The IDEA Protocol and performance weighting for expert elicitation

Victoria Hemming
Centre for Environmental and Economic Research
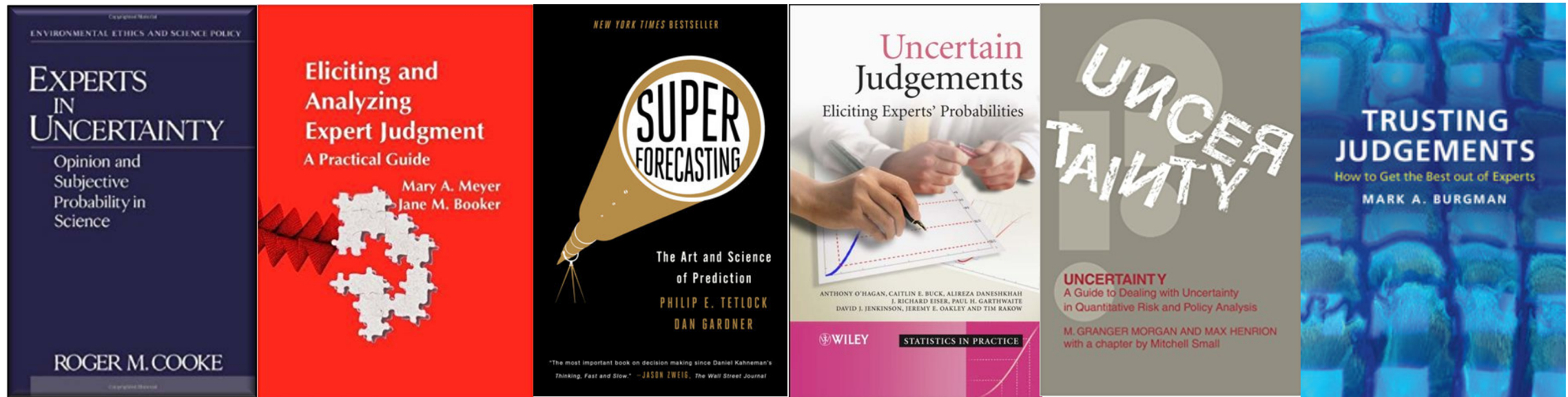+
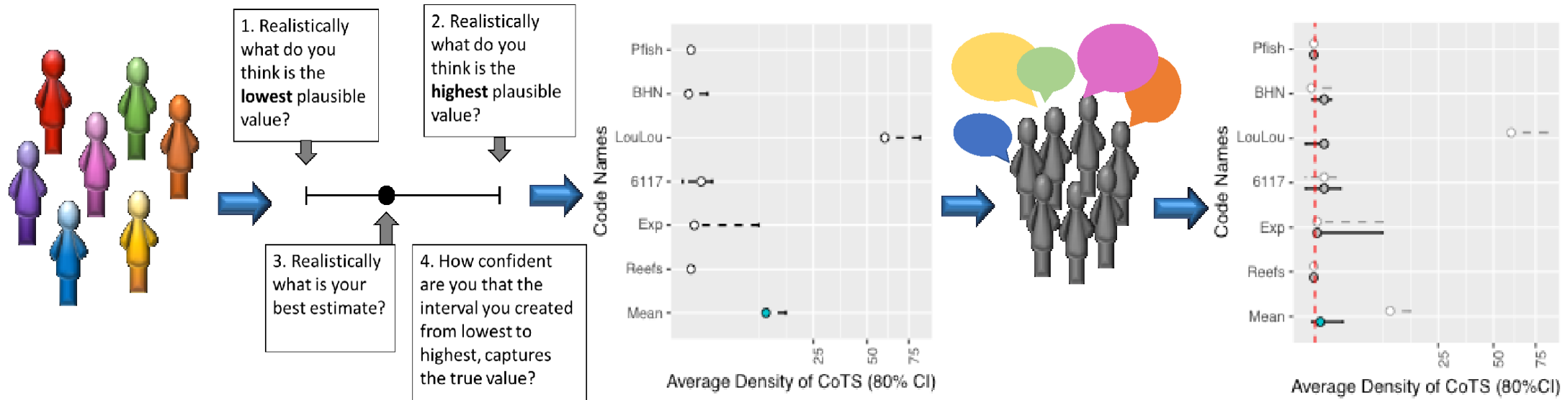Centre of Excellence for Biosecurity Research
The University of Melbourne
vhemming@unimelb.edu.au

Structured Elicitation Protocols

The IDEA protocol (Investigate, Discuss, Estimate, Aggregate)

