Does Permitting Uncertain Estimates Help or Hinder the Wisdom of Crowds?

Zack Ellerby
Lab for Uncertainty in Data and Decision Making (LUCID)
School of Computer Science, University of Nottingham, UK
zack.ellerby@nottingham.ac.uk

Christian Wagner (Senior Member, IEEE)
Lab for Uncertainty in Data and Decision Making (LUCID)
School of Computer Science, University of Nottingham, UK
christian.wagner@nottingham.ac.uk

***Abstract*—This paper adds to a growing body of research into the practical utility of using interval-valued (IV) response modes to efficiently capture richer quantitative data from people—e.g., through surveys. Specifically, IV responses offer a cohesive method of capturing uncertainty, vagueness, or range associated with individual quantitative responses. In turn, IV data provide a strong foundation for subsequent fuzzy set based modelling—e.g., using the Interval Agreement Approach. The present paper focuses on the impact of soliciting IV estimates upon accuracy of group perceptual judgements—the 'Wisdom of the Crowd'. We report results from two empirical studies, examining the utility of IV data in the context of estimating specific (i.e., discrete point) ground truths, and directly comparing IV perceptual judgements (quantity estimates) against more traditional point estimates. There were two main hypotheses concerning the potential impacts of permitting uncertain (i.e., IV) estimates. First, it is possible that when specific predictions are required, permitting deliberately imprecise responses may reduce (prediction) accuracy versus forcing each respondent to provide their single 'best guess'. Second, that capturing the uncertainty associated with individual predictions should permit improved aggregation of group estimates, through weighting individual estimates according to their certainty. We report findings from two studies designed to investigate these hypotheses, and outline proposals for future research in this area.**

I. Introduction and Background

In comparison with conventional point response modes, e.g., Likert-type [1], or Visual Analogue [2], [3] scales, interval-valued (IV) estimates offer the capacity to capture additional information in each response, concerning individual response uncertainty, vagueness, or range [4]. Recent evidence indicates that an ellipse response mode (cf. Fig. 1) is able to do so efficiently, in terms of user experience (i.e., perceived workload and complexity) [5], [6]. These positive attributes have contributed to a recent surge in interest in IV response elicitation—exploring a variety of real-world applications ([7]–[10]), leading to development of open-source software [11], and motivating associated theoretical advances [12]–[17].

The benefits in accuracy gained from aggregating estimates from multiple estimators have been long established, in an effect often termed the 'Wisdom of the Crowd' [18], [19]. The basic principle is that, in the absence of significant bias, individual estimate errors tend to balance out. The present paper explores the impact of soliciting IV estimates on the accuracy of group aggregate perceptual judgements—weighing the potential informational advantages against potential drawbacks of permitting estimators to provide explicitly uncertain estimates. There were two main hypotheses concerning the potential impacts of permitting uncertain (i.e., IV) estimates. First, it is possible that when specific predictions are required, permitting deliberately imprecise responses may reduce estimate accuracy versus forcing each respondent to provide their single 'best guess'. Second, that capturing the uncertainty associated with individual predictions should permit improved aggregation of group estimates, through weighting individual estimates according to their certainty.

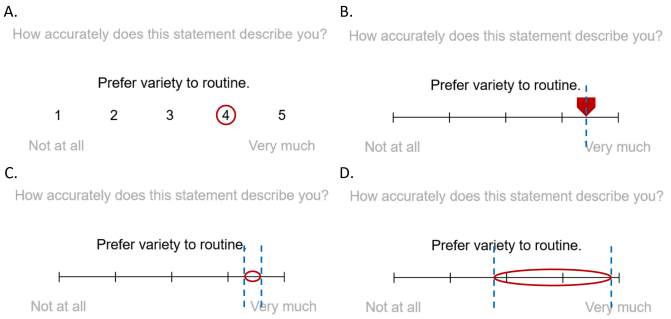

Fig. 1. Illustrative responses. A: Likert-type (ordinal). B: VAS-type. C: Ellipse (low uncertainty). D: Ellipse (high uncertainty).

In Section II we describe the participants, stimuli and procedures of two empirical studies—including details regarding data collection and analysis. In Section III we report descriptive and inferential results from each study. In Section IV we summarise key findings, discuss their implications, and describe planned future work.

II. Method

A. Study Participants

Study 1: Eighty participants completed the first study, recruited through opportunity sampling across three UK campuses of the University of Nottingham. Note that separate user feedback data obtained from this same experimental sample is reported in [5]. These were a mixture of academic and non-academic staff, as well as under- and post-graduate students. Participants volunteered approximately five minutes of their time to complete the study, in return for the option to enter a

This work was supported by UK Research and Innovation (UKRI) [grant numbers EP/P011918/1, EP/M02315X/1].

prize draw to win a jar of sweets (upon which they had made their judgements). Of these, 27 self-identified as female, 52 male and one declined to report their gender. Self-reported ages ranged from 17 to 57 (M=26.15, SD=10.08), though one participant declined to report their age. Fifty-five reported as native English speakers, and 25 as not.

