Judgment and Decision Making, Vol. 15, No. 5, September 2020, pp. 783–797

Effect of confidence interval construction on judgment accuracy

David R. Mandel* Robert N. Collins† Evan F. Risko‡ Jonathan A. Fugelsang§

Abstract

Three experiments (N = 550) examined the effect of an interval construction elicitation method used in several expert elicitation studies on judgment accuracy. Participants made judgments about topics that were either searchable or unsearchable online using one of two order variations of the interval construction procedure. One group of participants provided their best judgment (one step) *prior* to constructing an interval (i.e., lower bound, upper bound, and a confidence rating that the correct value fell in the range provided), whereas another group of participants provided their best judgment *last,* after the three-step confidence interval was constructed. The overall effect of this elicitation method was not significant in 8 out of 9 univariate tests. Moreover, the calibration of confidence intervals was not affected by elicitation order. The findings warrant skepticism regarding the benefit of prior confidence interval construction for improving judgment accuracy.

Keywords: judgment, accuracy, confidence intervals, order effect

1 Introduction

Improving the accuracy of our judgments represents a major effort in decision science and cognitive science, more broadly. This work has important implications for several applied domains where expert judgment is required, such as medicine (Berner & Graber, 2008; Dawson & Arkes, 1987), clinical practice (Dawes, 1979; Oskamp, 1965), law (Goodman-Delahunty, Granhag, Hartwig & Loftus, 2010), finance (Önkal, Yates, Simga-Mugan & Öztin, 2003), and geopolitical forecasting and strategic intelligence analysis (Dhami, Mandel, Mellers & Tetlock, 2015; Mandel & Barnes, 2018; Tetlock, 2005). For example, intelligence analysts often need to make rapid judgments under conditions of uncertainty and these judgments often inform mission-critical decisions (e.g., Fingar, 2011; Friedman, 2019; Mandel, 2019). One strategy for improving judgment involves designing structured elicitation methods that reduce potential bias or error in judgment (e.g., confirmation bias, overconfidence). That is, how individuals are probed for a given judgment is one potential target for intervention. If effective, then such methods could offer a reliable route to improving judgment accuracy.

A popular example of an elicitation method that appears to improve judgment accuracy is the "consider the opposite" approach. Lord, Lepper and Preston (1984) demonstrated that asking individuals to consider the opposite point of view reduced individuals' tendency to interpret evidence in terms of their prior beliefs. Herzog and Hertwig (2009) use a similar exhortation in their dialectical bootstrapping method in which individuals provide two estimates, the second of which follows instructions to imagine the first estimate was incorrect and the reasons why that might be. Herzog and Hertwig (2009) found that the accuracy of the average estimate was higher in that condition than a control condition in which individuals provided two estimates without the instruction to "consider the opposite" (see also Herzog & Hertwig, 2014; Müller-Trede, 2011). Herzog and Hertwig (2009) suggested that the consider-the-opposite instruction encourages more divergent estimates based on different sources of knowledge and, hence, greater benefits of aggregation. In a related vein, Williams and Mandel (2007) found that "evaluation frames", which make the complementary hypothesis explicit and hence contrastively evaluable (e.g., "How likely is x rather than $\neg x$ to occur?"), fostered greater judgment accuracy than "economy frames", which explicated only the focal hypothesis (e.g., "How likely is x to occur?").

A similar approach has been attempted in efforts to reduce another pervasive bias — overconfidence in interval estimates or "overprecision"; namely, the confidence intervals people generate are often too narrow or overly precise (Alpert & Raiffa, 1982; Soll & Klayman, 2004; Moore & Healy, 2008; Pitz, 1974; Teigen & Jørgenson, 2005). For example, Soll and Klayman (2004) demonstrated that individuals were more overconfident when asked to produce a single 80% confidence interval compared to providing two separate lower- and upper-bound 90% estimates. Soll and Klayman (2004)

This research was funded by Canadian Safety and Security Program project CSSP-2018-TI-2394. We thank Daniel Irwin for assistance with this research.

*Intelligence, Influence and Collaboration Section, Defence Research and Development Canada. ORCID: 0000-0003-1036-2286. Email: drmandel66@gmail.com.

†Intelligence, Influence and Collaboration Section, Defence Research and Development Canada. ORCID: 0000-0002-1714-7215.

‡HumanSystems® Incorporated. ORCID: 0000-0001-5702-0350.

§HumanSystems® Incorporated. ORCID: 0000-0002-6342-7023.

suggested that, like the consider-the-opposite approach, having to generate multiple point estimates (i.e., lower and upper bounds) yields a wider evidence base from which to generate the point estimates. Consistent with this idea, Teigen and Jørgensen (2005) asked one group of participants to provide a typical range estimate (i.e., both a lower and upper bound in the same query) and two other groups to each provide only a lower or upper bound. The latter interval was wider than when the same individual gave both bounds suggesting that thinking about each bound independently likely leads to more disparate evidence being retrieved than thinking about them simultaneously. Teigen and Jørgensen (2005) also found that individuals were less overprecise when they were allowed to assign a confidence level to an interval versus having to produce an interval of a specified confidence (e.g., 90%). However, the degree of overprecision appears to be related to the degree of confidence referenced in fixed confidence-level elicitations. For instance, Budescu and Du (2007) found that whereas 90% confidence intervals were overprecise, 70% intervals were well-calibrated, and 50% intervals were underprecise. Taken together, it appears that there are many ways in which elicitations can be structured to reduce bias and improve judgment accuracy.

2 The present investigation

In the present research, we continue this line of inquiry by examining a specific instantiation of interval elicitation, which prescribes that interval construction should precede the elicitation of best estimates. This approach has been adopted in both the three- and four-step methods (Speirs-Bridge et al., 2010), which have been used in several expert elicitation studies (e.g., Adams-Hosking et al., 2016; Burgman et al., 2011) and in the comprehensive structured elicitation IDEA protocol (for Investigate-Discuss-Estimate-Aggregate; e.g., Hanea et al., 2017; Hemming, Walshe, Hanea, Fidler & Burgman, 2018). Other methods such as the SHELF protocol (for Sheffield Elicitation Framework) also prescribe eliciting upper and lower bounds prior to best estimates (O'Hagan, 2019). Such methods are inspired by research showing the beneficial effect of interval estimation (e.g., Soll & Klayman, 2004; Teigen & Jørgensen, 2005). However, they go further by prescribing a fixed order in which estimates should be elicited from assessors. For instance, in the four-step method (Speirs-Bridge et al., 2010), after providing assessors with a query (e.g., what is the number of native plant species in a given region?) the assessors are asked to provide in the following order: (1) the lowest realistic numerical estimate, (2) the highest realistic numerical estimate, (3) a best estimate, and (4) a confidence judgment in the form of an estimate of the likelihood that the true value of the assessed quantity falls in the credible interval defined by the first two estimates provided.

These features of such "IBBE protocols" (for Intervals Before Best Estimates) can potentially support two ameliorative functions. First, because the bounds are estimated before the best estimate, this might prompt consideration of a wider range of relevant information. For instance, citing Morgan and Henrion (1990), Hemming et al. (2018) states, "asking for lowest and highest estimates first [in the three- or four-point methods], encourages consideration of counterfactuals and the evidence for relatively extreme values, and avoids anchoring on best estimates" (p. 174). As the quote suggests, eliciting bounds before best estimates may improve the accuracy of the latter by stimulating consideration of worst- and best-case scenarios or multiple viewpoints. However, to the best of our knowledge, this hypothesized order effect of interval construction on the accuracy of best estimates has not been empirically tested. Therefore, one aim of this research was to test the validity of this hypothesis.

