Do People Prefer to Give Interval-Valued or Point Estimates and Why?

Zack Ellerby, Christian Wagner (Senior Member, IEEE)
LUCID Research Group, University of Nottingham, UK

zack.ellerby; christian.wagner@nottingham.ac.uk

***Abstract*—Capturing interval-valued, as opposed to more conventional point-valued data, offers a potentially efficient method of obtaining richer information in individual responses. In turn, interval-valued data provide a strong foundation for subsequent fuzzy set based modelling—e.g., using the Interval Agreement Approach. In 2019, open-source software (DECSYS) was released to enable digital administration of interval-valued surveys using an ellipse response mode. This study follows on from an appraisal of this software and demonstration of practical value of the approach, reported last year, in one of many potential real-world applications (consumer preference research). A key ambition of ellipse-based interval elicitation is to maximise response efficiency—i.e., minimising workload and complexity in obtaining this richer information. User experience is therefore a vital consideration regarding potential for broader adoption. The present paper documents a direct empirical comparison between interval-valued response elicitation (using ellipses) and a conventional point-valued counterpart (using a Visual Analogue Scale), in terms of user experience during completion of a simple quantitative estimation task. We examine differences in perceived ease-of-use, unnecessary complexity and effective communication of desired responses, as well as overall liking—with positive outcomes for the interval-valued response mode in each case. We also report results of multiple regression analyses examining how the first three variables contribute to participants' overall liking of each response mode, as well as exploring differences driven by potentially important demographic factors (i.e., gender, age & native English speaking).**

I. Introduction and Background

Interval-valued survey responses offer potential to capture richer information than conventional point responses (such as Likert-type [1], or Visual Analogue [2] scales). However, as things stand they remain rarely used in either research, industry or wider society. There have been some good historical reasons for this, which may no longer be entirely justified. Specifically, potential barriers have included:

- A lack of clear evidence for real-world efficacy.
- Practical difficulties in administering interval-valued surveys at scale—with an absence of software to enable digital administration, and collation of this data following collection on paper being relatively time-consuming.
- Playing catch-up on breadth and accessibility of appropriate methods for statistical analysis, by comparison with point-valued data, which have benefited from decades of head start in mainstream development.
- Potential difficulties from the perspective of survey respondents—i.e., perceived increases in workload or complexity.

Despite these, a growing body of complementary theoretical, practical and empirical work has contributed to a recent surge in interest in interval-valued response elicitation. This research has worked to address each of these barriers—ranging from empirical studies to establish efficacy and potential value across a variety of real-world applications [3]–[7], to creation of open-source software tools to facilitate administration at scale [8], to development and evaluation of mathematical and statistical methods to best handle and interpret the richer information that is captured [9]–[17], cf. Fig. 1—building upon previous work in the fields of Interval Arithmetic, cf. [18]–[20] and Fuzzy Set Theory [21].

The present study aims to directly address the fourth item on this list. That is, whether or not intervals are well-received by users (i.e., survey respondents), who may find this type of response mode more difficult to understand, or perceive added complexity without appreciating its added informational capacity. Last year, in [4], we conducted an initial assessment of user feedback following primary data collection, but without any control condition. This paper extends upon this by direct empirical comparison between user feedback on the interval-valued response mode and an equivalent point-valued counterpart (the Visual Analogue Scale, or 'VAS'), following completion of a short perceptual judgement task.

It is important to note that in the present paper, when we refer to the interval-valued response format, we mean the ellipse response mode documented in [4], [5], [8] (cf. Fig. 2). This method was designed with response efficiency as a primary objective, leveraging the quick and intuitive nature of 'circling' areas of interest. It is intended to fill a niche, as an efficient compromise between the most prevalent conventional quantitative approaches, which predominantly elicit point data (e.g., Likert-type ordinal scales [1], or VAS [2], [22]) and alternatives of substantially greater complexity—e.g., qualitative interviews, and methods of eliciting more complex distributions [23]–[25], such as the 'Fuzzy Graphic Rating Scale' (FRS) [26]–[29], and the 'Sheffield Elicitation Framework' (SHELF) [30], [31].

As stated in [5]:

> "The ellipse approach is designed primarily to streamline the process of interval-valued data collec-

This work was funded in part by the UK EPSRC EP/P011918/1 grant.