Study 2: Twenty-two participants completed the second study, recruited through opportunity sampling of attendees at an interactive online workshop session held for 'Canberra AI Week' (Dec. 2020). These were a combination of under- and post-graduate students, academics, industry professionals, and the general public. Participants volunteered to take part during the session, with a study duration of approximately five minutes, in return for the chance to win a $100AUD prize. As this was a short pilot study, held during a time-sensitive workshop session, we did not collect participant demographic information.

B. Questions and Experimental Stimuli

Study 1: In this study, participants provided five perceptual estimates concerning a stimulus viewed in person. Each person was presented with a transparent plastic sweet jar, approximately half filled with coloured sweets (Bassett's Jelly Babies®). They were tasked with judging the overall weight of the jar, including contents, and the number of sweets that it contained of each of four colours: purple, yellow, green and orange. In fact, the jar weighed a total of 1669g, and contained 74 purple, 106 yellow, 26 green and 50 orange sweets. Responses were made on paper. The continuous scales upon which ratings were made ranged from [0, 4]kg, and from [0, 120] for each sweet colour.

Study 2: In this study, participants provided four perceptual estimates concerning a series of image stimuli, viewed on their own internet connected device. Each image was a photograph showing a group of penguins, and participants were tasked with reporting the number of penguins shown in each image—this was pitched as representing a 'citizen science' project, to monitor population size. In reality, the images contained 51, 212, 23, and 51 (again) penguins, respectively. Responses were made using 'DECSYS' [11]. The continuous scales upon which ratings were made ranged [0, 300].

C. Experimental Design

Study 1 used a between-subjects design, in which half of participants made point estimates (VAS), while the other half provided interval-valued estimates. *Study 2* used a within-subjects design, with all participants providing interval-valued estimates. In both studies, participants were instructed that they should provide their best estimates, and in the case of interval estimates, they were instructed that each interval should cover the area of the scale that they believed the correct value to fall within (a disjunctive interval).

Importantly, while in *Study 1* participants were entered into a random draw to win the prize (the jar of sweets), in *Study 2* the $100AUD prize was used to incentivise participants to give accurate interval estimates. Each participant was informed: *"The winner will be the respondent who gives the most answers where the correct value falls within their interval estimate. In the case of a draw (e.g., multiple respondents' estimates all contain the correct answer), the prize will go to whoever provided the more specific estimates (i.e., the narrowest correct intervals)"*.

D. Data Collection Procedure

Each study was approved by the University of Nottingham School of Computer Science Ethics Committee. For *Study 1*, participants were randomly allocated to a response condition (i.e., point or interval-valued), then shown either one or two information sheets. The first provided basic information about the study and use of resulting data. The second was shown only to respondents allocated to the interval response condition. This provided a brief explanation of the response mode—instructing them to mark each estimate with an ellipse, which could be made narrower or wider to indicate their uncertainty. Illustrative examples were provided here of both more and less certain responses. For *Study 2*, each participant was presented with equivalent study and response information over the internet, via DECSYS [11]. Having had the opportunity to review the study information, participants who wished to proceed recorded their consent, before beginning the task.

Study 1 responses were made on paper and encoded digitally afterwards. When making their perceptual estimates, participants were instructed that they were permitted to view the jar and its contents from different angles, but not to lift it to aid their weight judgements. After this, participants provided subjective user feedback on whichever of the two response modes they had used—these data were analysed in [5].

Study 2 participants viewed each image stimulus and made their perceptual estimates online, using DECSYS [11]. They were not asked subsequent user feedback questions due to time constraints in the workshop session.

Upon completing the survey, *Study 1* participants were given the opportunity to enter into a random draw to win the sweet jar and its contents, while *Study 2* participants were given the option to provide contact details in case of winning the performance-based prize. Each study took only around five minutes to complete for each respondent.

E. Analysis Procedure

We report a variety of descriptive and inferential statistics across both studies. First, we will report results from the larger *Study 1*, before proceeding to those from the pilot follow-up *Study 2*. The latter comprise a test of repeatability of certain *Study 1* findings under selectively different circumstances—i.e., a different task, in which:

- It is possible to be more accurate with added effort investment.
- There is a performance-based incentive to encourage making added effort investment.

More specifically, we first report IAA plots [15] illustrating grouped estimates for each *Study 1* stimulus, as well as mean interval and mean point estimates for the same stimuli. Then,

we report group-level (i.e., 'crowd') estimates and error for each question, according to a series of different aggregation metrics for each response mode. We then report individual level errors, both MBE and MAE, for each question and response mode. Following this, we examine three hypotheses:

1) That interval-valued estimates, once reduced to a point (in order to compare against a point ground truth), may be less accurate than point 'best guess' estimates.
2) That interval estimates offer an advantage over point estimates in that narrower (i.e., more certain) estimates are more accurate than wider (more uncertain) estimates—therefore permitting improved aggregation of group estimates by weighting individual estimates according to this additional information.
3) That uncertainty (i.e., interval widths) will scale positively with magnitude of the estimate.