1. Realistically what do you think is the **lowest** plausible value?

2. Realistically what do you think is the **highest** plausible value?

3. Realistically what is your best estimate?

4. How confident are you that the interval you created from lowest to highest, captures the true value?

1. Recruit a diverse group of experts.

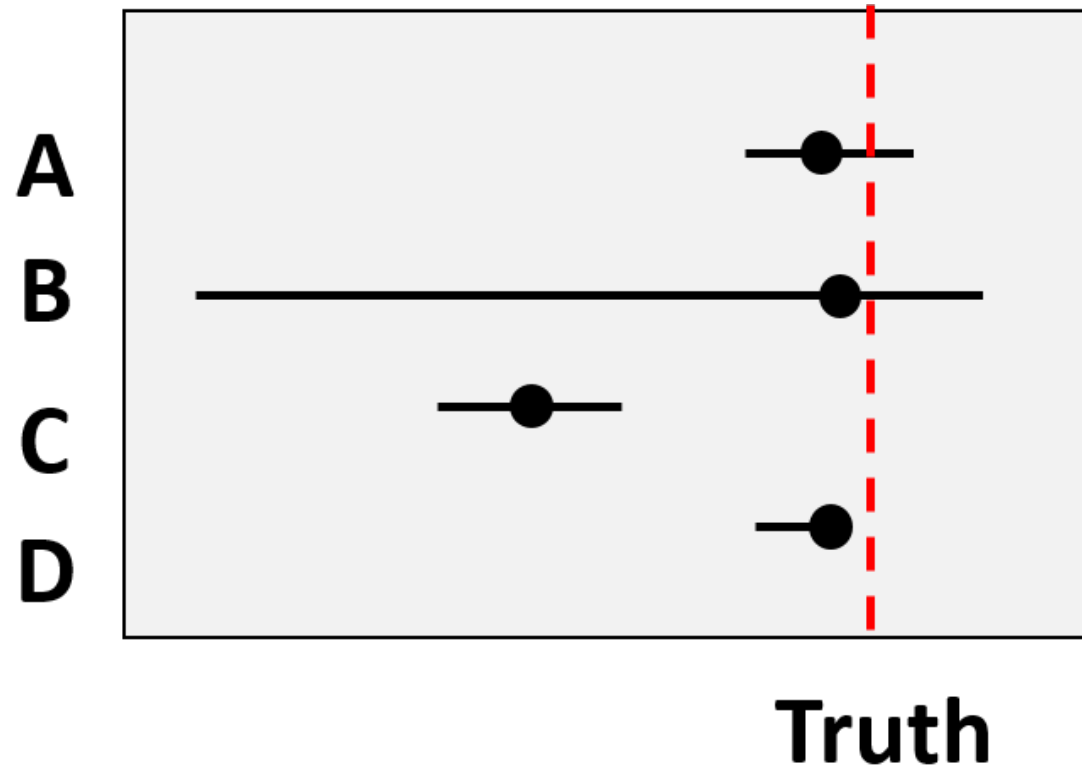2. Experts **INVESTIGATE** the problem independently. Then provide a private, initial and anonymous estimate.

3. Aggregate estimates. Provide feedback to individuals.

4. Facilitated **DISCUSSION**

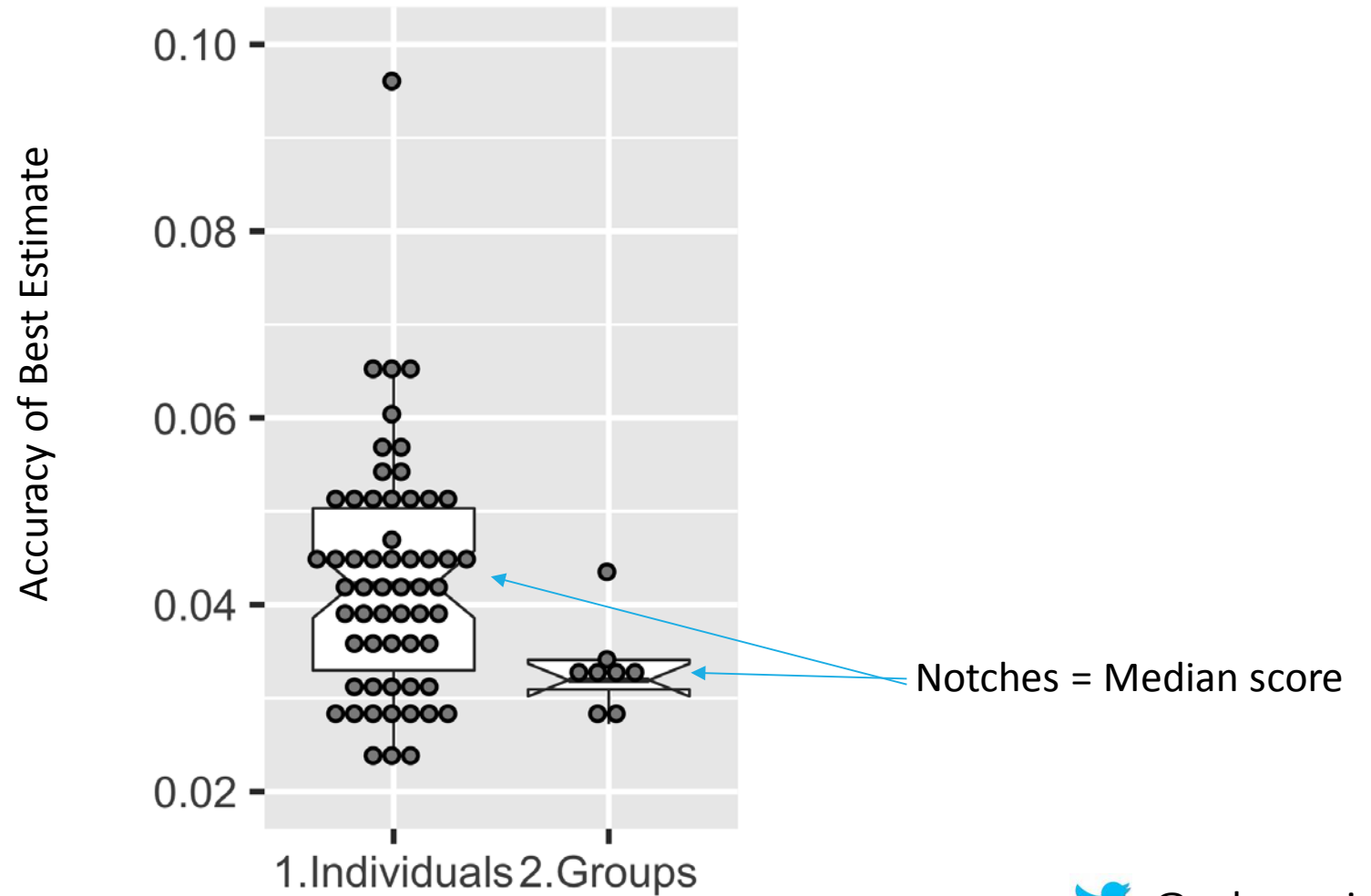5. Experts provide second anonymous **ESTIMATE.**

