medical Decision Making, 21*(1), 37-44. doi:10.1177/0272989X0102100105

Lipkus, I. M., & Peters, E. (2009). Understanding the role of numeracy in health: Proposed

theoretical framework and practical insights. *Health Education & Behavior, 36*(6), 1065-1081. doi:10.1177/1090198109341533

Makridakis, N., & Taleb, N. (2009). Living in a world of low levels of predictability. *International Journal of Forecasting, 25*(4), 840–844. doi:10.1016/j.ijforecast.2009.05.008

Mandel, D. R. (2005). Are risk assessments of a terrorist attack coherent? *Journal of Experimental Psychology: Applied*, *11*(4), 277-288. doi:10.1037/1076-898X.11.4.277

Mandel, D. R. (2008). Violations of coherence in subjective probability: A representational and assessment processes account. *Cognition*, *106*(1), 130-156. doi:10.1016/j.cognition.2007.01.001

Mandel, D. R. (2015). Accuracy of intelligence forecasts from the intelligence consumer's perspective. *Policy Insights from the Behavioral and Brain Sciences, 2*(1)*,* 111-120*.* doi:10.1177/2372732215602907

Mandel, D. R., & Barnes, A. (2018). Geopolitical forecasting skill in strategic intelligence. *Journal of Behavioral Decision Making, 31*(1), 127-137*.* doi:10.1002/bdm.2055

Mandel, D. R., & Irwin, D. (2020a). Facilitating sender-receiver agreement in communicated probabilities: Is it best to use words, numbers or both? *PsyArXiv*. doi:10.31234/osf.io/hm7zu

Mandel, D. R., & Irwin, D. (2020b). Uncertainty, intelligence, and national security decisionmaking. *International Journal of Intelligence and CounterIntelligence*. Advance online publication. doi:10.1080/08850607.2020.1809056

Mandel, D. R., Wallsten, T. S., & Budescu, D. V. (2020). Numerically-bounded language schemes are unlikely to communicate uncertainty effectively. *Earth's Future*. Advance online publication. doi:10.1029/2020EF001526

Meder, B., & Mayrhofer, R. (2017). Diagnostic causal reasoning with verbal information. Cognitive Psychology, 96, 54-84. doi:10.1016/j.cogpsych.2017.05.002Moore, R. E., Kearfott, R. B., & Cloud, M. J. (2009). *Interval analysis*. Philadelphia, PA: Society for Industrial and Applied Mathematics.

Mellers, B. A., Baker, J. D., Chen, E., Mandel, D. R., & Tetlock, P. E. How generalizable is good judgment? A multi-task, multi-benchmark study. *Judgment and Decision Making, 12*(4), 369-381.

Morgan, M. G. (1998). Commentary: Uncertainty analysis in risk assessment. *Human and Ecological Risk Assessment, 4*(1), 25-39. doi:10.1080/10807039.1998.11009680

Olson, M. J., & Budescu, D. V. (1997). Patterns of preference for numerical and verbal probabilities. *Journal of Behavioral Decision Making, 10*(2), 117–131. doi:10.1002/(SICI)1099-0771(199706)10:2<117::AID-BDM251>3.3.CO;2-Z

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*(4), 867-872. doi:10.1016/j.jesp.2009.03.009

Perrow, C. (1984). *Normal accidents: Living with high-risk technologies*. New York: Basic Books.

Rapoport, A., Wallsten, T. S., Erev, I., & Cohen, B. L. (1990). Revision of opinion with verbally and numerically expressed uncertainties. *Acta Psychologica, 74*(4), 61-79. doi:10.1016/0001-6918(90)90035-E

Rottenstreich, Y., & Tversky, A. (1997). Unpacking, repacking, and anchoring: Advances in support theory. *Psychological Review, 104*(2), 406–415.