A second ameliorative function of IBBE protocols is to improve the calibration of confidence by allowing assessors to assign a confidence level to their credible interval after the latter has been estimated. As noted earlier, prior research has found evidence to support this method (Soll & Klayman, 2004; Teigen & Jørgensen, 2005) and we do not pursue that issue further here. However, little research has examined whether the elicitation of a best estimate prior to interval construction has any effect on the latter process. We conducted three sets of analyses. First, we examined whether elicitation order influenced the *range* of participant' credible intervals. One hypothesis suggested by the work of Tversky and Kahneman (1974) and Epley and Gilovich (2006) is that intervals would be narrower if best estimates are elicited first because participants would anchor on the estimate and adjust until a plausible value is reached. In contrast, participants who generate intervals before their best estimates might anchor on the lower and upper limits of the probability scale and adjust away from those limits until a plausible value is reached. Second, we examined whether elicitation order affected participants' level of *confidence* in their credible range. If credible intervals were narrower after the elicitation of best estimates than if prior to their elicitation, one might expect confidence to be lower in the "best-first" case because estimate precision (and informativeness) is greater (Yaniv & Foster, 1995, 1997). That is, it should be easier to be confident that one's interval captures the true value if it is wider. Finally, we examined whether elicitation order affected *calibration* of confidence. The anchoring-and-adjustment hypothesis suggests that overprecision might be amplified by generating the best estimate first because confidence intervals will tend to be narrower than when the intervals are constructed before providing a best estimate. However, contrary to the anchoring-and-adjustment hypothesis, Block and Harper (1991) found that generating an explicit best estimate prior to a confidence interval improved the calibration of confidence by reducing overprecision, whereas Soll and

Klayman (2004, Experiment 3) found that hit rates did not depend on whether the best estimate (defined in their study as the median estimate) was elicited before, in the middle or after lower and upper bounds. Therefore, substantial uncertainty regarding the effect of generating best estimates on confidence intervals remains.

We report three experiments that prompt participants to use one of two variants of a modified four-step method to answer general-knowledge and estimation-type questions depending on the condition to which they were randomly assigned. In the "best-first" condition, best estimates were elicited before eliciting lower-bound, upper-bound, and confidence estimates, respectively. In the "best-last" condition, participants provided their best estimates after generating the three estimates required for confidence interval construction. We used three different types of questions. All questions required responses in a percentage format (i.e., between 0% and 100%). A third of the items were general-knowledge questions (e.g., "What percentage of a typical adult human's bones are located in its head?"). However, when research is conducted online, as in the present research, individuals could search for answers on the Internet even if instructed not to do so. To address this concern, the remaining items were unsearchable. Following Gaertig and Simmons (2019), this was achieved by splitting our samples into two groups, each of which had a set of queries that required them to estimate the percentage of respondents who exhibited a certain behavior or that correctly answered a certain general-knowledge question. The correct values for these queries were unsearchable and were determined based on the values in the survey sample.

3 Experiment 1a and 1b

3.1 Methods

3.1.1 Sampling strategy and participants

The sample size was set to ensure sufficient power to detect effects of medium size using a multivariate analysis of variance (MANOVA) of judgment accuracy. We further oversampled in anticipation of having to exclude participants for various reasons, as we describe in the Case exclusions subsection of the Results. Accordingly, 350 participants in Experiment 1a and 417 participants in Experiment 1b completed our study online via Qualtrics Panels. The study was available to adults between the ages of 18 and 60 years of age who have English as their first language, and were Canadian or American citizens (self-reported). After case exclusions, we retained a sample of 299 participants in Experiment 1a (mean age = 41.75; 166 females and 133 males; 173 Canadian citizens, 114 US, and 12 dual) and 357 participants in Experiment 1b (mean age = 38.37; 205 females, 149 males, 1 who preferred not to say, and 2 missing responses; 139 Canadian citizens, 209 US, and 9 dual). Participants were compensated from the panel provider (i.e., Qualtrics) for this study. The specific type of compensation varied (e.g., cash, gift card) and had a maximum value of $5 US (or equivalent in Canadian dollars).

3.1.2 Design

We used a between-groups design wherein the order in which participants provided their confidence intervals and best estimates was manipulated. In the best-last condition, participants were asked to estimate a lower bound, an upper bound, a confidence level that their interval captured the true value before providing their best estimate. In the best-first condition, participants provided their best estimate before providing the three estimates pertinent to confidence interval construction.

3.1.3 Materials and procedures

Supplementary materials including the full experimental protocol, data, and other supporting files are available from the Open Science Foundation project page https://osf.io/k3jhq/.

After reviewing the information letter and consent form, participants answered basic demographic questions (i.e., age, sex, education, nationality, citizenship, first language) before completing a series of tasks in a counterbalanced order. Following completion of the tasks, participants were debriefed about the purpose of the study. The other tasks included an eight-item version of the International Cognitive Ability Resource (ICAR; Condon & Revelle, 2014, see Appendix E in the online supplementary materials), an eight-item version of the Actively Open-Minded thinking scale (Baron, Scott, Fincher & Metz, 2015), and one of two sets of seven items from the 14-item Bias Blind Spot scale (Scopelliti et al., 2015). The focus of the present report is on the estimation task. However, ICAR was used to assess data quality. ICAR is a test of general cognitive ability. It is a forced choice, six-option multiple-choice test. The items are designed to tap into verbal reasoning, numeracy, matrix reasoning, and mental rotation abilities (Condon & Revelle, 2014). Participants completed a shorter, eight-item version of this test. They received a score of 0–8, one point for each correct answer. No partial marks are given.

The latter two tasks are the focus of a separate investigation and are not presented here.

Estimation task. Participants in the best-first condition first received the following instructions explaining the elicitation protocol to be used in providing their estimates (see also Appendix A in the online supplementary materials):

> In the following you will be presented with a series of questions (e.g., what percentage of stuffed

> animals are teddy bears?) for which we will ask you to provide four different estimates.
>
> Realistically, what is your BEST response?
>
> Realistically, what do you think the LOWEST plausible value could be?
>
> Realistically, what do you think the HIGHEST plausible value could be?
>
> How confident are you that the interval you created, from LOWEST to HIGHEST, could capture the true value? Please enter a number between 50 and 100%?
>
> Each estimate will be in the form of a percentage. Please try to be as accurate as possible. PLEASE DO NOT LOOK UP THE ANSWERS (e.g., on the Internet). In the next few pages we will review how to respond to each of the questions above. It is important that you understand how to respond to each. Please read carefully.

In the best-last condition, the questions were ordered such that the first question presented above was presented last. These instructions were followed by two instruction comprehension questions and feedback (see Appendix B in the online supplementary materials).

Participants then made four judgments regarding each of 18 questions. There were two sets of 18 questions and each participant received one set (328 participants received Set A and the other 328 received Set B across Experiments 1a and 1b). Each set of items was composed of three different question types: (a) searchable knowledge, (b) unsearchable knowledge, and (c) unsearchable behavior. There were six items of each type. A general description of these items is provided in the Introduction and all of the items are presented in Appendix C in the online supplementary materials. For each question, participants provided answers to the following queries: (a) "Realistically, what do you think the LOWEST plausible value could be?", (b) "Realistically, what do you think the HIGHEST plausible value could be?", (c) "How confident are you that the interval you created, from LOWEST to HIGHEST, could capture the true value? Please enter a number between 50 and 100%?", and (d) "Realistically, what is your BEST response?". Responses to (a), (b), and (d), were provided on a 101-point sliding scale ranging from 0% to 100%; responses to (c) used a sliding scale between 50% and 100%. The slider had a default position of the lowest value on the scale for all four questions. In both orders, the interval construction questions (i.e., a, b, c) were always presented on the same page and the best estimate was always elicited on a separate page (an example is shown in Appendix A).

Answers for unsearchable questions. To generate answers for the unsearchable questions, each participant completed six knowledge and six behavioral questions (see Appendix D in the online supplementary materials for a complete list of items). There was a total of 24 items, and each participant received one set of 12 (i.e., Lists A and B). The particular set received determined the estimation set they would receive, such that participants receiving List A would estimate on List B and vice versa. For example, one set of items would include the estimation question "What percentage of survey respondents reported having pet a cat in the last 3 days?" and the other set would include the question "Have you pet a cat in the last 3 days?" Each item required a yes/no (e.g., Have you visited a country in Europe in the last ten years?) or true/false (e.g., Volvo is a Swedish car manufacturer) response.