AGGREGATE estimates

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,** e0198468.

Scoring interval judgements

The wisdom of the crowd

The average of the group

Worse

Better

Accuracy of Best Estimate

Notches = Median score

1.Individuals 2.Groups

@v_hemming

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,** e0198468.

Self-rating and performance?

Worse / Better (Accuracy axis)

Lower → Higher (Self-rating axis)
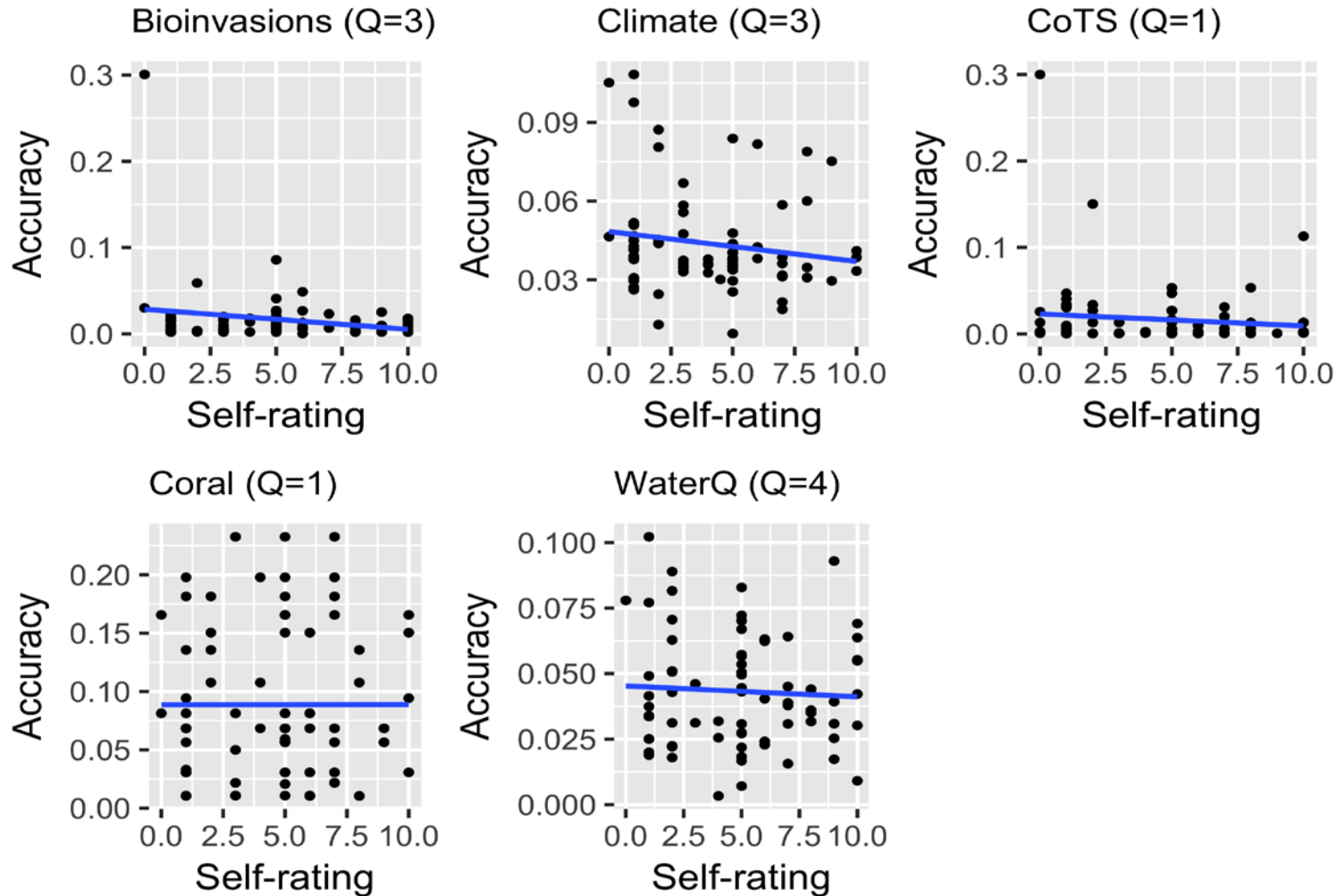
Spearman's rank correlations between -.01 and -0.17.

p> 0.05

@v_hemming

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,** e0198468.