978-1-6654-4407-1/21/$31.00 © 2021 IEEE

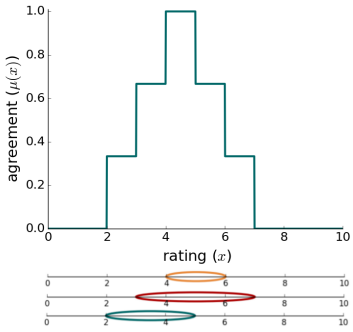

Fig. 1. Example of an IAA fuzzy set constructed from three intervals. The membership assigned to x is the degree of agreement between the intervals.

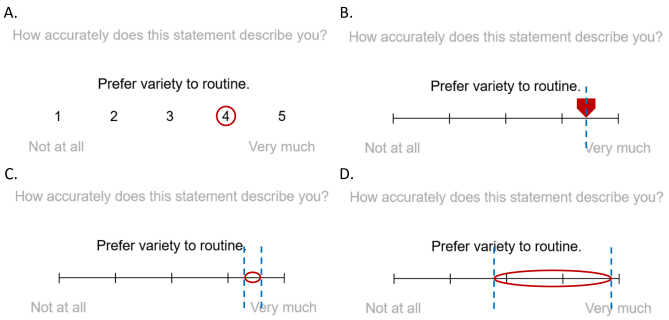

Fig. 2. Illustrative responses. A: Likert-type (ordinal). B: VAS-type. C: Ellipse (low uncertainty). D: Ellipse (high uncertainty). Divisions (sub markers) on continuous scales are illustrative here only and questions of appropriate scale design apply as for traditional scales.

> tion so far as to provide an alternative to point-valued response modes (e.g., Likert-type or VAS), where it offers a substantial informational advantage at a minimally increased, or potentially even reduced, workload—e.g., by counteracting choice paralysis when selecting between multiple potentially appropriate discrete alternatives or requiring fewer questions to be asked."

The ultimate objective is to establish whether ellipse responses could provide richer quantitative information—concerning either (or both) response uncertainty (i.e., epistemic, or disjunctive set-valued information), and inherent range in the appropriate response (i.e., ontic, or conjunctive set-valued information) [32], [33]—without sacrificing response efficiency or user experience. If so, they could be used in situations where, conventionally, intervals would be avoided 'for simplicity's sake'. This could bring the informational advantages associated with interval-valued data not only to situations where these may be expected a priori, but also to the multitude of cases where the benefits of better understanding uncertainty or variability in the data may not have been anticipated in advance—only becoming evident in retrospect. The study reported here addresses the second part of this objective, directly examining potential benefits and penalties in user experience by comparison with a traditional point response mode.

In Section II we describe the participants, stimuli and procedure of the experimental study—including details regarding data collection and analysis. In Section III we report the results of both descriptive and inferential analyses of the relevant data obtained in the study. In Section IV we summarise key findings and discuss their implications.

II. Method

A. Study Participants

A total of 80 participants completed this study, recruited through opportunity sampling across three UK campuses of the University of Nottingham. 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 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.

B. Statistical Power Estimates

A priori power calculations were made using G*Power [34]. When considering pairwise comparisons between subjective feedback ratings made for each experimental condition, these indicated power of .94 to detect a large effect, of .60 for a medium effect, and .14 for a small effect (difference between two independent means, two-tailed, α = .05, d=.8, .5, .2 respectively—c.f. [35]). For multiple regression analyses (used to investigate putative factors of overall liking), with all seven predictors (including all two-way interaction terms), power was estimated to be .98 to detect a large effect, .66 for a medium effect and .11 for a small effect (R^2 deviation from zero, α = .05, F^2 =.35, .15, .02 respectively).

C. Questions and Experimental Stimuli

In order to provide subjective feedback on a response mode, participants first had to use it. They did so on a short perceptual judgement task, in which they provided five estimates. Specifically, participants were each presented with a transparent plastic sweet jar, approximately half filled with sweets of four different colours (Bassett's Jelly Babies), and were tasked with judging its overall weight, as well as the number of sweets of each colour.