doi:10.1037/0033-295X.104.2.406
Stanovich, K. E., and West, R. F. (2000). Individual differences in reasoning: Implications for the rationality debate? *Behavioral and Brain Sciences*, *23*(5), 645–665. doi:10.1017/S0140525X00003435
Toplak, M. E., West, R. F., & Stanovich, K. E. (2017). Real-world correlates of performance on heuristics and biases tasks in a community sample: Heuristics and biases tasks and outcomes. *Journal of Behavioral Decision Making, 30*(2), 541-554. doi:10.1002/bdm.1973
Tversky, A., & Koehler, D. J. (1994). Support theory: A nonextensional representation of subjective probability. *Psychological Review*, *101*(4), 547–567. doi:10.1037//0033-295X.101.4.547
Wallsten, T. S., & Budescu, D. V. (1995). A review of human linguistic probability processing: General principles and empirical evidence. *The Knowledge Engineering Review, 10*(1), 43-62. doi:10.1017/S0269888900007256
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. doi:10.1287/mnsc.39.2.176
Wallsten, T. S., Budescu, D. V., Zwick, R., & Kemp, S. M. (1993). Preferences and reasons for communicating probabilistic information in verbal or numerical terms. *Bulletin of the Psychonomic Society, 31*(2), 135–138. doi:10.3758/BF03334162
Wallsten, T. S., Fillenbaum, S., & Cox, J. A. (1986). Base rate effects on the interpretations of probability and frequency expressions. *Journal of Memory and Language, 25*(5), 571-587. doi:10.1016/0749-596X(86)90012-4
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. doi:10.1037/0096-3445.115.4.348
Weber, E. U., & Hilton, D. J. (1990). Contextual effects in the interpretations of probability words: Perceived base rate and severity of events. *Journal of Experimental Psychology: Human Perception and Performance, 16*(4), 781-789. doi:10.1037/0096-1523.16.4.781
Wiles, M. D., Duffy, A., & Neill, K. (2020). The numerical translation of verbal probability expressions by patients and clinicians in the context of peri-operative risk communication. *Anaesthesia*, *75* (Suppl. 1), e39–e45. doi:10.1111/anae.14871
Wintle, B. C., Fraser, H., Wills, B. C., Nicholson, A. E., & Fidler, F. (2019). Verbal probabilities: Very likely to be somewhat more confusing than numbers. *PloS One, 14*(4), e0213522. doi:10.1371/journal.pone.0213522
Zadeh, L. A. (1975). The concept of a linguistic variable and its application to approximate reasoning—III. *Information Sciences*, *9*(1), 43-80. doi:10.1016/0020-0255(75)90017-1
Zimmer, A. C. (1984). A model for the interpretation of verbal predictions. *International Journal of Man-Machine Studies, 20*(1), 121-134. doi:10.1016/S0020-7373(84)80009-7
Zwick, R., Budescu, D. V., & Wallsten, T. S. (1988). An empirical study of the integration of linguistic probabilities. In T. Zétényi (Ed.), *Advances in Psychology* (Vol. 56, pp. 91-125). Amsterdam: North-Holland. doi:10.1016/S0166-4115(08)60483-5

David R. Mandel, a senior DRDC scientist and Adjunct Professor of Psychology, York University, studies human judgment and decision-making. Mandel was Chairman of a NATO scientific team that received the 2020 SAS Panel Excellence Award. He serves on the editorial boards of *Decision*, *Futures and Foresight Science*, and *Judgment and Decision Making*.

Mandeep K. Dhami, PhD is Professor in Decision Psychology, Middlesex University, London. Mandeep applies her expertise to the criminal justice and intelligence analysis domains. She has authored 130 scholarly publications and edited 'Judgment and Decision Making as a Skill' (Cambridge University Press). Mandeep is co-Editor of *Judgment and Decision Making*.

Serena Tran is currently completing the final year of her B.A. (Honours) in Psychology at the University of Waterloo. She has held a coop position at DRDC in 2019 and worked at the Ministry of Transportation in 2020. Tran was also awarded the 2020 Psychology Memorial Scholarship.

Daniel Irwin holds an M.S. in Applied Intelligence from Mercyhurst University and works for Canada's Department of National Defence, conducting research on intelligence analysis and uncertainty communication. In 2020, Irwin was awarded the North Atlantic Treaty Organization System Analysis and Studies Panel Excellence Award.

Table 1. Analysis of covariance on mean standardized absolute error (MSAE) in Experiment 1

Source	*MSE*	*F*	*p*	η_p^2
Intercept	8.70	99.03	.000	.324
Numeracy	0.46	5.24	.023	.025
PVRT	0.19	2.19	.141	.010
AOT	0.16	1.80	.182	.009
Format	0.82	9.44	.002	.043
Error (between subjects)	0.87			
Operation	0.77	17.85	.000	.079
Format × Operation	0.00	0.00	.972	.000
Error	0.04			

Note. *df* = 1, 208. Model based on Type III sum of squares. For brevity, interactions with covariates are not reported.