To address the first hypothesis, we report two mixed-model ANOVAs, examining whether there were significant differences in error between estimates made for each response mode, or between task stimuli. For our main effect of interest (difference in accuracy between point and interval response modes), a priori statistical power calculations, made using G*Power [20], indicated power of $>.99$ to detect a large effect, of .81 for a medium effect, and .21 for a small effect (ANOVA: repeated measures, between factors, α = .05, f=.4, .25, .1 respectively, corr. among rep. measures = .5—cf. [21]).

We then examine a second hypothesis, reporting a linear mixed effects model (LMEM) designed to examine the impact of interval width upon estimate accuracy. We also include estimate magnitude (i.e., interval position along the response scale) as a second fixed effect, along with a two way interaction term, in order to elucidate whether the relationship between estimate width and accuracy varies depending upon magnitude of estimate (e.g., narrower intervals around smaller estimates may indicate more reliable information than narrower intervals around larger estimates). We include random intercepts, to permit differing baseline accuracy in relation to each participant and question. The model formula is therefore:

$$\gamma^{e}_{i,j} = \beta_0 + \beta_1 x^{w}_{i,j} + \beta_2 x^{m}_{i,j} + \beta_3 (x^{w}_{i,j} \cdot x^{m}_{i,j}) + \mu_i + \mu_j + \epsilon_{i,j} \tag{1}$$

where $\gamma^{e}_{i,j}$ is the outcome variable of estimate error, β is the coefficient, $x^{w}_{i,j}$ is estimate width w, $x^{m}_{i,j}$, is estimate magnitude m, and $(x^{w}_{i,j} \cdot x^{m}_{i,j})$ is the interaction between these two factors—for a given participant i and question j. β_0 denotes the fixed intercept; μ_i and μ_j denote respective random intercepts for participant and question; and ϵ represents the residual error.

To examine the third hypothesis, we then report one further LMEM, with the outcome variable of interval width, $\gamma^{w}_{i,j}$. This model is as follows:

$$\gamma^{w}_{i,j} = \beta_0 + \beta_1 x^{m}_{i,j} + \mu_i + \mu_j + \epsilon_{i,j} \tag{2}$$

Following these analyses, we will proceed to report findings from *Study 2*. These will comprise IAA plots [15] to illustrate grouped estimates for each *Study 2* stimulus, as well as a linear mixed effects analysis designed to examine whether the results of model (1)—concerning hypothesis 2—are replicated on this new task, which differs from the first in the two respects reported near the beginning of this section.

F. Aggregation Methods

We compare a total of seven aggregation measures to determine group-level ('crowd') estimates—two of these concern point estimates, and five interval estimates. Each measure is listed below, with brief explanation.

First, *Point Mean*—this is simply the arithmetic mean of all point estimates.

Second, *Point Median*—this is the median value of all point estimates. In fact, as there were 40 point estimates for each question, this was the mean of the twentieth and twenty-first estimates, ordered by magnitude.

Third *Interval Mean*—the most common method to compute the mean of a set of intervals is to calculate the mean of all left endpoints and the mean of all right endpoints, as shown below [22]:

$$\left[\frac{1}{N} \sum_{i}^{N} \underline{x}_i, \ \frac{1}{N} \sum_{i}^{N} \overline{x}_i \right] \tag{3}$$

Where $\underline{x}_i$ is each left endpoint and $\overline{x}_i$ is each right endpoint, and N is the number of responses. This value is itself an interval. In order to reach a point-valued mean—for straightforward comparison against a point-valued ground truth—the midpoint of this mean interval can be taken. However, doing so discards information concerning interval width, which is valuable in many circumstances. Note that this is then mathematically equivalent to taking the mean of all individual interval midpoints, as shown below:

$$\frac{\left(\frac{1}{N}\sum_{i}^{N} \underline{x}_i\right) + \left(\frac{1}{N}\sum_{i}^{N} \overline{x}_i\right)}{2} = \frac{\sum_{i}^{N} \left(\frac{\underline{x}_i + \overline{x}_i}{2}\right)}{N} \tag{4}$$

Fourth and fifth *Interval Median*—unlike the point-valued mean of a set of intervals, when calculating the median of all intervals, different values are obtained (we report both) depending upon whether the midpoint of the median interval is taken, or the median of the interval midpoints, as shown below (for odd numbered dataset):

$$\frac{\left\{\underline{x}_{\left(\frac{N+1}{2}\right)}\right\} + \left\{\overline{x}_{\left(\frac{N+1}{2}\right)}\right\}}{2} \neq \left(\frac{\underline{x} + \overline{x}}{2}\right)_{\left(\frac{N+1}{2}\right)} \tag{5}$$

Sixth and seventh *IAA Mean of Maxima* and *IAA Centroid*—the Interval Agreement Approach (IAA [15]), provides a group-level of agreement across the (group of) intervals in the form of a 2-dimensional distribution, effectively a fuzzy set [23], where the degree of membership at each x-value depends directly upon the degree of 'agreement' of the data—i.e., overlap between intervals. Different measures of central