The subsequent subjective feedback section of the questionnaire comprised four items, adapted initially from the System Usability Scale [36]. Respondents were first instructed: "Please think about the method used to answer the previous questions, we call this the 'response-format'. Then mark a single box to provide your feedback on the following questions." They were then asked the extent to which they agreed to the following questions: "I found the response-format easy to use.", "I found the response-format unnecessarily complex.", "I found that the response-format allowed me to effectively communicate my desired response.", "Overall, I liked the response-format". All responses to this section were made using a traditional 5-point

ordinal response scale, ranging from 1—Strongly Disagree, to 5—Strongly Agree, to minimise complexity.

Finally, participants were asked to provide some basic demographic information about themselves. They were asked three further questions here: what was their gender (self-identified), what was their age, and whether or not they considered themselves to be a native English speaker.

D. Experimental Design

The study used a between-subjects design, in which half of participants provided feedback on use of discrete estimates, having used a point response format (VAS) in the preceding task. The other half provided feedback on their use of interval-valued estimates—in this case participants were instructed that each interval should cover the area of the scale that they believed the correct value to fall within (a disjunctive interval).

E. Data Collection Procedure

The study procedure was approved by the Ethics Committee at the University of Nottingham School of Computer Science. Before the task, participants were randomly allocated to a response condition (i.e., point or interval-valued); they were then shown either one or two information sheets. The first provided general information about the study, use of resulting data, and participants' freedom to withdraw at any time. 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 the degree of response uncertainty. Illustrative examples were provided here of both more and less certain responses. Having had the opportunity to review these information sheets and ask questions, participants who wished to proceed signed the consent form and began the study questionnaire.

Note that all responses in this study were made on paper and encoded digitally afterwards, due to the short nature of the task. First, respondents provided their perceptual estimates—when making these, 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. Following the perceptual judgement task, participants were asked to provide their level of agreement with four statements concerning their subjective user experience, in relation to whichever of the two response formats that they had just used. They were also asked three basic demographic questions (cf. subsection C for specific questions). Upon completing the survey, participants were given the opportunity to enter into a random draw to win the sweet jar and its contents.The whole process took around five minutes for each respondent.

F. Analysis Procedure

Descriptive and inferential statistics are used to report and interpret study results. Descriptive statistics present average ratings for each response format. 95% confidence intervals (representing between-subjects variance in estimates, and not to be confused with response intervals) are also provided, from which basic inferences can be drawn about agreement or disagreement with each feedback statement. In addition, independent samples t-tests are applied to inform differences between the two response modes on each rated attribute, as well as in terms of overall liking.

TABLE I
REGRESSION VARIABLES, NOTATIONS AND ASSOCIATED QUESTIONS.

Not.	Variable	Survey Item
r	Response Mode	N/A (Point-1, Interval-valued-2)
e	Ease-of-use	'I found the response-format easy to use.' (1–5)
u	Unnecessary complexity	'I found the response-format unnecessarily complex.' (1–5)
c	Effective communication	'I found that the response-format allowed me to effectively communicate my desired response.' (1–5)
g	Gender	'What is your gender?' (Male-1, Female-2)
a	Age	'What is your age?' (0–99)
n	Native speaker	'Are you a native English speaker?' (Yes-1, No-2)
o	Overall liking	'Overall, I liked the response-format.' (1–5)

Variable coding shown in brackets.

Following this, we apply multiple linear regression modelling to assess the influence of each response mode attribute and respondent demographic factor (cf. Table I) upon overall liking. We conduct two separate multiple regression analyses—using different combinations of factors to address two research questions, respectively:

- Q1: How did the three attributes (ease-of-use, unnecessary complexity, effective communication) explain differences in overall liking—and did significant unexplained variance remain between response modes, potentially attributable to other factors?
- Q2: Were there significant differences in overall liking of the two response modes depending upon respondent demographics (gender, age, native English speaking)?

Multiple regression (cf. [37]) estimates the contribution of each factor in the model together upon the outcome variable (i.e., overall liking), with the size and direction of each contribution reflected in the form of β weights. The model's predictor variables (i.e., response mode & rated attributes for analysis one, or response mode & demographic factors for analysis two) are entered as fixed effects, alongside two-way interaction terms in each model ($x^1 \cdot x^2$) between response mode and the other factors. These represent combined effects, which permit estimation of any differential contributions of either the attributes or demographic factors between response modes (e.g., older participants may have liked the discrete response mode more, but younger participants may have liked the ellipse response mode more).