Data quality We took several steps to improve data quality. First, we included three attention-check items, one prior to the survey, and two within the survey. The pre-survey attention check was employed to assess the degree to which a participant was likely to do the task (and not just quickly speed through the survey randomly clicking responses). Participants responded to the following question, where they needed to respond "No" to proceed to the survey:

> The survey that you are about to enter is longer than average, and will take about 30 to 60 minutes. There is a right or wrong answer to some of the questions, and your undivided attention will be required. If you are interested in taking this survey and providing your best answers, please select "No" below. Thank you for your time!

Within the survey, we presented participants with two additional attention checks to be used as a means of excluding data from participants during data analysis who were not attending to the task. These were the following:

1. The value of a quarter is what percentage of a dollar?
2. In the following alphanumeric series, which letter comes next? A B C D E with options (1) I (2) H (3) G (4) F (5) J.

The (1) attention check was embedded in the estimation task and the (2) attention check was embedded in one of the individual difference measures. Lastly, we also monitored speed of responding. We set the minimum plausible duration to 500 seconds, and only retained data from participants who spent more than 500 seconds completing the survey.

Second, we analyzed only those items for which participants provided a complete and coherent set of estimates. Abstaining from one or more of the lowest, highest, best, or confidence judgments resulted in the exclusion of the item, as did the violation of the following constraint:

$$L \leq B \leq U$$

That is, lower bounds had to be less than or equal to the upper bounds and the best estimate had to fall between (or precisely upon) those bounds.

Finally, we performed two additional data quality checks. First, we tested whether participants performed better than chance for each of the six combinations of questions type (SK, UK, UB) and list (A, B). To generate accuracy estimates for chance responses, we simulated 200,000 participants (100,000 for List A, 100,000 for List B) who selected randomm responses between 0 and 100 for each of the 18 best-estimate questions and scored these against the same truth vector used to compute accuracy for participant data. We then compared participants' accuracy to the random-response accuracy using six one-tailed t tests. Second, we examined the correlation between performance on ICAR and estimation accuracy.

Statistical procedure Accuracy of best estimates was measured using mean absolute error (MAE; the mean absolute difference between the participant's best estimate and the correct answer over items within each question type). We refer to the grand mean of MAE computed over participants as GMAE. Question type GMAE was calculated only where participants provided complete and coherent estimates for at least half of the items. Total GMAE was calculated only for those with a complete set of question type GMAEs.

3.2 Results

3.2.1 Data quality

Participants had to meet the following inclusion criteria to be included in the analyses: (1) pass the initial pre-study attention check, (2) pass the letter sequence attention check, (3) report "English first language" in the demographics, and (4) report Canadian and/or American citizenship. We did not use one of the attention checks (i.e., the value of a quarter) as an exclusion criterion because of very poor performance that possibly indicated that a significant portion of the participants misunderstood the question. In addition, participants with a large number of missing responses, inappropriate responses (e.g., to open text questions), and/or overly systematic response patterns were removed. In Experiment 1a, 37, 12, and 2 participants were excluded for demographic reasons, failure of the attention check, and missing or inappropriate responses, respectively. In Experiment 1b, 30, 18, and 17 participants were excluded for these same reasons, respectively. Because some participants were excluded for multiple reasons, the filtering procedure resulted in the exclusion of 51 participants for Experiment 1a and 60 for Experiment 1b.

In Experiment 1a, 74.21% (SD = 34.02%) of participants' sets of estimates were complete and coherent, whereas that figure was 75.12% (SD = 31.92%) in Experiment 1b. The complete and coherent requirement excluded an additional 120 participants in Experiment 1a and a final sample of 178 participants in Experiment 2b, leaving final sample sizes of 179 and 217 in Experiments 1a and 1b, respectively. Participants excluded on the basis of incoherence scored significantly lower on ICAR in Experiments 1a than the coherent participants retained in our samples ($r_{pb}[299] = .36, p < .001$) and 1b ($r_{pb}[357] = .53, p < .001$).

Participants' accuracy was significantly above chance for each question type in Lists A and B in both Experiment 1a and Experiment 1b (all $p < .001$). The full analysis is reported in Appendix F in the online supplementary materials.

A final data quality check revealed that ICAR (M = 3.59, SD = 1.99, after exclusion of the incoherent participants) correlated with GMAE in both Experiment 1a ($r[177] = -.197$, $p = .008$) and Experiment 1b ($r[215] = -.202$, $p = .003$), indicating that participants who performed well on ICAR also had more accurate best estimates.

3.2.2 Best estimates

Figure 1 shows the distributions of GMAE by experiment, question type, and elicitation order, whereas Table 1 shows the corresponding means and standard deviations. We conducted a MANOVA in each experiment with elicitation order as a fixed factor, and the three MAE measures corresponding to question type as dependent measures. Table 2 summarizes the multivariate results and Table 3 shows the parameter estimates for the univariate results. As can be seen in Table 2, the effect of elicitation order was not significant in Experiment 1a but was significant in Experiment 1b. The univariate parameter estimates in Table 3, however, show that there was only a significant effect of elicitation order for unsearchable knowledge questions. No other univariate parameters were significant in either experiment.

3.2.3 Credible intervals

We next analyzed the width of participants' credible intervals by subtracting the lower-bound estimate from the upper-bound estimate. We averaged the ranges within each question type. The three repeated measures were subjected to a MANOVA with elicitation order as a fixed factor. Table 4 shows the corresponding means and standard deviations, whereas Tables 5 and 6 summarize the multivariate effects and univariate parameter estimates from the models in each experiment, respectively. Neither the multivariate analysis nor the univariate parameter estimates were significant in either experiment.

3.2.4 Confidence judgments

We examined the effect of elicitation order on the confidence participants had that their credible intervals captured the correct value. We averaged participants' confidence ratings for

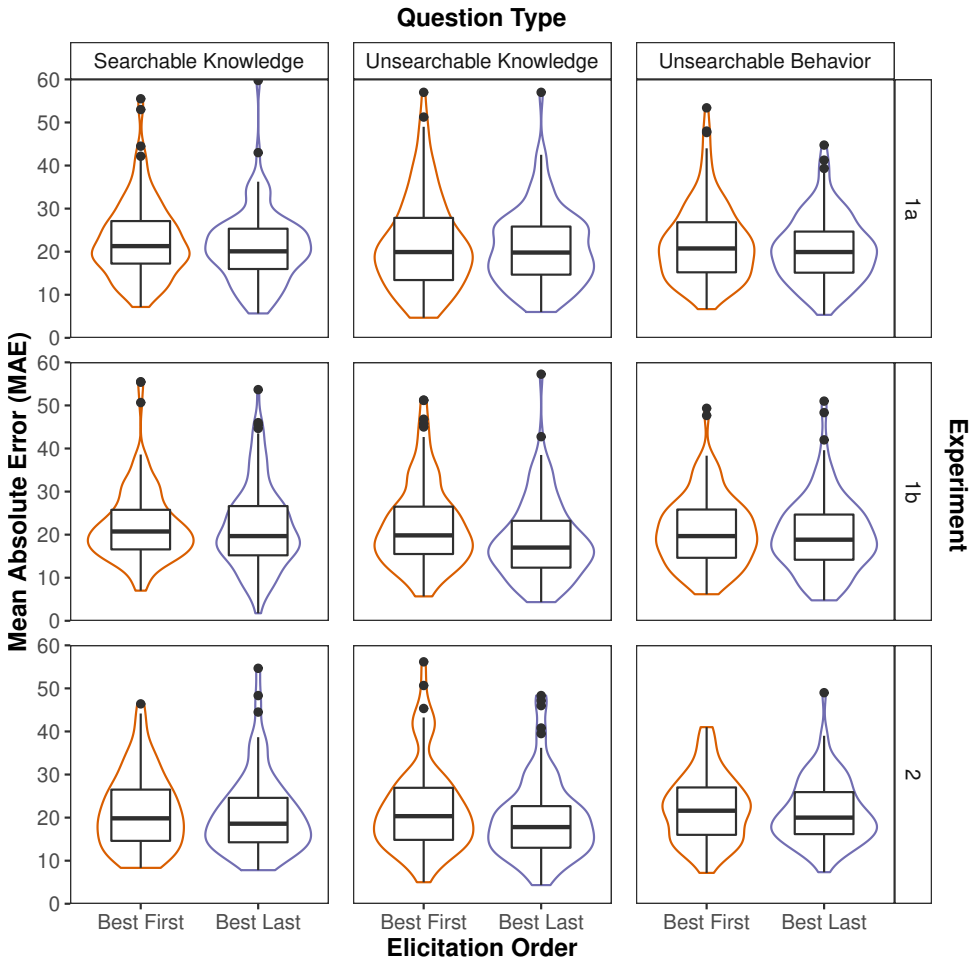

FIGURE 1: Distribution of MAE by experiment, question type and elicitation order.