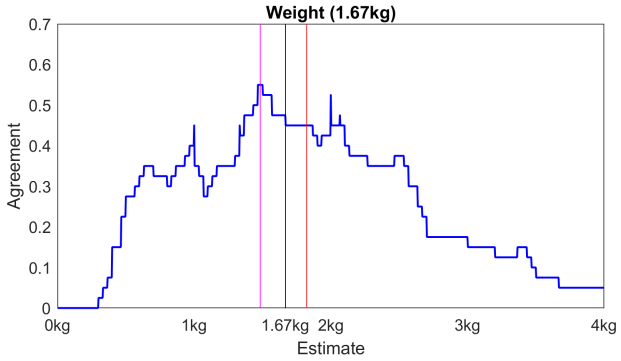

Fig. 2. IAA plot showing aggregated IV estimates for the overall weight of the sweets jar (Study 1), based upon 40 observations. Actual weight is shown in brackets and marked with a black vertical line. Magenta line shows IAA mean of maxima, and red line shows IAA centroid.

tendency of an IAA set can be taken. These may prioritise the degree of agreement, such as the point of maximum agreement between all intervals, or the mean of maxima if there are multiple such points. Alternative measures may give weight to the full range the set comprises; the centroid, for instance, represents the 'centre of mass' of the complete IAA set. It is notable however that this centroid is mathematically equivalent to a weighted average of all interval midpoints, with a greater weight given to broader intervals (see below).

$$C(x) = \frac{\sum_i^N \left(\frac{\underline{x}_i + \overline{x}_i}{2}\right) * \ (1 + \overline{x}_i - \underline{x}_i)}{\sum_i^N (1 + \overline{x}_i - \underline{x}_i)} \tag{6}$$

This runs counter to the more intuitive weighting approach, whereby narrower intervals—indicating greater confidence in the estimate—may be weighted more highly. It may therefore be considered unlikely that this metric will outperform others based on IV data, such as an unweighted mean. Note here that there also exist more complex approaches to derive central tendency from IAA sets, such as 'alpha-cut defuzzification' [24]. We purposely do not consider more complex approaches in the present analysis to limit the potential to identify a measure, or set of aggregation parameters, that performs well with the current data set by chance (i.e., overfitting).

III. Results

A. Descriptives—Study 1

IAA plots visualising *Study 1* interval estimates are shown in Figures 2 and 3. Mean *Study 1* estimates, both point and interval, are shown in Fig. 4.

Error for group-level estimates, according to each of the seven aggregation methods, is shown in Table I. Note here that we constrain each group-estimate to the same dimensions as the ground-truth for comparison. In each question of the present study the ground truth is a discrete value, and therefore when calculating accuracy on this basis we disregard some information inherent in the interval-valued estimates. This will accurately reflect some real-world circumstances, but not others.

TABLE I
GROUP-LEVEL 'CROWD' ERROR FOR EACH AGGREGATION METHOD AND QUESTION.

Agg. Method	Weight	Purple	Yellow	Green	Orange	MAE
Point Mean	+6.2	-28.7	-42.7	+2.2	-12.2	18.4
Point Median	+3.4	-34.0	-43.5	+2.5	-13.0	19.3
IV Mean	-0.7	-21.4	-42.2	+2.2	-11.1	15.5
IV Med. of Mid.	-3.1	-23.5	-43.5	-0.7	-14.2	17.0
IV Mid. of Med.	-4.8	-23.2	-45.5	-1.2	-15.0	17.9
IAA M. of Max.	-5.6	-34.0	-50.3	+2.0	-18.5	22.1
IAA Centroid	+4.4	-19.6	-42.1	+3.5	-10.2	16.0
Actual Value	50.1	74	106	26	50	

40 obs. per group. MAE is average group-level error across all questions.

TABLE II
AVERAGE INDIVIDUAL-LEVEL ERROR FOR EACH RESPONSE MODE AND QUESTION.

Response Mode	Weight	Purple	Yellow	Green	Orange	Avg.
MBE - Point	+6.2	-28.7	-42.7	+2.2	-12.2	-15.1
MBE - Interval	-0.7	-21.4	-42.2	+2.2	-11.1	-14.7
MAE - Point	20.4	31.0	43.4	7.5	15.5	23.5
MAE - Interval	19.0	26.7	43.1	9.1	16.2	22.8

40 obs. per group. Note that all MAE values are necessarily positive.

Errors (both mean bias, and mean absolute), across all individual estimates for each response mode, are shown in Table II. Note that the same constraints apply to individual interval estimates as acknowledged for group estimates, that is, we reduce intervals to their midpoints for a straightforward, but arguably naïve, assessment of accuracy.

Also note that in both Tables I and II weight estimates and error are converted to the same scale as the other questions [0,120], for ease of comparison.