Due to the high number of initial factors (seven), of which some may be redundant or irrelevant, an iterative process of backwards stepwise reduction was used to 'prune' the factor

set present within each model. This leaves only those that are found to contribute significantly to the outcome variable. Specifically, this process began by selection, from the pool of all non-significant effects, of the effect with the *t*-statistic closest to zero. This variable was then removed and the model re-calculated. This procedure was repeated until a final model was determined, within which all effects were statistically significant. This process was implemented for the purposes of increasing model interpretability, and power to detect significant effects—although it is important to bear in mind that this method can lead to inflation of the Type 1 error rate for factors retained in the final model, by comparison with retaining all initial factors. This should be considered when interpreting results. Here we report initial model outputs with all factors, as well as final models, to mitigate this issue.

Refer to Table I for variable notations. The initial model for effects of response mode and the three attribute ratings is

$$\begin{aligned}\gamma_i^o =&\beta_0 + \beta_1 x_i^r + \beta_2 x_i^e + \beta_3 x_i^u + \beta_4 x_i^c + \beta_5 (x_i^r \cdot x_i^e) \\ &+ \beta_6 (x_i^r \cdot x_i^u) + \beta_7 (x_i^r \cdot x_i^c) + \epsilon_i \end{aligned} \quad (1)$$

where β_z is the coefficient, x_i^r is the value coding for r (response mode—1,2), x_i^e is the rating for e (ease-of-use—1-5), and $(x_i^r \cdot x_i^e)$ is the interaction between these two factors—for a given participant i. β_0 denotes the fixed intercept and ϵ represents the error.

Likewise, the initial model for effects of response mode and respondent demographics is

$$\begin{aligned}\gamma_i^o =&\beta_0 + \beta_1 x_i^r + \beta_2 x_i^g + \beta_3 x_i^a + \beta_4 x_i^n + \beta_5 (x_i^r \cdot x_i^g) \\ &+ \beta_6 (x_i^r \cdot x_i^a) + \beta_7 (x_i^r \cdot x_i^n) + \epsilon_i \end{aligned} \quad (2)$$

Each of the initial models was then subjected to the backwards stepwise variable elimination procedure, described above, to remove non-significant effects.

III. Results

A. Descriptives and Pairwise Comparisons

Descriptive results, for both point and interval-valued response modes, are shown in Table II. Note that although subjective feedback ratings were collected using a conventional five point ordinal scale (ranging from 1—Strongly Disagree, to 5—Strongly Agree), these were re-scaled to the range -2, 2 (i.e., by subtracting three from each rating), so that negative values indicate disagreement and positive values agreement. It is clear from means and 95% confidence intervals that both groups of participants rated their agreement as significantly greater than zero on the three positive factors, and significantly lower than zero on the one negative factor ('unnecessarily complex')—this was true for both response modes.

However, p-values, obtained through independent samples t-tests (two-tailed), further indicate that respondents in the interval-valued response group rated their response mode as more effective in allowing them to communicate their desired responses, as well as liking it more overall. No significant differences were evident between response modes concerning either ease of use or unnecessary complexity.

TABLE II
Subjective feedback ratings (ranging from -2, 2). 95% Confidence Intervals shown in brackets.

Response Mode	Easy to use	Unnecessarily complex	Effectively communic.	Overall liking
Point	1.30 (.25)	-1.35 (.32)	0.88 (.30)	1.00 (.26)
Interval	1.40 (.25)	-1.35 (.30)	1.50 (.19)	1.53 (.22)
p-value	.59	1.00	$< .001$	$< .001$

40 obs. per group. p-values two-tailed, uncorrected for mult. comparisons.

TABLE III
Effects of response mode, three examined response mode attributes, and two-way interaction terms, on overall liking.