TABLE 1: GMAE by experiment, question type and elicitation order.

		Question type							
		SK		UK		UB		Overall	
Exp.	Order	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1a	Best first	21.89	9.05	22.19	10.82	22.40	9.02	22.14	7.31
1a	Best last	20.50	7.56	20.12	7.77	20.63	7.57	20.50	4.95
1b	Best first	21.32	7.97	22.42	9.77	21.01	8.24	21.58	6.30
1b	Best last	20.75	9.54	18.60	9.15	19.67	7.50	19.66	6.10
2	Best first	21.14	8.74	21.45	9.74	22.02	7.65	21.54	5.41
2	Best last	19.56	7.85	19.90	9.61	20.71	6.21	20.07	5.72

Note: S = searchable; U = unsearchable; K = knowledge; B = behavior.

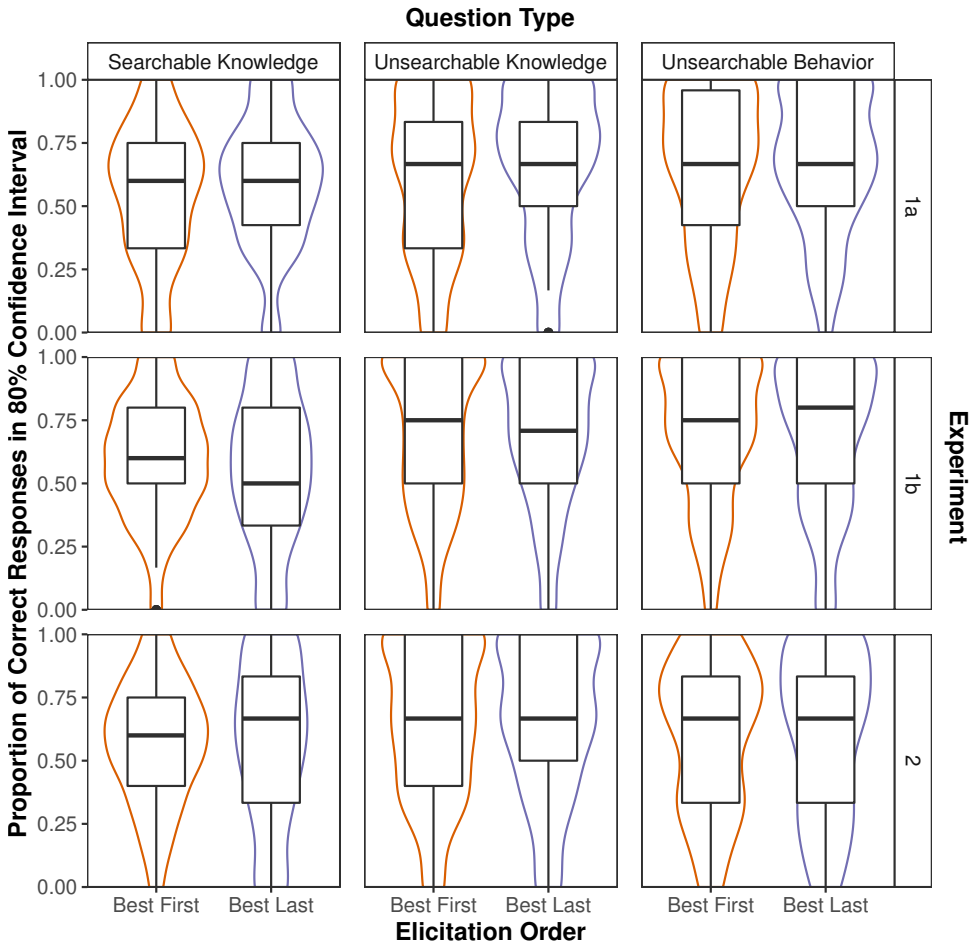

FIGURE 2: Distribution of proportion of correct responses in 80% confidence intervals by experiment, question type and elicitation order.

TABLE 2: Multivariate effects of elicitation order on GMAE by experiment.

Exp.	Effect	F	df	p	η_p^2
1a	Intercept	683.81	3, 175	<.001	.921
1a	Order	1.15	3, 175	.331	.019
1b	Intercept	793.11	3, 213	<.001	.918
1b	Order	3.01	3, 213	.031	.041
2	Intercept	805.53	3, 150	<.001	.942
2	Order	0.94	3, 150	.423	.018

each question type and subjected the three repeated measures to a MANOVA with elicitation order as a fixed factor. Table 7 shows the corresponding means and standard deviations. There was no effect of elicitation order in either experiment (both $p \geq .80$).

3.2.5 Calibrated confidence intervals

Our final analysis examined the accuracy of participants' confidence intervals when all were calibrated to a fixed confidence level of 80%. We constructed the lower and upper bound of the calibrated intervals using the following formulas (Hemming et al., 2018):

$$\text{Lower Bound: } B - [(B - L) \times (S/C)],$$

$$\text{Upper Bound: } B + [(U - B) \times (S/C)],$$

where B is the best estimate, L is the lower bound, U is the upper bound, C is the reported confidence, and S is the calibrated interval range. For each of the three question types, we computed for each participant the proportion of standardized confidence intervals that captured the correct answer. If the correct answer fell outside the interval, then it was scored as incorrect. Figure 2 shows the distributions of the proportion of correct judgments by experiment, question type, and elicitation order, whereas Table 8 shows the corresponding

TABLE 3: Univariate parameter estimates from MANOVA predicting GMAE by experiment.

Exp.	Type	Parameter	*B*	*SE*	*t*	*p*	η_p^2
1a	SK	Intercept	20.50	0.93	22.10	<.001	.734
1a	SK	Order	1.39	1.26	1.10	.271	.007
1a	UK	Intercept	20.12	1.05	19.08	<.001	.673
1a	UK	Order	2.08	1.43	1.45	.148	.012
1a	UB	Intercept	20.63	0.93	22.26	<.001	.737
1a	UB	Order	1.77	1.26	1.41	.160	.011
1b	SK	Intercept	20.75	0.86	24.06	<.001	.729
1b	SK	Order	0.57	1.19	0.48	.631	.001
1b	UK	Intercept	18.60	0.93	19.91	<.001	.648
1b	UK	Order	3.82	1.29	2.96	.003	.039
1b	UB	Intercept	19.67	0.78	25.28	<.001	.748
1b	UB	Order	1.34	1.07	1.25	.212	.007
2	SK	Intercept	19.56	0.97	20.22	<.001	.729
2	SK	Order	1.58	1.34	1.18	.242	.009
2	UK	Intercept	19.90	1.12	17.69	<.001	.673
2	UK	Order	1.55	1.56	0.99	.324	.006
2	UB	Intercept	20.71	0.81	25.48	<.001	.810
2	UB	Order	1.31	1.13	1.16	.247	.009

Note: Order = best first; S = searchable; U = unsearchable; K = knowledge; B = behavior.

means and standard deviations. The resulting three repeated measures were analyzed in a MANOVA with elicitation order as a fixed factor. Neither the multivariate analysis nor the univariate parameter estimates were significant in either experiment ($p \geq .25$).

It is evident from descriptive results in Figure 2 and Table 8 that participants were overprecise in both experiments, with their accuracy rates falling far short of 80% accuracy. We confirmed this by running one-sample *t*-tests against a test value of .8: In Experiment 1a, the grand mean accuracy rate was .65 ($SD = .21$, $t(178) = -9.54$, $p < .001$, $d = 0.71$); in Experiment 1b, the rate was .67 ($SD = .21$, $t(216) = -8.96$, $p < .001$, $d = 0.62$).