B. Analyses of Variance—Study 1

Descriptive results suggest that while there were substantial differences in error between questions, no large differences were apparent between response modes. In order to formally examine differences in estimate error, we conducted two 2x5 mixed model ANOVAs—with dependent variables of MBE and MAE respectively—each included the 2-level between subjects factor of point vs IV response mode, and the 5-level within subjects factor of task question.

Both models were Greenhouse-Geisser corrected for violations of sphericity. For MBE, results indicated a significant main effect of question, $F(2.798,218.216)=114.517$, $p<.001$, but neither significant main effect of response-type, $F<1.0$, nor a significant interaction term, $F(2.798,218.216)=1.966$, $p=.124$. The same pattern was evident for MAE. A significant main effect of question was evident, $F(2.278,177.711)=81.553$, $p<.001$, but neither a significant main effect of response type, $F<1.0$, nor a significant interaction term, $F<1.0$. Due to finding no significant main effect of response mode (our variable of interest), nor significant interaction effects, we did not proceed to conduct post-hoc pairwise comparisons.

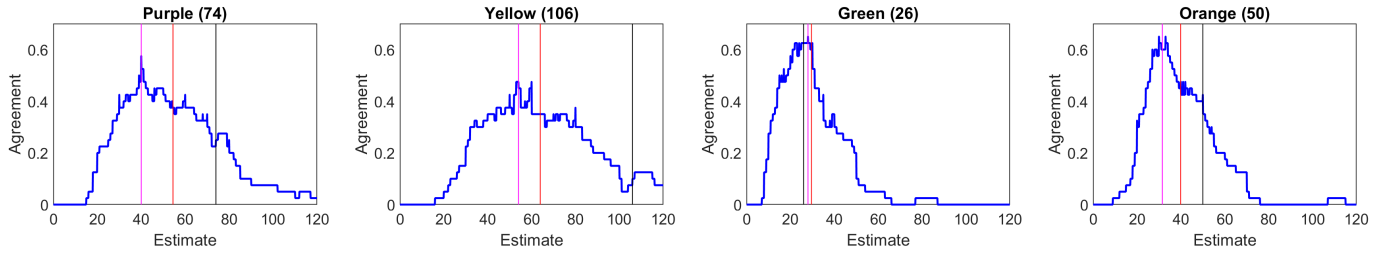

Fig. 3. IAA plots showing aggregated IV estimates for the number of sweets in the jar of each of the four colours (Study 1), based upon 40 observations. Actual values are shown in brackets and marked with a black vertical line. Magenta line shows IAA mean of maxima, and red line shows IAA centroid.

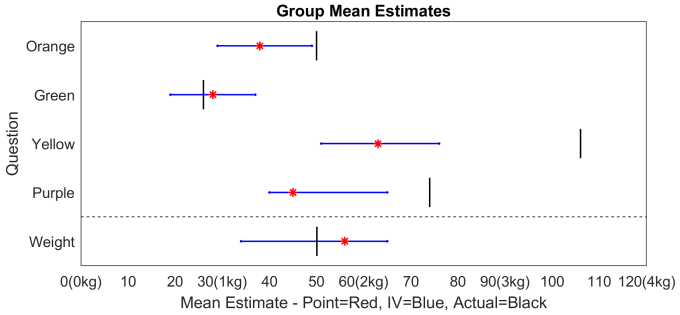

Fig. 4. Showing group mean estimates for both point and IV response conditions in Study 1. Mean point responses shown as red asterisks, Mean intervals as blue lines, and actual ground truth values as black vertical lines. x-axis labels represent sweet number, with weight in brackets.

TABLE III
MODEL SHOWING STUDY 1 EFFECTS OF IV ESTIMATE WIDTH (w), MAGNITUDE (m), AND TWO-WAY INTERACTION TERM ($w \cdot m$) ON ESTIMATE ERROR (MAE) (e).

Fixed Effects Estimates	β	SE	t	p
Intercept : $(_0)$	38.381	8.693	4.415	<.001
Width w : $(x^w_{i,j})$	-.197	.205	-.960	.338
Magnitude m : $(x^m_{i,j})$	-.281	.085	-3.304	.001
W.*M. int. $w \cdot m$: $(x^w_{i,j} \cdot x^m_{i,j})$	.002	.003	.571	.569
Random Effects Estimates	μ			
Participant intercept $(_i)$	4.041			
Question intercept $(_j)$	16.269			
Residual $\epsilon_{i,j}$	12.881			

Fitted using REML, N = 200, DF = 196, AIC = 1647.2, BIC = 1670.1

C. Linear Mixed Effects Models—Study 1

To examine the hypothesis that narrower intervals were more accurate on this task than wider intervals, as well as whether this effect varies in relation to estimate magnitude, we implemented the LMEM shown in (1). Results are reported in Table III.

To examine the hypothesis that larger interval estimates (i.e., of greater magnitude) were wider, reflecting greater uncertainty—phrased differently, that uncertainty scaled positively with the size of the estimate—we implemented the LMEM shown in (2). Results are reported in Table IV.

D. Descriptives—Study 2

IAA plots visualising *Study 2* interval estimates are shown in Figure 5.