Table 2. Analysis of covariance on incoherence scores in Experiment 1

Source	*MSE*	*F*	*p*	η_p^2
Intercept	74.26	167.98	.000	.447
Numeracy	7.14	16.16	.000	.072
PVRT	2.49	5.64	.018	.026
AOT	0.93	0.21	.647	.001
Format	13.35	30.21	.000	.127
Error (between subjects)	0.44			
Operation	7.97	16.52	.000	.074
Format × Operation	2.57	5.33	.022	.025
Error	0.48			

Note. *df* = 1, 208. Model based on Type III sum of squares. For brevity, interactions with covariates are not reported.

Table 3. Analysis of covariance on mean standardized absolute error (MSAE) in Experiment 2

Source	*MSE*	*F*	*p*	η_p^2
Intercept	5.11	118.85	.000	.419
Numeracy	0.23	5.23	.023	.026
PVRT	0.15	3.39	.067	.017
AOT	0.00	0.03	.857	.000
Format	1.40	27.12	.000	.123
Operation	7.64	32.62	.000	.144
Format × Operation	0.31	7.17	.008	.036
Error	0.04			

Note. *df* = 1, 194. Model based on Type III sum of squares.

Table 4. Analysis of covariance on coherence scores in Experiment 2

Source	*MSE*	*F*	*p*	η_p^2
Intercept	42.89	110.52	.000	.363
Numeracy	2.75	7.08	.008	.035
PVRT	0.59	1.53	.218	.008
AOT	0.44	1.13	.288	.006
Format	22.07	56.88	.000	.227
Operation	44.25	114.02	.000	.370
Format × Operation	3.67	9.45	.002	.046
Error	0.39			

Note. *df* = 1, 194. Model based on Type III sum of squares.

Table 5. Analysis of covariance on mean standardized absolute error (MSAE) in Experiment 4

Source	*df*	*MSE*	*F*	*p*	η_p^2
Intercept	1	7.75	143.26	.000	.300
Numeracy	1	0.51	9.38	.002	.027
PVRT	1	0.17	3.10	.079	.009
AOT	1	0.03	0.49	.484	.001
Format	2	0.28	5.22	.006	.030
Operation	1	4.81	88.91	.000	.210
Format × Operation	2	0.02	0.31	.733	.002
Error	334	0.05			

Note. Model based on Type III sum of squares.

Table 6. Percentage of participants' self-reported strategy use by format.

	Format			
Strategy	Verbal	Range	Point	Total
Mental calculation	33.3	40.4	47.7	40.2
Rough estimation	66.7	59.6	52.3	59.8

Table 7. Analysis of covariance on incoherence scale in Experiment 4

Source	*df*	*MSE*	*F*	*p*	η_p^2
Intercept	1	42.10	89.48	.000	.211
Numeracy	1	1.99	4.23	.040	.013
PVRT	1	0.44	0.93	.336	.003
AOT	1	0.35	0.75	.389	.002
Format	2	4.14	8.80	.000	.050
Operation	1	119.35	253.63	.000	.432
Format × Operation	2	2.07	4.39	.013	.026
Error	334	0.47			

Note. Model based on Type III sum of squares.

Figure 1. Mean standardized absolute error (MSAE) by format and operation in Experiment 1. Error plots show estimated marginal means and 95% confidence intervals from ANCOVA. Box-and-whisker plots show the distribution of MSAE for each condition based on sample data.