3.3 Discussion

The results of Experiments 1a and 1b were quite consistent. In Experiment 1a, none of the univariate tests of order were significant, and in Experiment 1b, only one (unsearchable knowledge) was significant. Moreover, consistent with Soll and Klayman (2004, Experiment 3), in each experiment, elicitation order had no significant effect on the accuracy of their calibrated confidence intervals. Nor did elicitation order have an effect on the range of credible intervals or confidence judgments in the range. The hypothesis we tested based on insights from work on anchoring-and-adjustment processes (Epley & Gilovich, 2006; Tversky & Kahneman, 1974) were unsupported in the present experiments. As well, our findings did not generalize Block and Harper's (1991) result that generating and writing down best estimates first improves calibration. It was evident that the inaccuracy of participants' confidence intervals was expressed in both experiments in the form of overprecision, with the deviations from perfect calibration of a medium to large effect size in both experiments. Therefore, it is not simply the case that there is little or no miscalibration to correct in the first place.

While Experiments 1a and 1b appear to provide clear evidence that the effect of the prior elicitation of confidence intervals on the accuracy of best estimates is minimal, we wanted to provide an additional test to put these results on even firmer footing. Thus, in Experiment 2 we set out to replicate Experiments 1a and 1b. In an attempt to improve participants' use of confidence interval construction we moved the critical estimation task to the beginning of the battery of tasks that the participants completed. In addition, we added additional instructions to remind participants about how to construct the intervals coherently. Lastly, we adopted a more stringent attention check (Oppenheimer, Meyvis & Davidenko, 2009) and placed it immediately after the estimation task. Each of these modifications was geared toward increasing participants' willingness and/or ability to use confidence interval construction and provide us with an additional, independent means of selecting those individuals who were more likely doing so.

4 Experiment 2

4.1 Methods

4.1.1 Participants

Five hundred and forty-five participants completed our study online via Qualtrics Panels. The sample characteristics were identical to Experiments 1a and 1b (i.e., between 18 and 60 years of age, English first language, US and/or Canadian citizen). After exclusions based on the same criteria used in Experiments 1a and 1b we retained a sample of 198 (mean age = 43.71; 124 females, 73 males, and 1 missing response; 112 Canadian citizens, 81 US, and 5 dual). In Experiment 2, 64, 286, and 74 participants were excluded for demographic reasons, failure of the attention check, and missing or inappropriate responses, respectively. Participants were compensated in the same manner as the prior experiments.

TABLE 4: Range of credible interval by experiment, question type and elicitation order.

		Question type							
		SK		UK		UB		Overall	
Exp.	Order	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1a	Best first	35.55	17.57	45.22	18.99	45.26	20.42	41.97	17.37
1a	Best last	36.17	17.72	46.56	18.22	45.14	18.34	42.72	16.34
1b	Best first	35.96	16.36	47.22	17.37	47.32	18.05	43.55	16.08
1b	Best last	33.83	18.61	45.36	18.35	44.92	18.43	41.43	16.30
2	Best first	32.49	16.16	42.57	16.68	40.63	16.94	38.60	15.37
2	Best last	35.88	22.60	45.56	21.75	44.25	22.03	41.94	20.93

Note: S = searchable; U = unsearchable; K = knowledge; B = behavior.

TABLE 5: Multivariate effects of elicitation order on credible interval ranges by experiment.

Exp.	Effect	*F*	*df*	*P*	η_p^2
1a	Intercept	376.24	3, 175	<.001	.866
1a	Order	0.25	3, 175	.860	.004
1b	Intercept	529.44	3, 213	<.001	.882
1b	Order	0.33	3, 213	.804	.005
2	Intercept	276.83	3, 150	<.001	.847
2	Order	0.45	3, 150	.716	.009

TABLE 6: Univariate parameter estimates from MANOVA predicting credible interval ranges by experiment.

Exp.	Type	Parameter	*B*	*SE*	*t*	*P*	η_p^2
1a	SK	Intercept	36.17	1.95	18.57	<.001	.661
1a	SK	Order	-0.62	2.65	-0.23	.816	.000
1a	UK	Intercept	46.56	2.06	22.62	<.001	.743
1a	UK	Order	-1.34	2.80	-0.48	.632	.001
1a	UB	Intercept	45.14	2.15	20.97	<.001	.713
1a	UB	Order	0.11	2.92	0.04	.969	.000
1b	SK	Intercept	33.83	1.72	19.66	<.001	.643
1b	SK	Order	2.13	2.37	0.90	.372	.004
1b	UK	Intercept	45.36	1.76	25.81	<.001	.756
1b	UK	Order	1.86	2.43	0.77	.444	.003
1b	UB	Intercept	44.92	1.80	25.01	<.001	.744
1b	UB	Order	2.40	2.48	0.97	.334	.004
2	SK	Intercept	35.88	2.27	15.81	<.001	.622
2	SK	Order	-3.39	3.15	-1.08	.284	.008
2	UK	Intercept	45.56	2.24	20.32	<.001	.731
2	UK	Order	-2.99	3.11	-0.96	.339	.006
2	UB	Intercept	44.25	2.27	19.47	<.001	.714
2	UB	Order	-3.63	3.15	-1.15	.252	.009

Note: Order = best first; S = searchable; U = unsearchable; K = knowledge; B = behavior.

4.1.2 Materials and procedures

The materials and procedures were identical to Experiments 1a and 1b, with the following exceptions. As noted previously, we moved the estimation task to the beginning of the survey rather than having it randomly intermixed. As well, we added the following reminder on the pages on which individuals constructed their intervals: "Remember your LOWEST plausible value should be LOWER than your BEST response and HIGHEST plausible value" when providing the lowest plausible value, and "Remember your HIGHEST plausible value should be HIGHER than your BEST response and your LOWEST plausible value" when providing the highest plausible value. We also removed the questions used to calculate the true response for unsearchable items. Rather, baseline behaviors/judgments were determined based on the responses provided by participants in Experiments 1a and 1b. Lastly, we replaced the second attention check item (i.e., the value of a quarter is what percentage of a dollar?) based on low performance in Experiments 1a and 1b. In its place, we included an "Instructional Manipulation Check" (adapted from Oppenheimer et al., 2009) whereby under the cover of a question about sports participation, participants were simply instructed to ignore the main question and click a button to proceed to the next screen. As in Experiments 1a and 1b, participants also completed a number of other tasks/scales (now all of which were administered after the estimation task). There were also minor changes to these other tasks/scales. There were now eight Bias Blindspot items (taken from Scopelliti et al., 2015 and West, Meserve

TABLE 7: Confidence in credible interval by experiment, question type and elicitation order.

		Question type							
		SK		UK		UB		Overall	
Exp.	Order	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1a	Best first	73.29	11.00	74.76	10.82	75.11	11.40	74.42	10.40
1a	Best last	71.19	10.41	73.03	10.38	74.04	11.12	72.78	9.88
1b	Best first	75.33	11.92	76.51	10.93	75.52	11.05	75.82	10.48
1b	Best last	74.72	9.96	75.45	10.21	75.45	8.52	75.14	8.54
2	Best first	73.01	11.64	73.74	10.84	73.83	10.65	73.49	10.22
2	Best last	71.73	12.34	73.57	12.08	72.94	11.56	72.77	11.40

Note: S = searchable; U = unsearchable; K = knowledge; B = behavior.

TABLE 8: Proportion of correct responses in 80% confidence intervals by experiment, question type and elicitation order.

		Question type							
		SK		UK		UB		Overall	
Exp	Order	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
1a	Best first	.576	.261	.631	.310	.670	.307	.624	.227
1a	Best last	.628	.233	.710	.263	.693	.275	.676	.195
1b	Best first	.606	.232	.699	.281	.720	.270	.675	.202
1b	Best last	.576	.272	.717	.276	.718	.276	.671	.216
2	Best first	.566	.251	.655	.305	.594	.298	.605	.225
2	Best last	.607	.307	.670	.304	.647	.297	.643	.258

Note: S = searchable; U = unsearchable; K = knowledge; B = behavior.

& Stanovich, 2012), plus six general heuristics and biases problems, adapted from a variety of sources (assessing Anchoring, Base-rate neglect, Conjunction Fallacy, and Outcome Bias). We also removed the Actively Open-Minded Thinking Scale. These latter tasks are the focus of another investigation and are not presented here.