TABLE IV
MODEL SHOWING STUDY 1 EFFECT OF IV ESTIMATE MAGNITUDE (m) ON ESTIMATE WIDTH (w).

Fixed Effects Estimates	β	SE	t	p
Intercept : $(_0)$	15.698	2.655	5.916	<.001
Magnitude m : $(x^m_{i,j})$	.165	.032	5.178	<.001
Random Effects Estimates	μ			
Participant intercept $(_i)$	8.899			
Question intercept $(_j)$	3.623			
Residual $\epsilon_{i,j}$	6.840			

Fitted using REML, N = 200, DF = 198, AIC = 1447.0, BIC = 1463.5

TABLE V
MODEL SHOWING STUDY 2 EFFECTS OF IV ESTIMATE WIDTH (w), MAGNITUDE (m), AND TWO-WAY INTERACTION TERM ($w \cdot m$) ON ESTIMATE ERROR (MAE) (e).

Fixed Effects Estimates	β	SE	t	p
Intercept : $(_0)$	-18.032	9.584	-1.881	.063
Width w : $(x^w_{i,j})$	.768	.268	2.866	.005
Magnitude m : $(x^m_{i,j})$	.474	.089	5.351	<.001
W.*M. int. $w \cdot m$: $(x^w_{i,j} \cdot x^m_{i,j})$	-.005	.002	-2.499	.014
Random Effects Estimates	μ			
Participant intercept $(_i)$	7.563			
Question intercept $(_j)$	6.387			
Residual $\epsilon_{i,j}$	33.898			

Fitted using REML, N = 88, DF = 84, AIC = 897.50, BIC = 914.52

E. Linear Mixed Effects Models—Study 2

We now report a replication of the former model (1) on the *Study 2* data set. Results are reported in Table V.

IV. SUMMARY, CONCLUSIONS AND FUTURE WORK

This paper documents two studies designed to empirically evaluate the impact of permitting interval-valued perceptual estimates—and thereby explicitly acknowledging the varying degrees of uncertainty in these estimates—upon the 'Wisdom of Crowds' (i.e., accuracy of group-level estimates).

Collecting intervals provides greater informational capacity within each response than collecting points. We propose that IV responses can capture response uncertainty (i.e., epistemic, or disjunctive set-valued information), and also inherent range in the appropriate response (i.e., ontic, or conjunctive set-valued information—cf. [25], [26]). The present paper focuses exclusively the former case—i.e., we established a single

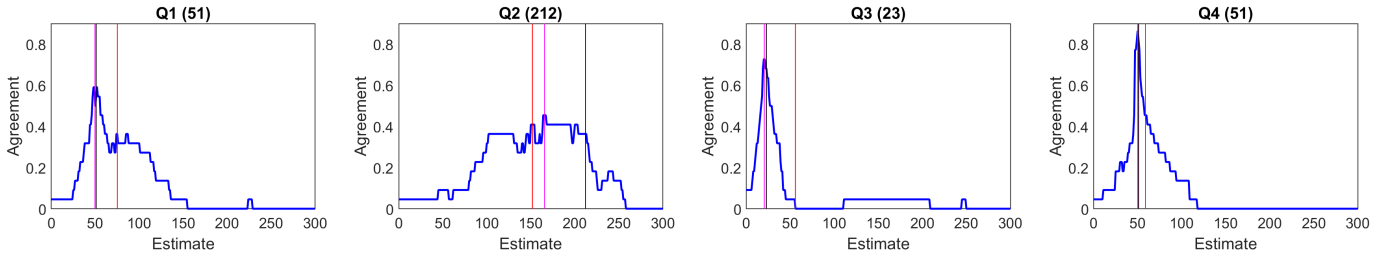

Fig. 5. IAA plots showing aggregated IV estimates for the number of penguins shown in each of the four stimuli (Study 2), based upon 22 observations. Actual values are shown in brackets and marked with a black vertical line. Magenta line shows IAA mean of maxima, and red line shows IAA centroid.

point ground truth for each experimental stimulus, therefore the information captured by interval estimate widths (if used correctly) should represent the level of epistemic uncertainty associated with each individual estimate. This will accurately reflect some real-world circumstances, but not others.

To summarise, results of *Study 1* found a tendency towards underestimation of larger true values, in both point and IV response modes. Group-level ('crowd') estimates based on IV estimates tended to slightly outperform those based on point estimates, the mean of intervals was the best performing measure. The IAA 'Mean of Maxima' was the exception to this rule, having the greatest error on three out of five questions, and overall. This poor performance was likely due to exacerbation of the general tendency to underestimate true values—taking the value of maximum agreement effectively eliminates the contribution of outliers to the group estimate, but these may be crucial to the accuracy of the 'Wisdom of Crowds' effect. Likewise, the bounding of the response scale (to 4kg & 120 sweets maxima) may have contributed to underestimation on this task, by precluding more extreme positive outliers. As well as comparing accuracy of 'crowd' estimates, we examined two primary research questions:

1) When specific predictions are required (i.e., with a point ground truth), does permitting imprecise IV estimates reduce accuracy of individual estimates (by comparison with point 'best estimates')?
2) Does capturing the additional information associated with each IV estimate (i.e., concerning uncertainty) permit more accurate group-level estimates, by weighting individual estimates according to their specificity?