Effect Estimates	β	SE	t	p
Intercept : $(_0)$	.235	.409	.575	.567
Response mode r : (x_i^r)	-.179	.294	-.608	.545
Easy to use e : (x_i^e)	.315	.243	1.295	.199
Unnecessarily complex u : (x_i^u)	.157	.187	.840	.404
Effectively communicate c : (x_i^c)	.521	.219	2.383	.020
r*e interaction : $(x_i^r \cdot x_i^e)$	.028	.151	.185	.854
r*u interaction : $(x_i^r \cdot x_i^u)$	-.162	.120	-1.359	.178
r*c interaction : $(x_i^r \cdot x_i^c)$	.040	.157	.257	.798
Residual ϵ_i	.486			

N = 80, DF = 7,72, F = 22.1, $p < .001$, $R^2 = .682$, $Adj.R^2 = .651$

B. Multiple Regression Analyses

Results of the first linear multiple regression analysis—focusing on importance of the three attribute ratings on overall liking—are shown in Table III. Table IV shows the final model, following the variable reduction process.

Two factors were retained in the final model, each identified as holding substantial influence over overall liking of each response mode. First, how effectively the response mode was perceived to allow communication of the desired response. This was found to have the most robust effect—with every point increase in this rating increasing the rating for overall liking by, on average, approximately .6. Perceived ease-of-use was the second significant factor—with this found to increase overall liking by approximately .4 for each point increase. By contrast, perceived unnecessary complexity was not found to explain significant variance in overall liking once accounting for the two aforementioned factors. In addition, the non-significant effect of response mode (x^r) indicates that no significant variation in overall liking between the two response modes remained unexplained beyond that accounted for by the first two factors. No two-way interaction terms (i.e., $x^1 \cdot x^2$) were found to be significant, indicating that the

TABLE IV
Model one effects on overall liking, following variable reduction process.

Effect Estimates	β	SE	t	p
Intercept : $(_0)$	.026	.120	.216	.830
Easy to use e : (x_i^e)	.392	.070	5.611	$< .001$
Effectively communicate c : (x_i^c)	.596	.066	9.008	$< .001$
Residual ϵ_i	.489			

N = 80, DF = 2,77, F = 73.5, $p < .001$, $R^2 = .656$, $Adj.R^2 = .647$

TABLE V
EFFECTS OF RESPONSE MODE, THREE EXAMINED DEMOGRAPHIC FACTORS, AND TWO-WAY INTERACTION TERMS, ON OVERALL LIKING.

Effect Estimates	β	SE	t	p
Intercept : $(_0)$	-2.107	1.376	-1.532	.130
Response mode r : (x_i^r)	1.878	.885	2.123	.037
Gender g : (x_i^g)	1.158	.550	2.104	.039
Age a : (x_i^a)	.019	.026	.715	.477
Native English speaker n : (x_i^n)	.454	.603	.753	.454
r*g interaction : $(x_i^r \cdot x_i^g)$	-.560	.357	-1.570	.121
r*a interaction : $(x_i^r \cdot x_i^a)$	-.017	.018	-.933	.354
r*n interaction : $(x_i^r \cdot x_i^n)$	-.154	.384	-.400	.690
Residual ϵ_i	.773			

N = 80, DF = 7,72, F = 2.51, p=.023, $R^2 = .196$, $Adj.R^2 = .118$

TABLE VI
MODEL TWO EFFECTS ON OVERALL LIKING, FOLLOWING VARIABLE REDUCTION PROCESS.

Effect Estimates	β	SE	t	p
Intercept : $(_0)$	.475	.277	1.713	.091
Response mode r : (x_i^r)	.525	.175	2.994	.004
Residual ϵ_i	.784			

N = 80, DF = 1,78, F = 8.96, p=.004, $R^2 = .103$, $Adj.R^2 = .092$

influence of each of the three factors upon overall liking did not substantially differ between the two response modes.

Results of the second multiple regression analysis—focusing on importance of three demographic factors on overall liking—are shown in Table V. Table VI shows the final model, following variable reduction process.

Here, only one factor was retained in the final model. This was the response mode used (x^r), with use of the interval response mode associated with an increase in overall liking of just over .5 (consistent with descriptive results shown in Table II). By contrast, none of the three demographic factors were found to explain significant variance in overall liking in the final model—being female was associated with significantly higher overall liking ratings in the initial model, but this effect did not survive the pre-specified variable reduction process. In addition, no significant two-way interaction terms (i.e., $x^1 \cdot x^2$) were evident, indicating that the (lack of) influence of the demographic factors on overall liking did not vary substantially between response modes.