Peer recommendation

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,** e0198468.
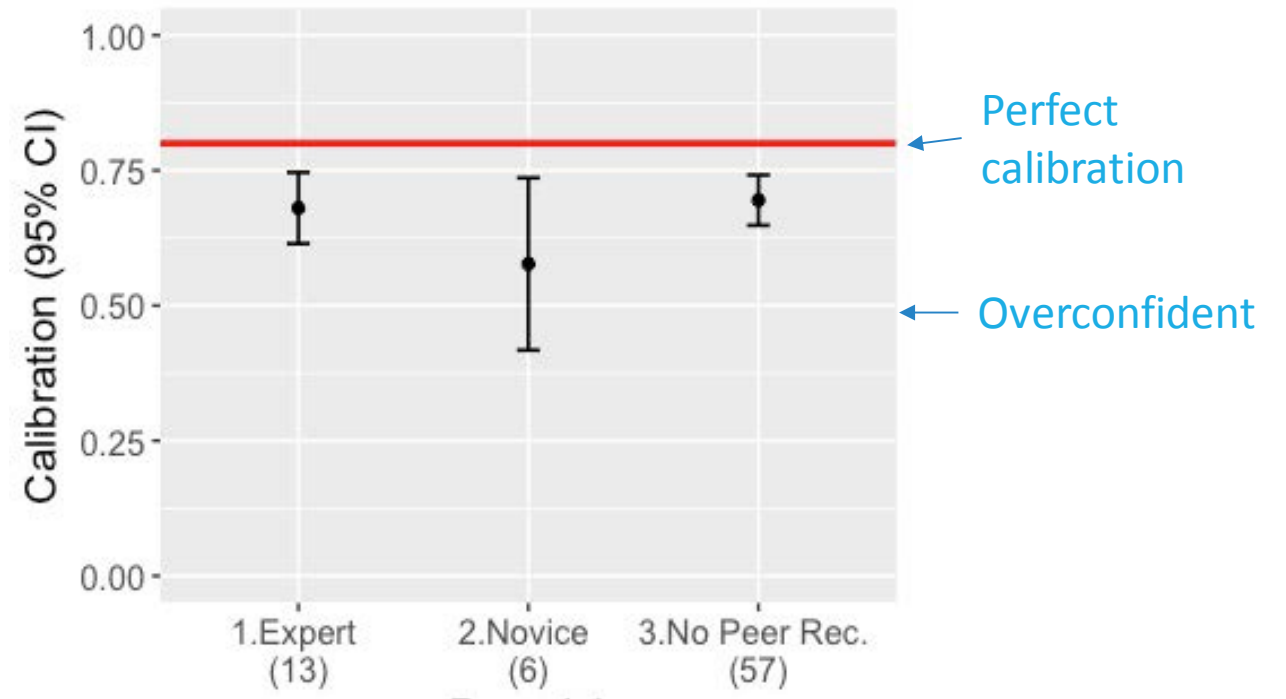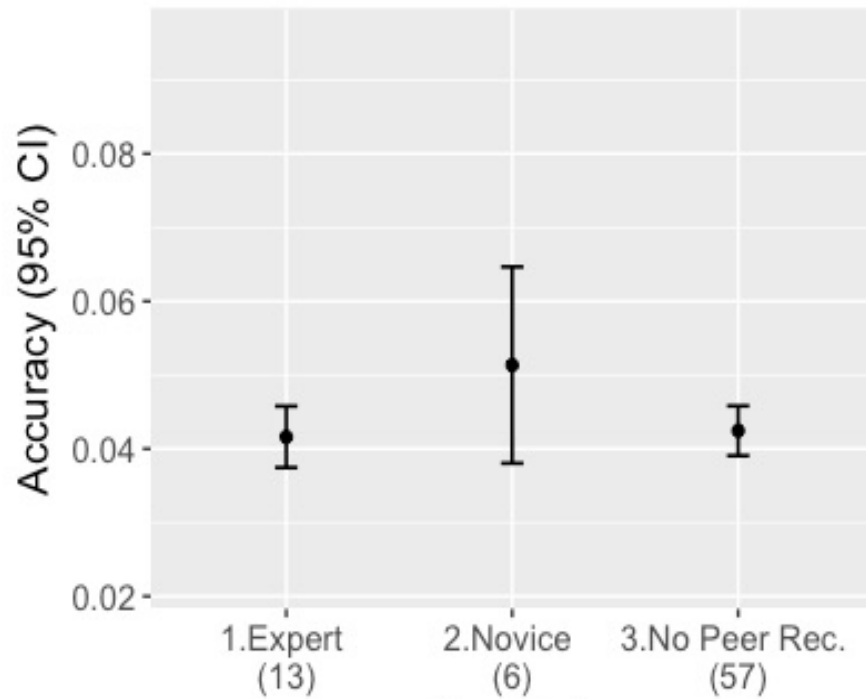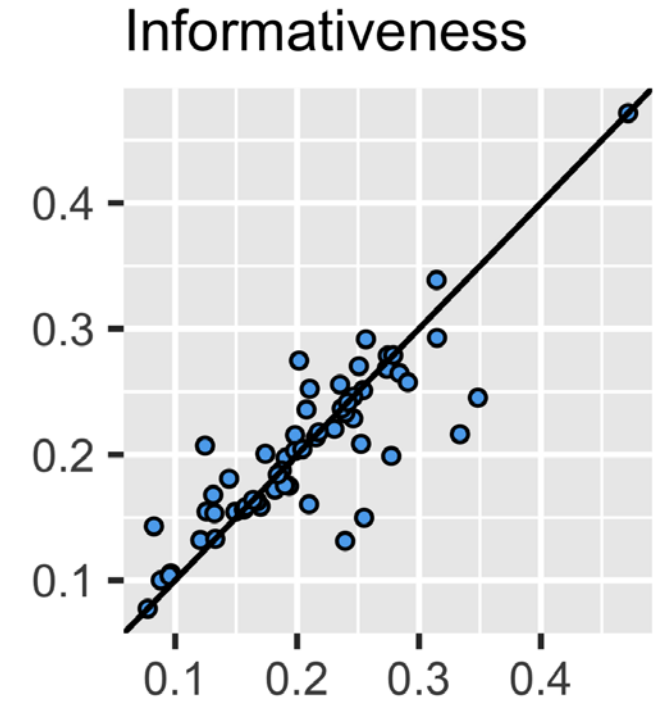