4.2 Results

4.2.1 Data quality

In Experiment 2, 89.05% (*SD* = 31.92%) of participants' sets of estimates were complete and coherent. This was significantly greater than the pooled coherence rate from Experiment 1a and 1b (*M* = 74.7%, *SD* = 31.92%, $t(350.22) = -5.42$, $p < .001$, $d = 0.43$). Therefore, it appears the modifications to Experiment 2 had the intended effect of improving the coherence of participants' judgments. The complete and coherent requirement excluded an additional 44, leaving a final sample of 154. As in the prior experiments, coherence was significantly related to ICAR ($r_{pb}[198] = .34$, $p < .001$).

As in the earlier experiments, participants' accuracy was significantly above chance for each question type in Lists A and B, all $p < .001$. The full analysis is reported in Appendix F in the online supplementary materials.

The final data quality check revealed that ICAR (*M* = 3.92, *SD* = 1.92) correlated with GMAE ($r[152] = -.279$, $p < .001$); participants who performed well on ICAR tended to be more accurate.

4.2.2 Best estimates

Figure 1 and Table 1 show the descriptive results. We conducted a MANOVA with elicitation order as a fixed factor and the three MAE measures corresponding to question type as dependent measures. Consistent with Experiment 1a, the effect of elicitation order was not significant in either the multivariate analysis or any of the univariate parameter estimates (see Tables 2 and 3).

4.2.3 Credible intervals

We calculated the range of the credible intervals and averaged them within question type (see Table 4 for descriptive results). We then computed a MANOVA on the repeated measures with elicitation order as a fixed factor. Consistent with the earlier experiments, there was no effect of elicitation order in either the multivariate analysis or any of the univariate parameter estimates (see Tables 5 and 6).

4.2.4 Confidence judgments

As in the prior experiments, we averaged participants' confidence ratings for each question type and subjected the three repeated measures to a MANOVA with elicitation order as a fixed factor (see Table 7 for descriptive results). Consistent with the earlier experiments, there was no significant effect of elicitation order ($p = .76$).

4.2.5 Calibrated confidence intervals

We examined the accuracy of participants' confidence intervals when all were calibrated to a fixed confidence level of 80% (Hemming et al., 2018), using the same proportion-correct metric used in Experiments 1a and 1b (see Figure 2 and Table 8 for descriptive results). The three repeated measures were first analyzed in a MANOVA with elicitation order as a fixed factor. The effect of elicitation order was not significant (p = .66). Finally, as in the earlier experiments, participants were overprecise. Their grand mean accuracy rate was .62 (SD = .24), significantly lower than the .8 criterion required for perfect calibration (t[153] = −9.09, p < .001, d = 0.75).

4.3 Discussion

In spite of the changes in procedure to improve participants' focus on the central task and to estimate judgments coherently, the results of Experiment 2 are highly consistent with those of Experiments 1a and 1b. Elicitation order did not have a significant effect on the accuracy of best estimates or the calibration of confidence intervals. As well, consistent with the earlier experiments, participants showed a substantial degree of overprecision across the judgment tasks.

5 General discussion

In the present investigation, we examined the effect of prior confidence interval construction on the accuracy of best estimates as well as the effect of best estimate construction on the calibration of confidence intervals. Previous research has provided support for the potential effectiveness of interval construction where separate lower and upper bounds are constructed and where confidence levels are assigned to credible intervals by assessors rather than by experimental fiat (Soll & Klayman, 2004; Teigen & Jørgenson, 2005). Various IBBE protocols such as the four-step method (Speirs-Bridge et al., 2010), implement these attributes, but further specify that elicitation order is relevant and prescribes that credible intervals be constructed prior to the elicitation of best estimates. The potential benefits of confidence interval construction are often explained in terms of the beneficial influence of taking multiple samples from memory and/or taking multiple perspectives on a given judgment (i.e., similar to various consider-the-opposite approaches; Herzog & Hertwig, 2009; 2014; Hirt & Markman, 1995; Koriat, Lichtenstein & Fischhoff, 1980; Lord et al., 1984; Williams & Mandel, 2007). That is, by generating intervals before best estimates, assessors might be prompted to consider a wider range of relevant evidence that might, in turn, improve the accuracy of the best estimates. As well, by generating intervals before best estimates, assessors might escape the biasing effect that the best estimates might have if assessors are prone to anchor on them and then insufficiently adjust (Epley & Gilovich, 2006; Tversky & Kahneman, 1974). Therefore, IBBE protocols can be viewed as a debiasing method for judgment, one that addresses Lilienfeld et al.'s (2009) call for research on correcting errors in judgment.

Overall, our results painted an uninspiring picture of the effectiveness of prior confidence interval construction on the accuracy of assessors' best estimates. In two of the three experiments, elicitation order had no significant effect on accuracy, and in the one experiment (1b) where there was a multivariate effect, that effect was due to only one significant univariate effect. Fully 8 out of 9 univariate tests of the effect of order on best estimate accuracy failed to find a significant effect. The null effect of elicitation order was even more stable for accuracy of calibrated confidence intervals, where not one univariate parameter estimate (out of 9 tests across the 3 experiments) was significant. As well, in Experiments 1a and 1b, we included participants who failed one of our attention checks because we thought that item might have been misunderstood by many, given the high error rate we observed. In Experiment 2, however, we also observed a high error rate on an attention check that has been used in other studies (Oppenheimer et al., 2009). The probable net effect is that we were much more liberal in our inclusion criteria in Experiments 1a and 1b than we were in Experiment 2, and yet we obtained highly consistent results.

Furthermore, recall that confidence interval construction is hypothesized to benefit the accuracy of best estimates by improving the recruitment of relevant evidence pertinent to testing the equivalent of best- and worst-case scenarios or multiple viewpoints (Hemming et al., 2018). Presumably, the benefit afforded to best-estimate accuracy depends on how accurately the preceding intervals are constructed. Following this line of reasoning, one might also expect the correlation between best-estimate accuracy and (calibrated) confidence interval accuracy to be stronger if interval construction preceded best-estimate construction than if the best estimates were constructed first. However, we do not find support for that prediction either. Across experiments and question types, the correlation between GMAE for the best estimates and the proportion of correct responses in the calibrated confidence interval was $r(289) = -.40$ ($p < .001$) when best estimates were elicited first and $r(257) = -.34$ ($p < .001$) when they were elicited after the confidence intervals were constructed. The difference is not significant ($z = -0.79$, p = .21).

From a practical perspective, the present results do not indicate the utility of prior confidence interval construction for improving the accuracy of best estimates. As we noted in the Introduction, improving the accuracy of judgments represents a major effort that has important implications for several domains. An important consideration in these efforts is cost. IBBE protocols require significant additional time (e.g.., in the case of the four-step method, three additional

judgments). Thus, even a small benefit may not justify the added effort, given that other methods that require a similar number of elicitations have yielded large improvements in probability judgment accuracy. Notwithstanding the risks associated with internal meta-analyses (Ueno, Fastrich & Murayama, 2016; Vosgerau, Simonsohn, Nelson & Simmons, 2019), it is useful to estimate the overall effect size of the elicitation order manipulation we conducted across three experiments. There is a small positive effect (Cohen's $d = 0.285$, 95% CI [0.117, 0.453]) of the modified four-step method we tested on the accuracy of best estimates.[1] Moreover, in some elicitation contexts, such as decision analysis (Clemen, 1996; von Winterfeldt & Edwards, 1986), it may be highly desirable, if not necessary, to collect lower- and upper-bound estimates, in which case there may be a small benefit to following the ordering prescribed by IBBE protocols. However, if the aim of the method is to improve best estimates, then query intensive IBBE protocols do not compare favorably with alternative methods for improving judgment accuracy that require similar increases in elicitation.