In relation to hypothesis one, inferential analyses revealed significant differences in individual estimate error (both MBE & MAE) between questions, but no evidence of a substantial difference between point and 'collapsed' interval-valued estimates (i.e., midpoints) on this task, despite a priori calculations indicating high statistical power to detect either a large or medium effect with this sample size. As discussed earlier, comparing the accuracy of individual interval-valued estimates versus point ground truths—by collapsing these to points—is a naïve approach to assessing their value, as this process disregards the additional information that they capture. Nonetheless, it is interesting to observe that permitting respondents to provide deliberately imprecise (i.e., uncertain) responses did not lead to any significant loss in estimate accuracy, even when reducing these responses to their midpoints.

In relation to hypothesis two, for *Study 1*, linear mixed effects analysis revealed no significant relationship between widths of interval estimates and the accuracy of these estimates (see Table III). This model did identify a significant negative effect of estimate magnitude on error, consistent with the overall tendency towards underestimation in this task, but also found no significant two-way interaction between these factors, indicating that estimate width did not reliably relate to accuracy irrespective of the magnitude of the estimate. These results mean that weighting interval estimates by their specificity (i.e., inversely by their width) would not offer any improvement on group-level estimates for the *Study 1* data set.

It is surprising that those respondents who felt confident to provide more specific estimates were not, in fact, any more accurate than those who felt less able to do so. This result implies one of two things—on this task, either intervals didn't capture uncertainty, or uncertainty didn't relate to accuracy. The former seems inconsistent with previous research, which has established a reliable relationship between interval response width and respondent uncertainty [4], the latter also seems to run counter to both intuitions and existing evidence [27]. In either case, the findings may represent some specific issue relating to the *Study 1* task and/or sample. We propose two potential explanations for further evaluation—relating to participant motivation and task difficulty, respectively.

1) Due to insufficient incentives—i.e., the prize was not high value, and the draw was random—many participants did not invest enough effort to provide high quality estimates.
2) Due to the difficulty of the task—i.e., not all sweets were visible, so participants couldn't be genuinely certain—there was insufficient inter-rater variance in certainty to capture a significant effect.

In short, either participants weren't trying or simply weren't able to give precise estimates on this task—each leading to low-fidelity responses. This may reflect generally poor quality data, or it may be driven by a confounding subset of participants who failed to engage with the task, thereby providing both inaccurate estimates and failing to effectively use the interval response mode by defaulting to narrow, 'point-like', responses. Best practice in identifying 'careless respondents'

to IV survey questions is another area that will require substantial consideration and development. We planned a follow-up study to explore these hypotheses. *Study 2* represents a pilot for this study, the broad aim being to test the repeatability of *Study 1* findings on a different task, in which:

- It was possible to be more accurate with added effort investment.
- There was a performance-based incentive to encourage making added effort investment.

Importantly, the task incentivised the *narrowest correct* estimates, so it was in participants' interest for their intervals not only to be close to the true values, but also to accurately reflect their certainty. Also note that performance on this task (penguin counting) should scale positively with the potentially *intra*-rater variable factor of effort invested (i.e., care taken), rather than another factor that varies only on an inter-rater level (e.g., expertise), this should increase 'metacognitive awareness' [28] of differences in knowledge and performance, thereby increasing signal to noise ratio.

Results on the second task did not replicate those on the first (see Table V). A significant positive effect of estimate width on estimate error was found, as well as an interaction term indicating that this effect was significantly stronger when estimates were of lower magnitude (i.e., reported certainty was more reliable for lower estimates). This means that, on this task, weighting intervals by their certainty could offer significant improvements in group-level 'crowd' estimate accuracy. However, as task changes were made relating to both putative explanations for the null *Study 1* results, this pilot study is not able to differentiate the two accounts. We hope soon to publish results from a more comprehensive empirical study, with a substantially larger and more representative sample, additional experimental manipulations designed to tease apart the impacts of each factor, as well as a point response control condition, to permit comparison versus a conventional response mode on the updated task design.

To summarise, in this paper we report results and discuss implications of two empirical studies examining the efficacy of interval-valued perceptual estimates. Results suggest that intervals can capture more (useful) information in this context but, importantly, that this may depend on the nature of the task, as well as levels of participant motivation and engagement. These considerations are important more generally across quantitative data collection—however, it is possible that IV responses may be more sensitive to these factors, due to lack of familiarity and potential to revert to habitualised 'point-like' responses. Further work is ongoing.

REFERENCES

[1] R. Likert, "A technique for the measurement of attitudes.," *Archives of psychology*, 1932.

[2] R. C. Aitken, "A growing edge of measurement of feelings [abridged] measurement of feelings using visual analogue scales," 1969.