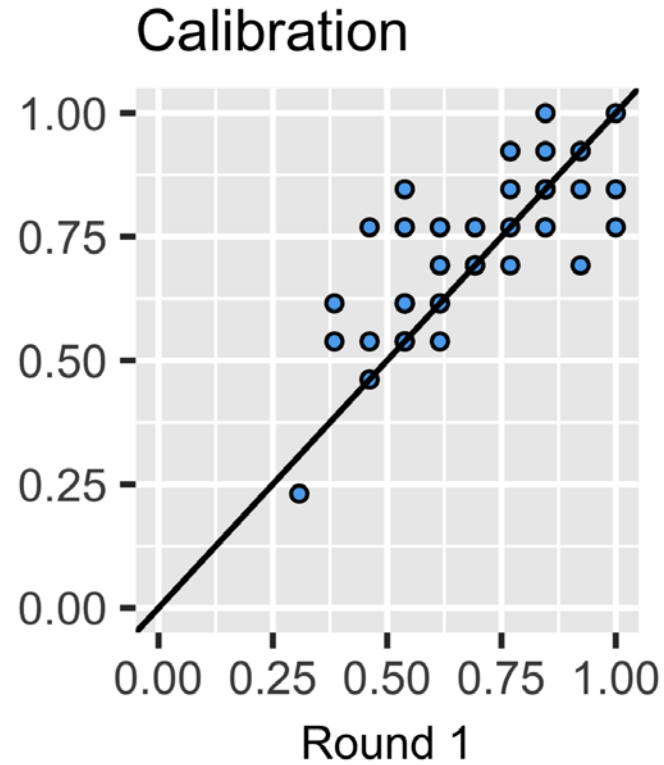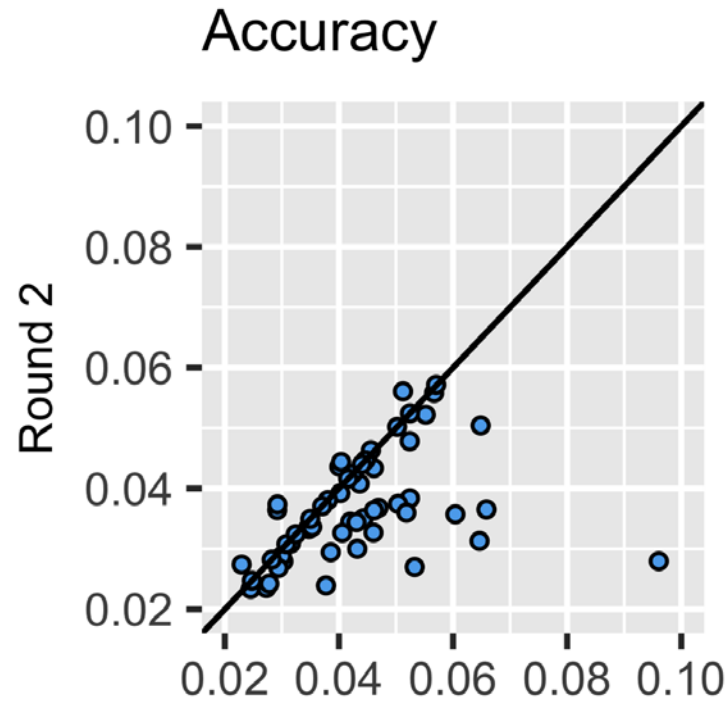
Discussion + Round 2
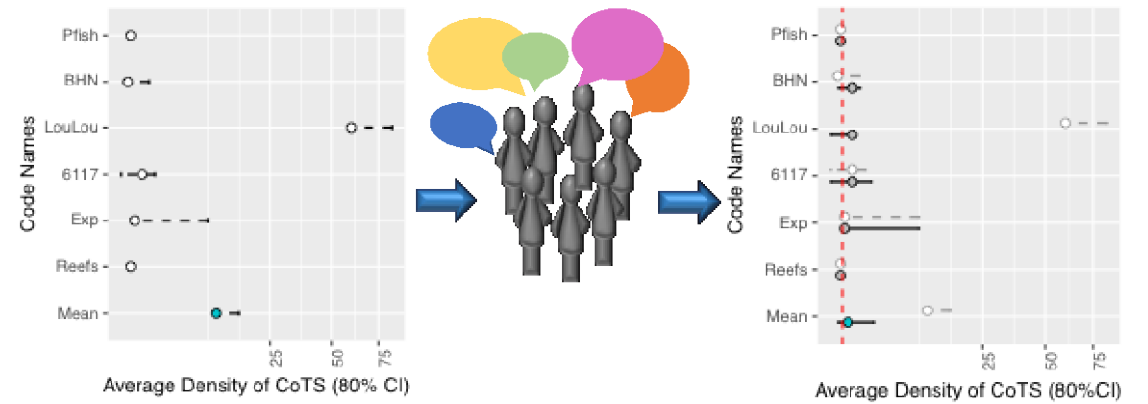
Estimates improved in Round 2