For instance, several studies have found that by eliciting a small set of logically-related judgments (typically 3–4 items per topic), accuracy can be substantially improved by recalibrating them using *coherentization* methods that constrain the estimates to respect certain logical criteria, such as the additivity and unitarity properties in probability calculus (e.g., Fan, Budescu, Mandel & Himmelstein, 2019; Karvetski et al., 2013). For example, Mandel et al. (2018) found a large (i.e., $d = 0.96$) improvement in accuracy on a probability judgment task after four related probability judgments were coherentized. Moreover, individual differences in the degree of incoherence have been effectively used in these studies and others to improve aggregation through performance weighting — namely, by giving more weight to assessors who are more coherent (e.g., Karvetski, Mandel & Irwin, 2020; Predd, Osherson, Kulkarni & Poor, 2008; Wang, Kulkarni, Poor & Osherson, 2011). Other techniques such as use of conditional rather than direct probability assessments (Kleinmuntz, Fennema & Peecher, 1996), using ratio rather than direct probability assessments (Por & Budescu, 2017), using contrastive evaluation frames that make complements explicit (Williams & Mandel, 2007) have been shown to improve judgment accuracy, whereas other methods such as eliciting probability estimates for ranges over entire distributions (Haran, Moore & Morewedge, 2010; Moore, 2019) or iteratively adjusting interval sizes until a pre-specified confidence level is matched by an assessor's subjective probability that the interval captures the true value (Winman, Hansson & Juslin, 2004) have shown promise for reducing overprecision.

[1]Estimation of the confidence interval on d was computed using the implementation of procedures by Smithson (2001) provided by Wuensch (2012).

That said, the present research solicited estimates to general-knowledge and behavior-related questions in a percentage format, and thus the possibility remains that there are contexts wherein this particular form of elicitation generates larger (and more justifiable) gains. For instance, our problems might not have been ideal for recruiting "for vs. against" evidence that would bear on the best estimate. Moreover, one might question whether the unsearchable items were effective for our research purposes. We believe they were for at least two reasons. First, we did not observe much difference between accuracy levels for the three types of questions (see Table 1 and Figure 1) and all three question types were answered with accuracy levels significantly above chance levels. Second, we did not find that order had an effect on the commonly employed general-knowledge items. Indeed, the only significant effect of order we observed was on the unsearchable knowledge items.

Thus, an important contribution of the present work is to introduce measured skepticism about expecting a *general* gain in terms of judgment accuracy from prior confidence interval construction. Future work aimed at locating contexts wherein or conditions under which such an elicitation method is beneficial would be valuable. For example, unlike the present research in which participants were compensated equally regardless of performance level, researchers could investigate whether incentivized conditions moderate the effect of confidence interval construction. It is possible that with performance-based incentives, the beneficial effect of confidence interval construction would be more pronounced. As well, research could examine the effect of instructions accompanying the elicitation of estimates. Perhaps order would have more of an effect if the instructions more strongly encouraged dialectical thinking (Herzog & Hertwig, 2009). Finally, experts and novices display different patterns of response in tasks involving interval construction (e.g., McKenzie, Liersch & Yaniv, 2008). The performance benefits that Speirs-Bridge et al. (2010) reported for the four-step method over a three-step variant that omitted the judgment of confidence level were observed in expert samples. Although Speirs-Bridge et al. did not compare these elicitations to a control condition in which neither intervals nor confidence levels was elicited, it may be that the medium effect size they observed across their studies is attributable in part to the expert samples employed. Although we did not use an expert sample, we took care to rule out (at a substantial cost to our sample sizes across the three experiments) participants who made blatantly incoherent responses, and we observed that performance in the resulting samples was correlated with intelligence. Nevertheless, it would be useful in future research to conduct similar tests with expert samples.

Finally, it is worth noting that the average confidence level that participants assigned to their credible intervals was remarkably stable across question types and experiments, ranging from 71% to 77%. Recall that Budescu and Du (2007)

found that participants directed to construct 70% confidence intervals were better calibrated than those required to construct either 50% or 90% confidence intervals. Although participants may have chosen confidence levels that offer relatively good prospects for calibration, evidently in the present research this tendency did not buffer them from overprecision, which they exhibited in moderate to large degree.

6 Conclusion

The present investigation provided a strong test of the effect of IBBE methods that require the prior construction of credible intervals on judgment accuracy. There was weak evidence that eliciting confidence intervals prior to best estimates increases the accuracy of those estimates, at least with respect to the types of judgment we evaluated and the type of sample we recruited. Taken together, these findings call for greater skepticism regarding the effectiveness of interval construction as an elicitation method for improving judgment accuracy. By the same token, we found no evidence that generating best estimates before confidence intervals improves the calibration of the intervals, as Block and Harper (1991) reported. Nor did we find support for the contrary anchoring-and-adjustment hypothesis (Epley & Gilovich, 2006; Tversky & Kahneman, 1974), which predicted that generating prior best estimates would, if anything, aggravate overprecision. Rather, in line with Soll and Klayman (2004), we found meager evidence that the order in which best estimates and confidence intervals are elicited matters much to accuracy and calibration. However, the generalizability of this finding should be tested in future research.

References

Adams-Hosking, C., McBride, M.F., Baxter, G., Burgman, M., de Villiers, D., Kavanagh, R., et al. (2016). Use of expert knowledge to elicit population trends for the koala (Phascolarctos cinereus). *Diversity and Distributions*, *22*, 249–262.

Alpert, M., & Raiffa, H. (1982). A progress report on the training of probability assessors. In D. Kahneman, P. Slovic, & A. Tversky (Eds.), *Judgment under uncertainty: Heuristics and biases* (pp. 294–305) Cambridge, MA: Cambridge University Press.

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*, 265–284.

Berner, E. S., & Graber, M. L. (2008). Overconfidence as a cause of diagnostic error in medicine. *American Journal of Medicine*, *121*(5, Suppl), S2–S23.

Block, R. A., & Harper, D. R. (1991). Overconfidence in estimation: Testing the anchoring-and-adjustment hypothesis. *Organizational Behavior and Human Decision Processes, 49*, 188–207.

Budescu, D. V., & Du, N. (2007). Coherence and consistency of investors' probability judgments. *Management Science, 53*, 1731–1744.

Burgman, M. A., McBride, M., Ashton, R., Speirs-Bridge, A., Flander, L., Wintle, B., et al. (2011). Expert status and performance. *PLoS ONE 6*(7): e22998. https://doi.org/10.1371/journal.pone.0022998.

Clemen, R. T. (1996). *Making hard decisions: An introduction to decision analysis* (2nd ed.). Boston, MA: PWS-Kent.

Condon, D. M., & Revelle, W. (2014). The international cognitive ability resource: development and initial validation of a public-domain measure. *Intelligence, 43*, 52–64.

Dawes, R. M. (1979). The robust beauty of improper linear models in decision making. *American Psychologist, 34*, 571–582.

Dawson, N. V., & Arkes, H. R. (1987). Systematic errors in medical decision making: Judgment limitations. *Journal of General Internal Medicine, 2,*183–187.

Dhami, M. K., Mandel, D. R., Mellers, B. A., & Tetlock, P. E. (2015). Improving intelligence analysis with decision science. *Perspectives on Psychological Science, 10*, 753–757.

Epley, N., & Gilovich, T. (2006). The anchoring-and-adjustment heuristic: Why the adjustments are insufficient. *Psychological Science, 17,* 311–318.

Fan, Y., Budescu, D. V., Mandel, D. R., & Himmelstein, M. (2019). Improving accuracy by coherence weighting of direct and ratio probability judgments. *Decision Analysis*, *16*, 197–217.

Fingar, T. (2011). *Reducing uncertainty: Intelligence analysis and national security*. Stanford: Stanford Security Studies.

Friedman, J. A. (2019). *War and chance: Assessing uncertainty in international politics*. New York: Oxford University Press.

Gaertig, C., & Simmons, J. P. (2019, November). Does dialectical bootstrapping improve the wisdom of the inner crowd [Paper presentation]? *40th Annual Conference of the Society for Judgment and Decision Making*, Montreal, QC.

Goodman-Delahunty, J., Granhag, P. A., Hartwig, M., & Loftus, E. F. (2010). Insightful or wishful: Lawyers' ability to predict case outcomes. *Psychology, Public Policy, and Law, 16,* 133–157.

Hanea, A., McBride, M., Burgman, M., Wintle, B., Fidler, F., Flander, L., et al. (2017). I nvestigate D iscuss E stimate A ggregate for structured expert judgement. *International Journal of Forecasting*, *33*, 267–279.