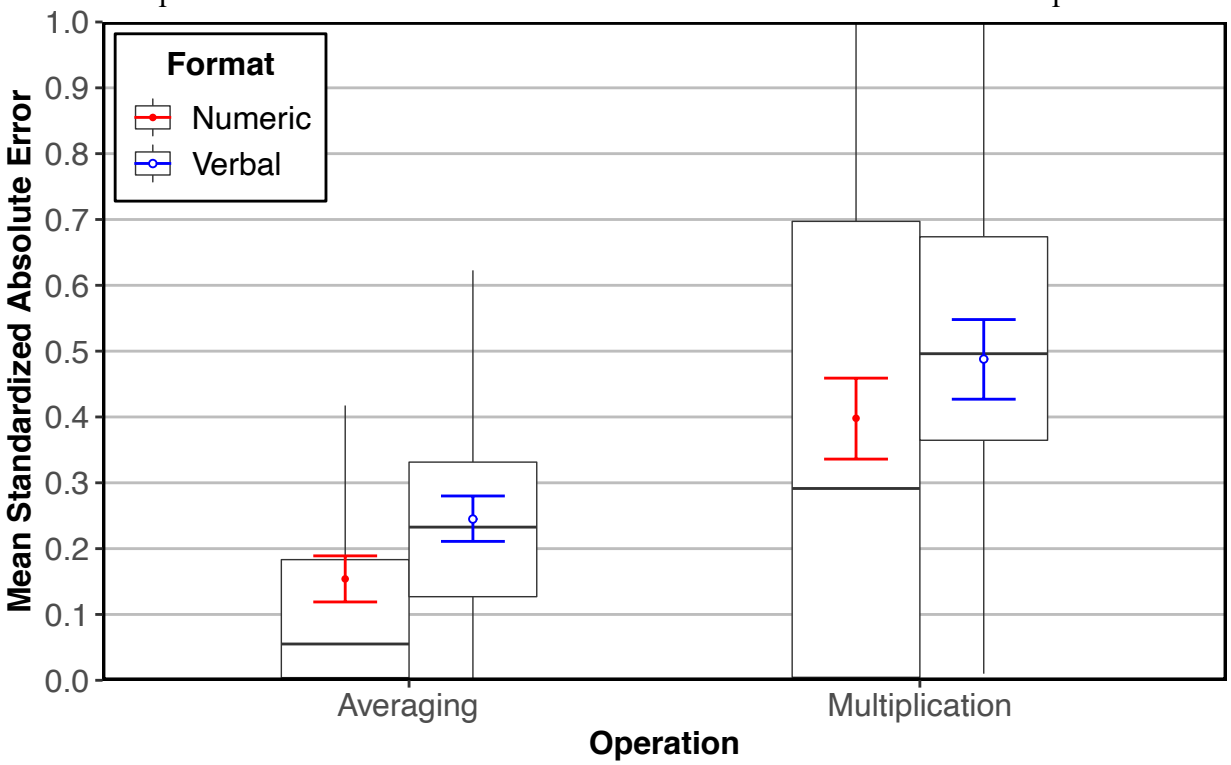

Figure 2. Mean standardized absolute error (MSAE) by format and operation in Experiment 2. Error plots show estimated marginal means and 95% confidence intervals from ANCOVA. Box-and-whisker plots show the distribution of MSAE for each condition based on sample data.