[3] M. Freyd, "The graphic rating scale.," *Journal of educational psychology*, vol. 14, no. 2, p. 83, 1923.

[4] Z. Ellerby, C. Wagner, and S. B. Broomell, "Capturing richer information: On establishing the validity of an interval-valued survey response mode," *Behavior Research Methods*, pp. 1–23, 2021.

[5] Z. Ellerby and C. Wagner, "Do people prefer to give interval-valued or point estimates and why?," in *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–6, IEEE, 2021.

[6] J. Zamora and V. Kreinovich, "Why ovals in eliciting intervals?," 2021.

[7] Z. Ellerby, J. McCulloch, M. Wilson, and C. Wagner, "Exploring how component factors and their uncertainty affect judgements of risk in cyber-security," in *International Conference on Critical Information Infrastructures Security*, pp. 31–42, Springer, 2019.

[8] Z. Ellerby, O. Miles, J. McCulloch, and C. Wagner, "Insights from interval-valued ratings of consumer products—a decsys appraisal," in *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, IEEE, 2020.

[9] J. Navarro, C. Wagner, U. Aickelin, L. Green, and R. Ashford, "Exploring differences in interpretation of words essential in medical expert-patient communication," in *2016 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 2157–2164, IEEE, 2016.

[10] K. J. Wallace, C. Wagner, M. J. Smith, *et al.*, "Eliciting human values for conservation planning and decisions: a global issue," *Journal of environmental management*, vol. 170, pp. 160–168, 2016.

[11] Z. Ellerby, J. McCulloch, J. Young, and C. Wagner, "Decsys–discrete and ellipse-based response capture system," in *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, IEEE, 2019.

[12] T. C. Havens, C. Wagner, and D. T. Anderson, "Efficient modeling and representation of agreement in interval-valued data," in *2017 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, IEEE, 2017.

[13] J. McCulloch, Z. Ellerby, and C. Wagner, "On comparing and selecting approaches to model interval-valued data as fuzzy sets," in *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, IEEE, 2019.

[14] J. McCulloch, Z. Ellerby, and C. Wagner, "Choosing sample sizes for statistical measures on interval-valued data," in *2020 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, IEEE, 2020.

[15] C. Wagner, S. Miller, J. M. Garibaldi, D. T. Anderson, and T. C. Havens, "From interval-valued data to general type-2 fuzzy sets," *IEEE Transactions on Fuzzy Systems*, vol. 23, no. 2, pp. 248–269, 2015.

[16] S. Kabir and C. Wagner, "Interval-valued regression-sensitivity to data set features," in *2021 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, pp. 1–7, IEEE, 2021.

[17] S. Kabir, C. Wagner, and Z. Ellerby, "Towards handling uncertainty-at-source in ai–a review and next steps for interval regression," *arXiv preprint arXiv:2104.07245*, 2021.

[18] F. Galton, "Vox populi (the wisdom of crowds)," *Nature*, vol. 75, no. 7, pp. 450–451, 1907.

[19] K. F. Wallis, "Revisiting francis galton's forecasting competition," *Statistical Science*, pp. 420–424, 2014.

[20] F. Faul, E. Erdfelder, A.-G. Lang, and A. Buchner, "G* power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences," *Behavior research methods*, vol. 39, no. 2, pp. 175–191, 2007.

[21] J. Cohen, "Statistical power analysis for the behavioural sciences. hillsdale, nj: Laurence erlbaum associates," 1988.

[22] S. Ferson, V. Kreinovich, J. Hajagos, W. Oberkampf, and L. Ginzburg, "Experimental uncertainty estimation and statistics for data having interval uncertainty," *Sandia National Laboratories, Report SAND2007-0939*, vol. 162, 2007.

[23] L. A. Zadeh, "Fuzzy sets," in *Fuzzy sets, fuzzy logic, and fuzzy systems: selected papers by Lotfi A Zadeh*, pp. 394–432, World Scientific, 1996.

[24] A. Pourabdollah, J. M. Mendel, and R. I. John, "Alpha-cut representation used for defuzzification in rule-based systems," *Fuzzy Sets and Systems*, vol. 399, pp. 110–132, 2020.

[25] I. Couso and D. Dubois, "Statistical reasoning with set-valued information: Ontic vs. epistemic views," *International Journal of Approximate Reasoning*, vol. 55, no. 7, pp. 1502–1518, 2014.

[26] D. Dubois and H. Prade, "Gradualness, uncertainty and bipolarity: making sense of fuzzy sets," *Fuzzy sets and Systems*, vol. 192, pp. 3–24, 2012.

[27] B. H. Bornstein and D. J. Zickafoose, "I know i know it, i know i saw it: The stability of the confidence–accuracy relationship across domains.," *Journal of experimental psychology: Applied*, vol. 5, no. 1, p. 76, 1999.

[28] G. Schraw and R. S. Dennison, "Assessing metacognitive awareness," *Contemporary educational psychology*, vol. 19, no. 4, pp. 460–475, 1994.