@v_hemming

Results repeated

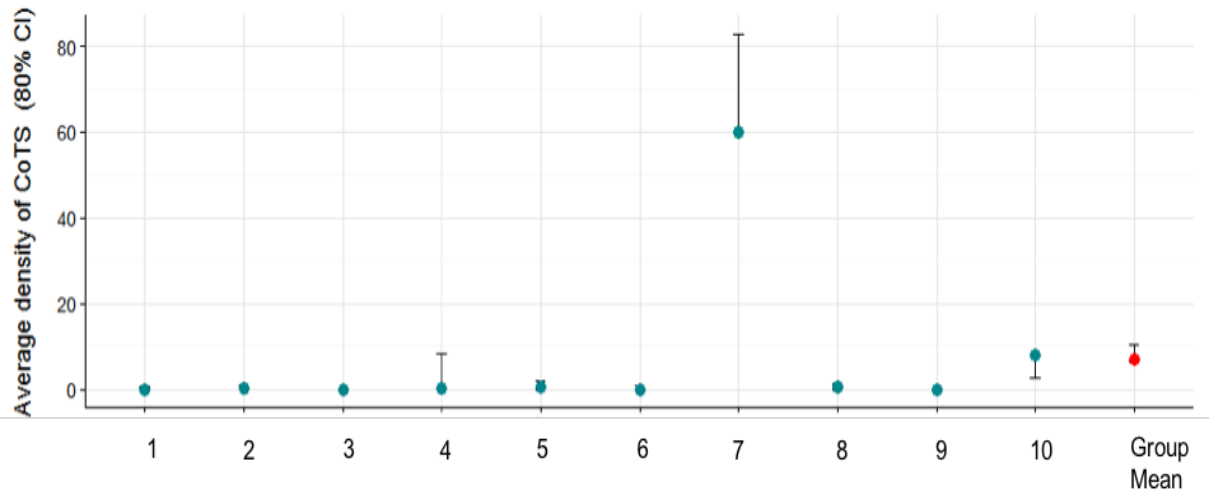# A case study from engineering

This slide has been removed as I don't have permission to share it publicly at this time.

(Hemming, Hanea et al. in revision)

Additional benefit: Rationales

Average density of CoTS at Rib Reef

	1	2	3	4	5	6	7	8	9	10	Group Mean
Best	0.00	0.20	0.10	0.20	0.50	0.05	60.00	0.50	0.10	8.00	6.96
Low	0.00	0.00	0.05	0.18	0.06	0.01	60.00	0.00	0.05	2.67	6.30
High	0.49	1.05	0.12	8.52	1.83	0.89	82.86	1.30	0.15	8.00	10.52

Name	Comments	Round / Date
Participant 2	COTS still seem to be only sporadically present in the Innisfail sector north and upstream of the Rib Reef.	Round 1
Participant 4	Based on data from link given	Round 1
Participant 8	It appears that this is a very broad technique that could be biased by the trained eye of the diver and how conspicuous is the organism.	Round 1
Facilitator	Some good comments here. I'd like to hear from people at the lower and higher ends of this spectrum. Can you elaborate on your reasoning?	Round 1
Participant 3	COTS are moving south but the numbers (as estimated by the LTMP technique) were still very low in 2015. I expect an increase over the 2015 counts (which were 0.05 per tow according to the web page), but not by >10-times	Discussion 21/03/2016
Participant 7	Excuse me, but fortunately I was wrong to write 60. Whereas the percentage of coral cover is around 40, and analyzing the data, I correct my answer: better value 0.6 and lowest 0.06.	Discussion 22/03/2016
Participant 10	The COTS are traveling down the GBR. I thought Rib Reef was closer to Innisfail rather than Townsville on reviewing I would lower my best guess to 4	Discussion 29/03/2016

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,** e0198468.

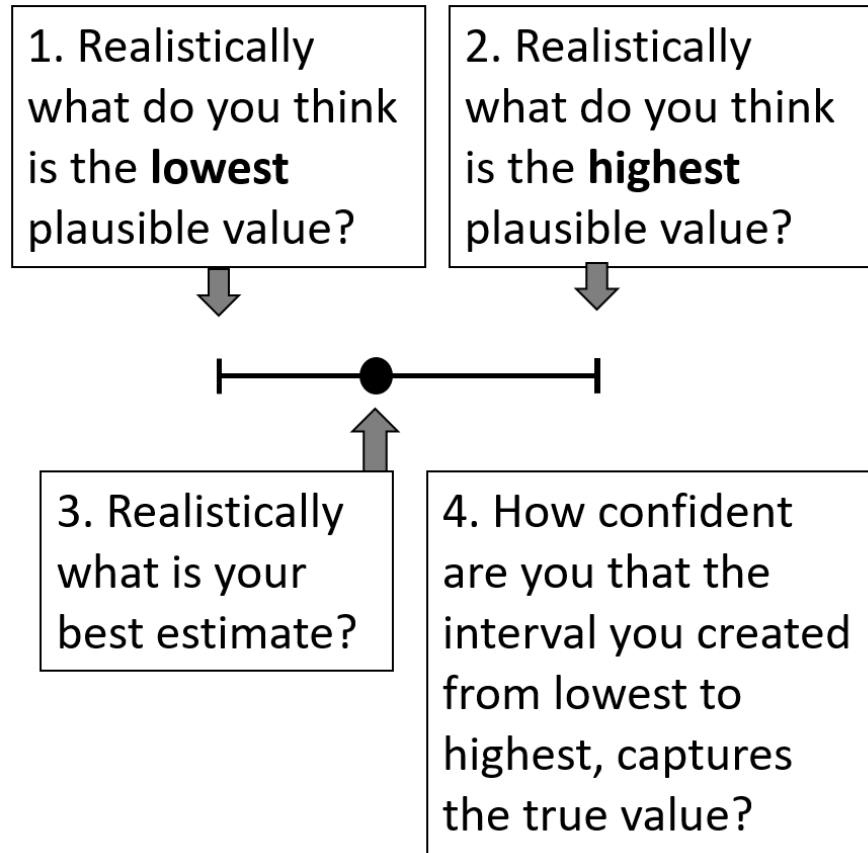# Additional benefit: Flexible elicitation formats