Haran, U., Moore, D. A., & Morewedge, C. K. (2010). A simple remedy for overprecision in judgment. *Judgment and Decision Making, 5*, 467–476.

Hemming, V., Walshe, T. V., Hanea, A. M., Fidler, F., & Burgman, M. A. (2018). Eliciting improved quantitative judgements using the IDEA protocol: A case study in natural resource management. *PloS One*, *13*(6), e0198468. https://doi.org/10.1371/journal.pone.0198468

Herzog, S. M., & Hertwig, R. (2009). The wisdom of many in one mind: Improving individual judgments with dialectical bootstrapping. *Psychological Science*, *20*, 231–237.

Herzog, S. M., & Hertwig, R. (2014). Harnessing the wisdom of the inner crowd. *Trends in Cognitive Sciences*, *18*, 504–506.

Hirt, E. R., & Markman, K. D. (1995). Multiple explanation: A consider-an-alternative strategy for debiasing judgments. *Journal of Personality and Social Psychology, 69*, 1069–1086.

Karvetski, C. W., Mandel, D. R., & Irwin, D. (2020). Improving probability judgment in intelligence analysis: From structured analysis to statistical aggregation. *Risk Analysis, 40,* 1040–1057.

Karvetski, C. W., Olson, K. C., Mandel, D. R., & Twardy, C. R. (2013). Probabilistic coherence weighting for optimizing expert forecasts. *Decision Analysis, 10*, 305–326.

Kleinmuntz, D. N., Fennema, M. G., & Peecher, M. E. (1996). Conditioned assessment of subjective probabilities: Identifying the benefits of decomposition. *Organizational Behavior and Human Decision Processes, 66*, 1–15.

Koriat, A., Lichtenstein, S., & Fischhoff, B. (1980). Reasons for confidence. *Journal of Experimental Psychology: Human Learning and Memory, 6*, 107–118.

Lilienfeld, S., Ammirati, R., & Landfield, K. (2009). Giving debiasing away: Can psychological research on correcting cognitive errors promote human welfare. *Perspectives on Psychological Science, 4,* 390–398.

Lord, C. G., Lepper, M. R., & Preston, E. (1984). Considering the opposite: A corrective strategy for social judgment. *Journal of Personality and Social Psychology, 47*, 1231–1243.

Mandel, D. R. (2019). Can decision science improve intelligence analysis? In S. Coulthart, M. Landon-Murray, & D. Van Puyvelde (Eds.), *Researching national security intelligence: Multidisciplinary approaches* (pp. 117–140). Washington, DC: Georgetown University Press.

Mandel, D. R., & Barnes, A. (2018). Geopolitical forecasting skill in strategic intelligence. *Journal of Behavioral Decision Making, 31*, 127–137.

Mandel, D. R., Karvetski, C. W., & Dhami, M. K. (2018). Boosting intelligence analysts' judgment accuracy: What works, what fails? *Judgment and Decision Making*, *13*, 607–621.

McKenzie, C. R. M., Liersch, M. J., & Yaniv, I. (2008). Overconfidence in interval estimates: What does expertise buy you? *Organizational Behavior and Human Decision Processes*, *107*, 179–191.

Moore, D. A. (2019). *Overprecision is a property of thinking systems*. Preprint available from https://psyarxiv.com/fxswm/.

Moore, D. A., & Healy, P. J. (2008). The trouble with overconfidence. *Psychological Review, 115*, 502–517.

Morgan, M. G., & Henrion, M. (1990). Uncertainty: A guide to dealing with uncertainty in quantitative risk and policy analysis. New York, NY: Cambridge University Press.

Müller-Trede, J. (2011). Repeated judgment sampling: Boundaries. *Judgment and Decision Making, 6*, 283–294.

Murphy, A .H., & Winkler, R. L. (1970). Scoring rules in probability assessment and evaluation. *Acta Psychologica*, *34*, 273–286.

O'Hagan, A. (2019). Expert knowledge elicitation: Subjective but scientific. *The American Statistician, 73,* 69–81.

Önkal, D., Yates, J. F., & Simga-Mugan, C., & Öztin, S. (2003). Professional vs. amateur judgment accuracy: The case of foreign exchange rates. *Organizational Behavior and Human Decision Processes, 91*, 169–185.

Oppenheimer, D. M., Meyvis, T., & Davidenko, N. (2009). Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of Experimental Social Psychology, 45*, 867–872.

Oskamp, S. (1965). Overconfidence in case-study judgments. *Journal of Consulting Psychology, 29*, 261–265.

Pitz, G. F. (1974). Subjective probability distributions for imperfectly known quantities. In L. W. Gregg (Ed.), *Knowledge and cognition* (pp. 29–41). Potomac, MD: Erlbaum.

Por, H., & Budescu, D. V. (2017). Eliciting subjective probabilities through pair-wise comparisons. *Journal of Behavioral Decision Making*, *30*, 181–196.

Predd, J. B., Osherson, D. N., Kulkarni, S. R., & Poor, H. V. (2008). Aggregating probabilistic forecasts from incoherent and abstaining experts. *Decision Analysis, 5*, 177–189.

Scopelliti, I., Morewedge, C. K., McCormick, E., Min, H. L., Lebrecht, S., & Kassam, K. S. (2015). Bias blind spot: Structure, measurement, and consequences. *Management Science, 61*, 2468–2486

Smithson, M. (2001). Correct confidence intervals for various regression effect sizes and parameters: The importance of noncentral distributions in computing intervals. *Educational And Psychological Measurement, 61*, 605–632.

Soll, J.B., & Klayman, J. (2004). Overconfidence in interval estimates. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 299–314.

Speirs-Bridge, A., Fidler, F., McBride, M., Flander, L., Cumming, G., & Burgman, M. (2010). Reducing overconfidence in the interval judgments of experts. *Risk Analysis, 30*, 512–523.

Teigen, K. H., & Jørgensen, M. (2005). When 90% confidence intervals are 50% certain: on the credibility of credible intervals. *Applied Cognitive Psychology, 19*, 455–475.

Tetlock, P. E. (2005). *Expert political judgment: How good is it? How can we know?* Princeton, NJ: Princeton University Press.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124–1131.

Ueno, T., Fastrich, G. M., & Murayama, K. (2016). Meta-analysis to integrate effect sizes within an article: Possible misuse and Type I error inflation. *Journal of Experimental Psychology: General, 145*, 643–654.

von Winterfeldt, D., & Edwards, W. (1986). *Decision analysis and behavioral research*. Cambridge, UK: Cambridge University Press.

Vosgerau, J., Simonsohn, U., Nelson, L. D., & Simmons, J. P. (2019). 99% impossible: A valid, or falsifiable, internal meta-analysis. *Journal of Experimental Psychology: General, 148,* 1628–1639.

Wang, G., Kulkarni, S. R., Poor, H. V., & Osherson, D. N. (2011). Aggregating large sets of probabilistic forecasts by weighted coherent adjustment. *Decision Analysis*, *8*, 128–144.

West, R. F., Meserve, R. J., & Stanovich, K. E. (2012). Cognitive sophistication does not attenuate the bias blind spot. *Journal of Personality and Social Psychology, 103*, 506–519.

Williams, J. J., & Mandel, D. R. (2007). Do evaluation frames improve the quality of conditional probability judgment? In D. S. McNamara & J. G. Trafton (Eds.), *Proceedings of the 29th Annual Meeting of the Cognitive Science Society* (pp. 1653–1658), Mahwah, NJ: Erlbaum.

Winman, A., Hansson, P., & Juslin, P. (2004). Subjective probability intervals: How to reduce overconfidence by interval evaluation. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 30*, 1167–1175.

Wuensch, K. L. (2012). *Using SPSS to obtain a confidence interval for Cohen's d.* http://core.ecu.edu/psyc/wuenschk/SPSS/CI-d-SPSS.pdf.

Yaniv, I., & Foster, D. P. (1995). Graininess of judgment under uncertainty: An accuracy-informativeness trade-off. *Journal of Experimental Psychology: General, 124,* 424–432.

Yaniv, I., & Foster, D. P. (1997). Precision and accuracy in judgmental estimation. *Journal of Behavioral Decision Making, 10,* 21–32.