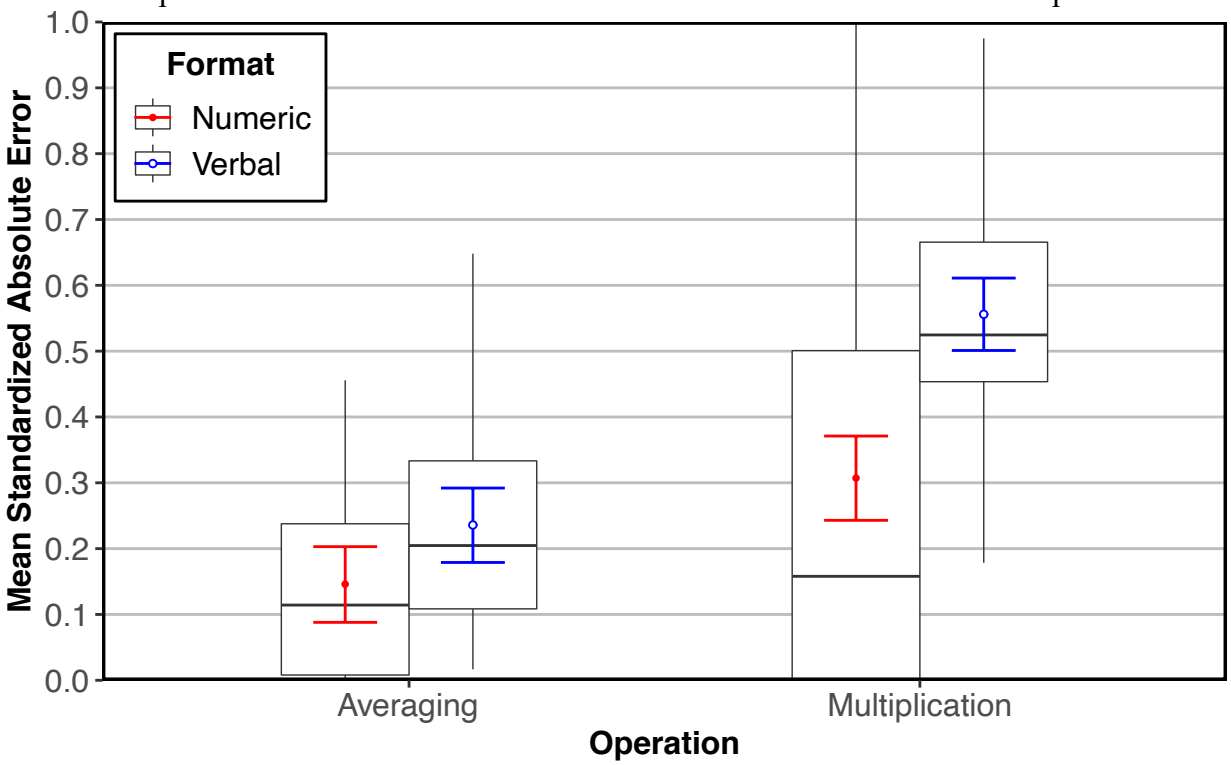

Figure 3. Mean standardized absolute error (MSAE) by format and operation in Experiment 4. Error plots show estimated marginal means and 95% confidence intervals from ANCOVA. Box-and-whisker plots show the distribution of MSAE for each condition based on sample data.