Where is it being applied?

- Australia's Biosecurity
- IUCN Red List of Ecosystems
- Biodiversity offsets
- UK Food security
- CIA research on judgement
- Australian Department of Defence procurement
- New Zealand seismic models
- Koala research priorities in NSW, Australia
- Monitoring for prescribed burning and fuel preparation in River Red Gum forests

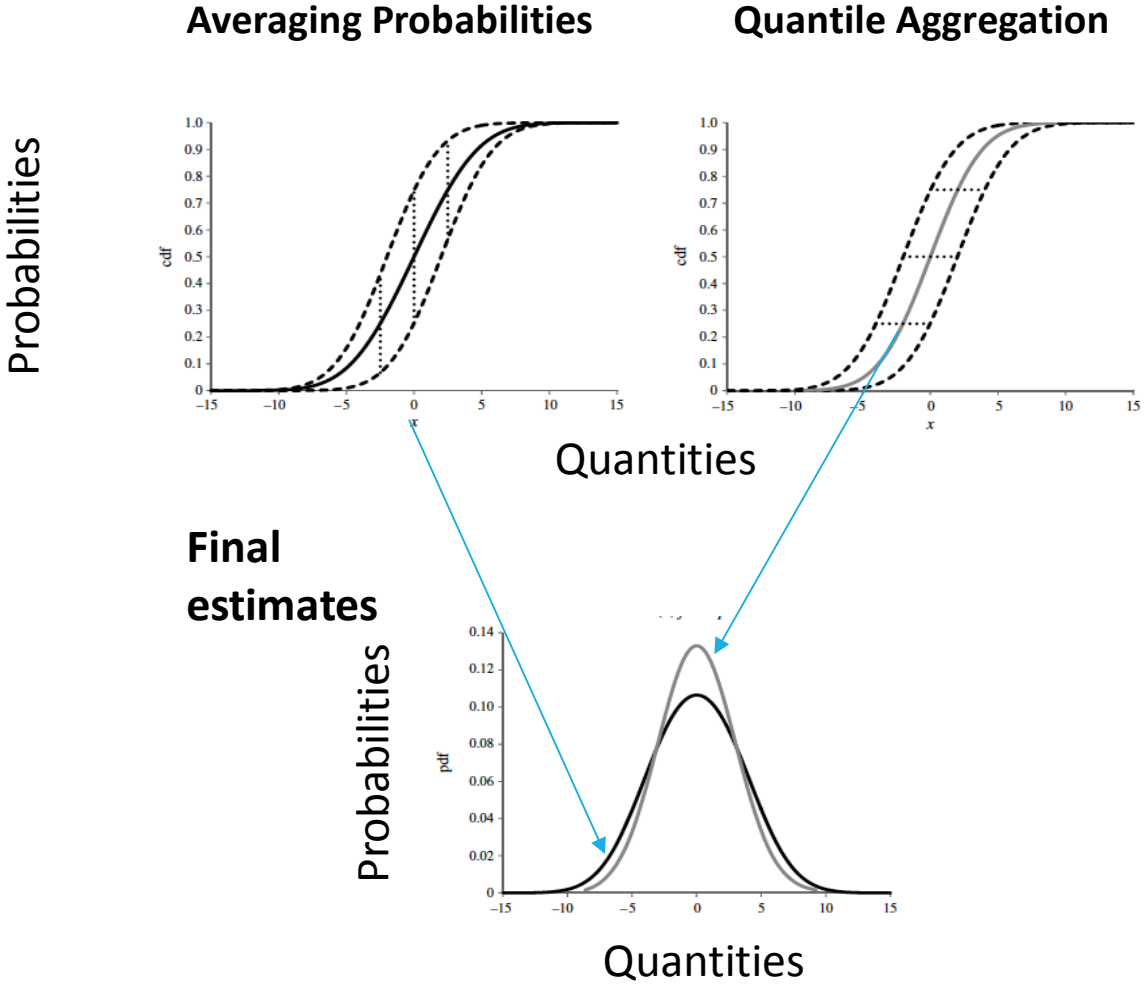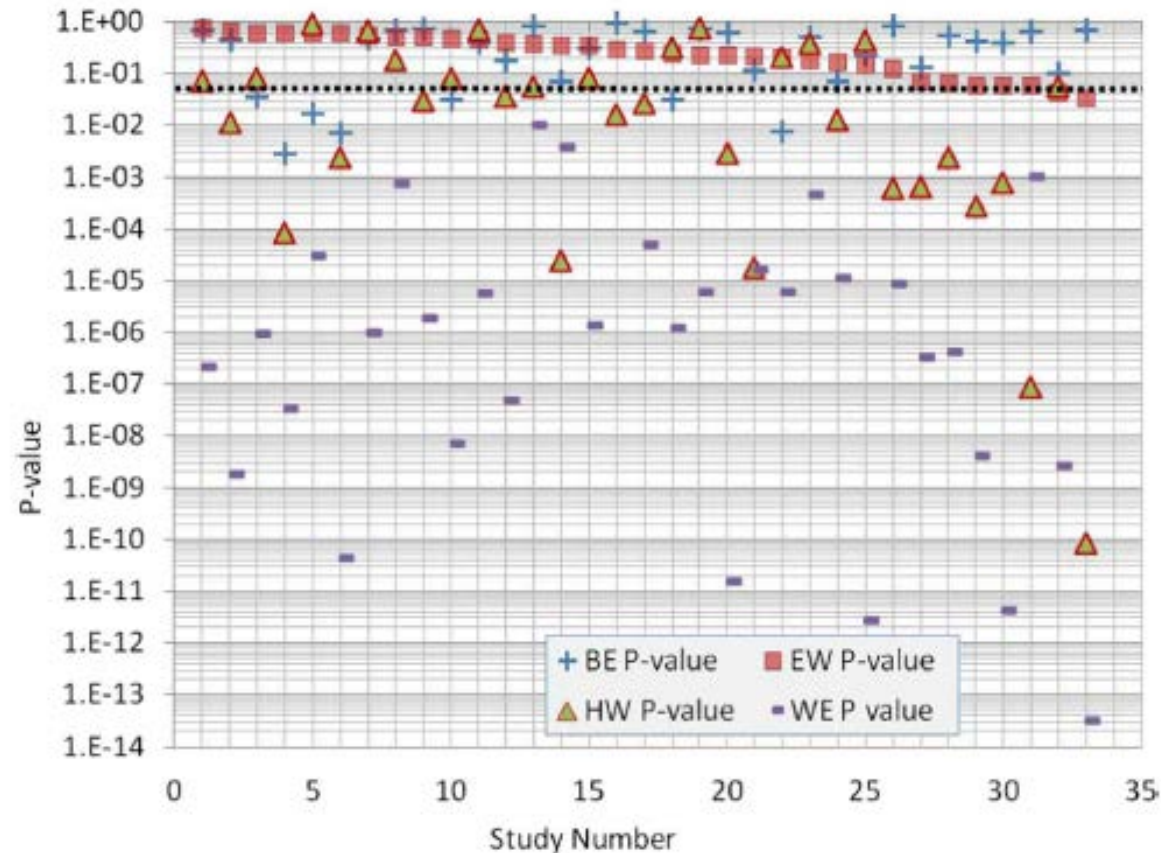# Improvement via aggregation methods?