IV. SUMMARY, CONCLUSIONS AND FUTURE WORK

This paper documents a study designed to empirically assess user experiences of an interval-valued response mode (cf. [4], [5], [8], [14], [16]), with direct reference to a conventional point alternative (VAS—cf. [2], [22]), to inform its usability and potential for future uptake. It also addresses two further research questions. First, what were the influences of perceived ease-of-use, unnecessary complexity, and capacity for effective communication upon overall liking? Second, what were the impacts of gender, age, and whether the respondent was a native English speaker on user feedback ratings?

Collecting intervals, rather than points, provides greater informational capacity within individual responses. We propose that ellipse responses can capture response uncertainty (i.e., epistemic, or disjunctive set-valued information), and inherent range in the appropriate response (i.e., ontic, or conjunctive set-valued information—cf. [32], [33]), and we hypothesise that they can achieve this without sacrificing a substantial degree of response efficiency, or user experience. In this study we test this hypothesis, examining just how efficient respondents find the ellipse response mode to be in practice.

A similar preliminary analysis of user feedback was performed in a previous study [4], finding initial evidence that:

> "Participants reported that they found the survey easy to use, that it was not unnecessarily complex, that it allowed them to effectively communicate their desired responses, and that they liked it overall. Of course, these ratings should not be over-interpreted in the absence of comparable ratings for traditional, or other alternative response formats."

In this paper we replicate these positive findings, with no clear differences in ratings of the ellipse response mode by comparison with this preceding study (i.e., 95% CIs overlap in each case). Crucially, we also extend these findings, through direct empirical comparisons with a conventional (i.e., point response) control condition. These revealed that, on the present task, ellipse responses were rated as neither less easy to use nor more unnecessarily complex than point responses. By contrast, intervals were rated as permitting significantly more effective communication of participants' desired responses, and were significantly preferred overall.

These results are promising for future general acceptance of efficient interval-valued response capture. They suggest that respondents did not see this added dimension to their responses as unnecessary or redundant—even on the simple task described here—and that they can instead appreciate the added richness of response that it allows them to provide.

As promised in the title, this paper not only examines whether respondents preferred to give interval or point estimates, but also informs why. Two analyses were conducted to examine this question, focusing on response mode attributes and demographic factors, respectively. These revealed no significant differences in liking of the two response formats relating to either gender, age, or native English speaking. Rather, they indicate that overall liking is determined by two primary factors: perceived ease-of-use, and perceived communicative effectiveness (confirming preliminary findings in [4]). Interpreted together with earlier comparative results (cf. Table II), these findings suggest that the observed overall preference for interval-valued responses is explained by their significantly greater capacity for effective communication.

This paper represents a valuable extension of evidence concerning the actual efficiency of the ellipse-based interval elicitation method, relative to a prevalent conventional alternative. However, work in this area is by no means finished. It will be important to further develop this line of investigation, by comparing efficiency of interval elicitation techniques, and alternatives, using objective measures of workload, across a wider variety of tasks, and with broader experimental

samples—including to establish whether different groups hold different response preferences. In the future, it will also be vital to compare the ellipse response mode, in terms of both relative workload and real-world information capture, against alternatives of greater complexity—e.g., qualitative interviews, and methods of eliciting more complex distributions [23]–[25], such as the FRS [26]–[29], and SHELF [30], [31]—to better establish and inform the putative 'effort-information trade-off'.

Finally, initial results suggest that the capture in particular of conjunctive intervals, rather than disjunctive, offers a pathway to capture human insight that avoids some of the pitfalls of disjunctive sets [5]. The latter, which include the commonly used (elicitation of) confidence intervals, rely on respondents' understanding of the statistical underpinnings of such intervals—whereas conjunctive intervals enable respondents to rely on their intuitive ability to reason about (conjunctive) sets. This is of particular relevance in explainable AI and, more broadly, human-AI interaction, where effective information exchange between human and machine is paramount.

To summarise, an absence of methods for easily obtaining interval-valued responses has held back their use in wider research and society. This study provides evidence that interval elicitation using ellipses is not only effective (cf. [3]–[5]), but efficient. Results found that respondents preferred to give interval rather than point estimates, because they perceived them as no more difficult to provide, while permitting more effective communication of their desired responses. We hope that this will encourage broader engagement with, research into, and ultimately uptake of efficient interval-valued response capture, modelling, analysis, and AI.

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] 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.