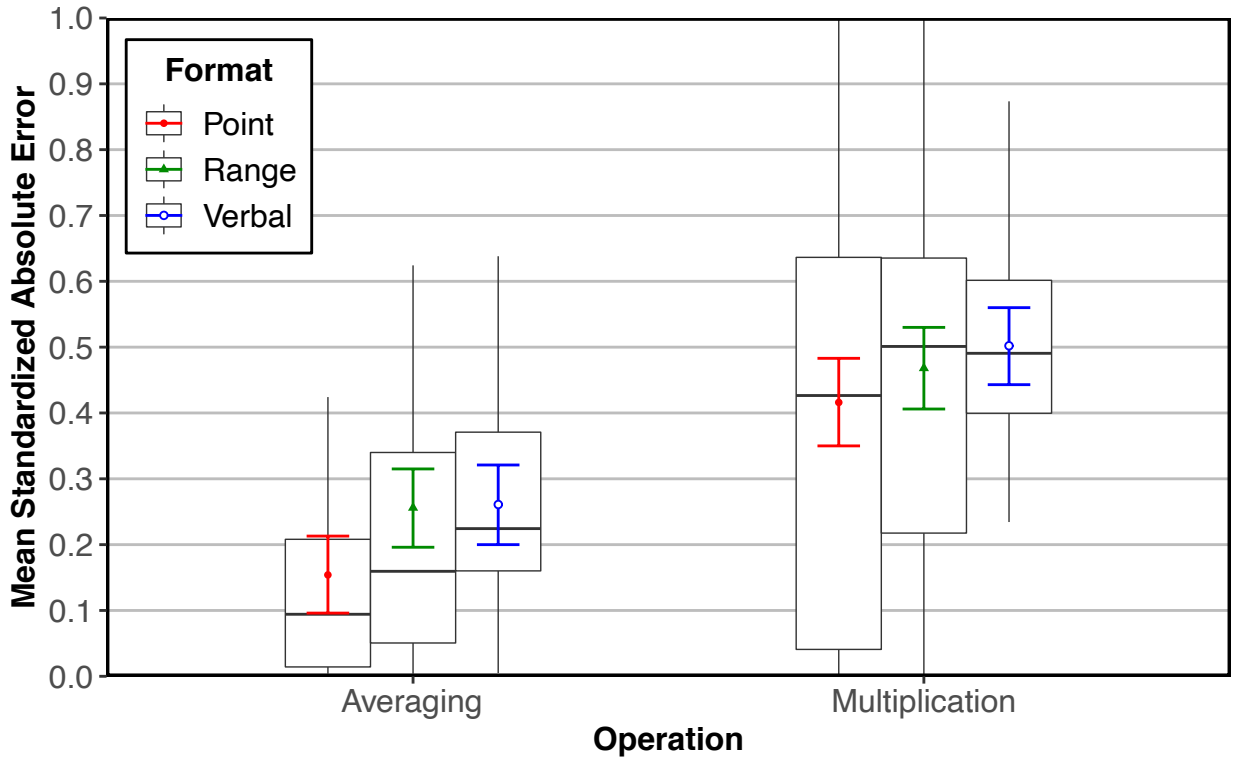

Figure 4. Mean incoherence by format and operation in Experiment 4. Error plots show estimated marginal means and 95% confidence intervals from ANCOVA.

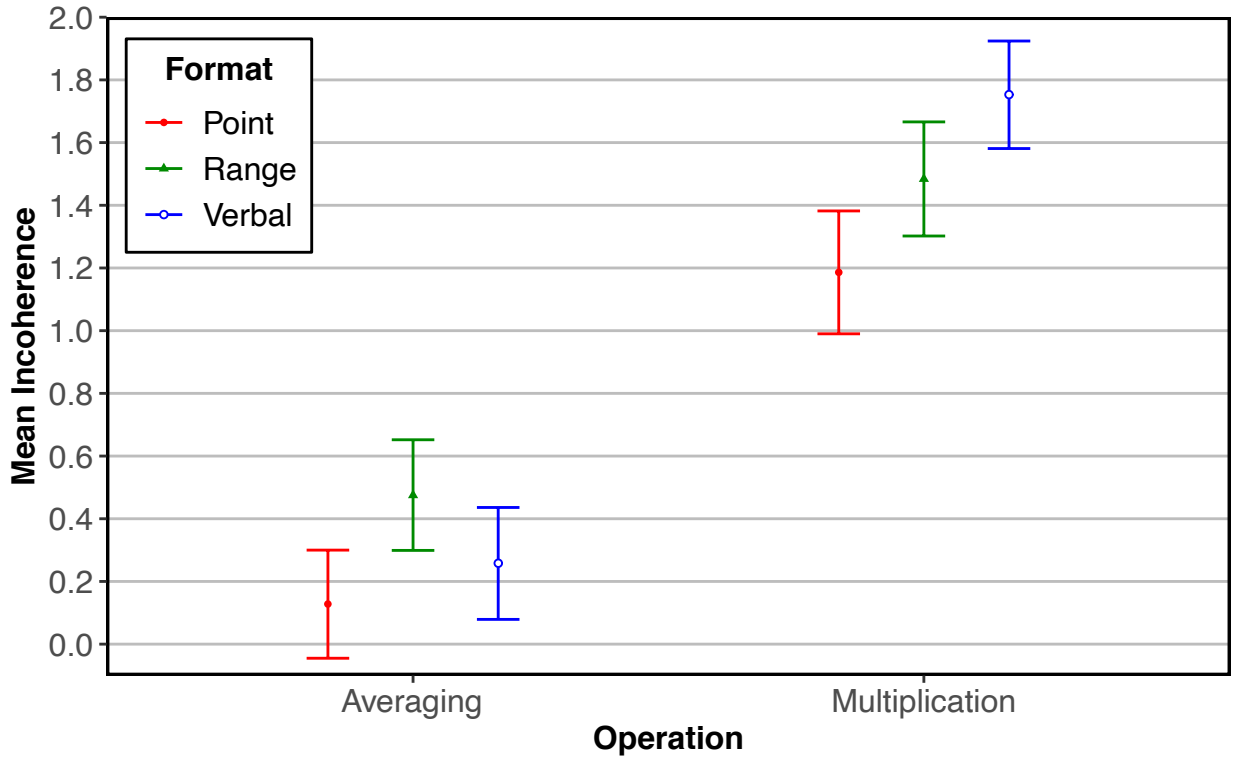

Figure 5. Mediation model in Experiment 4. Values are standardized regression coefficients and the value in parentheses controls for the mediator. *$p < .05$, **$p < .01$, ***$p < .001$.