[4] 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.

[5] Z. Ellerby, C. Wagner, and S. Broomell, "Capturing richer information–on establishing the validity of an interval-valued survey response mode," *arXiv preprint arXiv:2009.08456*, 2021.

[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] F. Liu and J. M. Mendel, "Encoding words into interval type-2 fuzzy sets using an interval approach," *IEEE transactions on fuzzy systems*, vol. 16, no. 6, pp. 1503–1521, 2008.

[11] 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.

[12] J. McCulloch, Z. Ellerby, and C. Wagner, "On the relationship between similarity measures and thresholds of statistical significance in the context of comparing fuzzy sets," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 8, pp. 1785–1798, 2019.

[13] 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.

[14] S. Miller, C. Wagner, J. M. Garibaldi, and S. Appleby, "Constructing general type-2 fuzzy sets from interval-valued data," in *2012 IEEE International Conference on Fuzzy Systems*, IEEE, 2012.

[15] C. Wagner, S. Miller, and J. M. Garibaldi, "Similarity based applications for data-driven concept and word models based on type-1 and type-2 fuzzy sets," in *2013 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*, IEEE, 2013.

[16] 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, 2014.

[17] D. Wu, J. M. Mendel, and S. Coupland, "Enhanced interval approach for encoding words into interval type-2 fuzzy sets and its convergence analysis," *IEEE Transactions on Fuzzy Systems*, vol. 20, no. 3, pp. 499–513, 2011.

[18] R. E. Moore, R. B. Kearfott, and M. J. Cloud, *Introduction to interval analysis*. SIAM, 2009.

[19] R. E. Moore, *Interval analysis*, vol. 4. Prentice-Hall Engle. Cliffs, 1966.

[20] H. T. Nguyen, V. Kreinovich, B. Wu, and G. Xiang, *Computing statistics under interval and fuzzy uncertainty*, vol. 130. Springer, 2012.

[21] 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.

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

[23] W. B. de Bruin, C. F. Manski, G. Topa, and W. Van der Klaauw, "Measuring consumer uncertainty about future inflation," tech. rep., Staff Report, 2009.

[24] D. E. Morris, J. E. Oakley, and J. A. Crowe, "A web-based tool for eliciting probability distributions from experts," *Environmental Modelling & Software*, vol. 52, 2014.

[25] A. Speirs-Bridge, F. Fidler, M. McBride, L. Flander, G. Cumming, and M. Burgman, "Reducing overconfidence in the interval judgments of experts," *Risk Analysis: An International Journal*, vol. 30, no. 3, pp. 512–523, 2010.

[26] T. Hesketh, R. Pryor, and B. Hesketh, "An application of a computerized fuzzy graphic rating scale to the psychological measurement of individual differences," *International Journal of Man-Machine Studies*, vol. 29, no. 1, pp. 21–35, 1988.

[27] B. Hesketh, K. McLachlan, and D. Gardner, "Work adjustment theory: An empirical test using a fuzzy rating scale," *Journal of Vocational Behavior*, vol. 40, no. 3, pp. 318–337, 1992.

[28] M. A. Lubiano, S. d. l. R. de Sáa, M. Montenegro, B. Sinova, and M. Á. Gil, "Descriptive analysis of responses to items in questionnaires. why not using a fuzzy rating scale?," *Information Sciences*, vol. 360, pp. 131–148, 2016.

[29] P. Quirós, J. M. Alonso, and D. P. Pancho, "Descriptive and comparative analysis of human perceptions expressed through fuzzy rating scale-based questionnaires," *International Journal of Computational Intelligence Systems*, vol. 9, no. 3, pp. 450–467, 2016.

[30] J. P. Gosling, "Shelf: the sheffield elicitation framework," in *Elicitation*, pp. 61–93, Springer, 2018.

[31] A. O'Hagan, "Expert knowledge elicitation: subjective but scientific," *The American Statistician*, vol. 73, no. sup1, pp. 69–81, 2019.

[32] 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.

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

[34] 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.

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

[36] J. Brooke *et al.*, "Sus-a quick and dirty usability scale," *Usability evaluation in industry*, vol. 189, no. 194, pp. 4–7, 1996.

[37] P. D. Allison, *Multiple regression: A primer*. Pine Forge Press, 1999.