Averaging Quantiles

| 1. Realistically what do you think is the **lowest** plausible value? | 2. Realistically what do you think is the **highest** plausible value? |

| 3. Realistically what is your best estimate? | 4. How confident are you that the interval you created from lowest to highest, captures the true value? |

Expert	5th (lower)	50th (best)	95th (upper)
1	2	12	34
2	4	15	50
3	7	9	40
4	20	22	23
Average	8.25	14.5	36.75

Averaging Probabilities versus Quantiles

Averaging Probabilities   **Quantile Aggregation**

Final estimates

Lichtendahl Jr, K.C., Grushka-Cockayne, Y. & Winkler, R.L. (2013).

Does it matter?

In only 18 of the 33 studies averaging quantiles is statistically accurate at the 5% level.

Conclusion:
"averaging quantiles" is still used by unwary practitioners, while an elementary performance analysis could easily predict its strong penchant for overconfidence". Colson 2017.

Colson, A.R. & Cooke, R.M. (2017)

Quantile Aggregation vs Linear Pooling

Hemming, V., Hanea, A.M., Burgman, M.A.
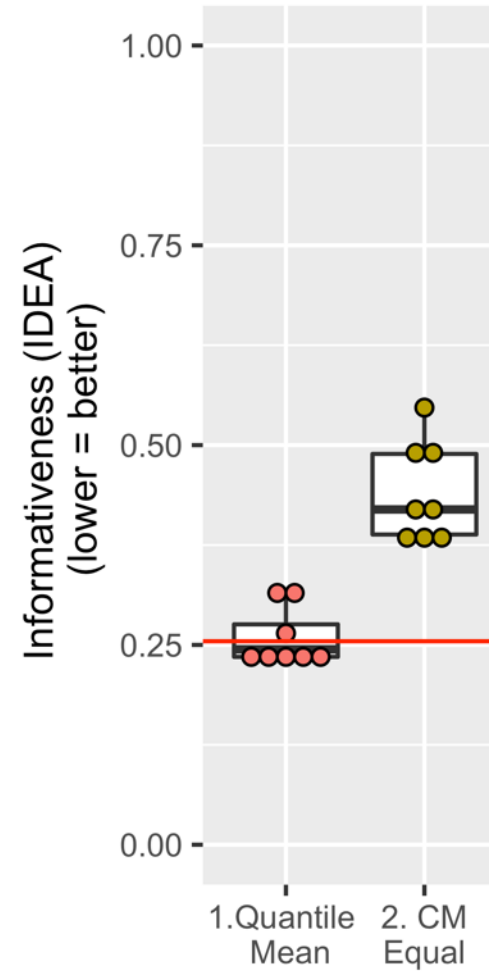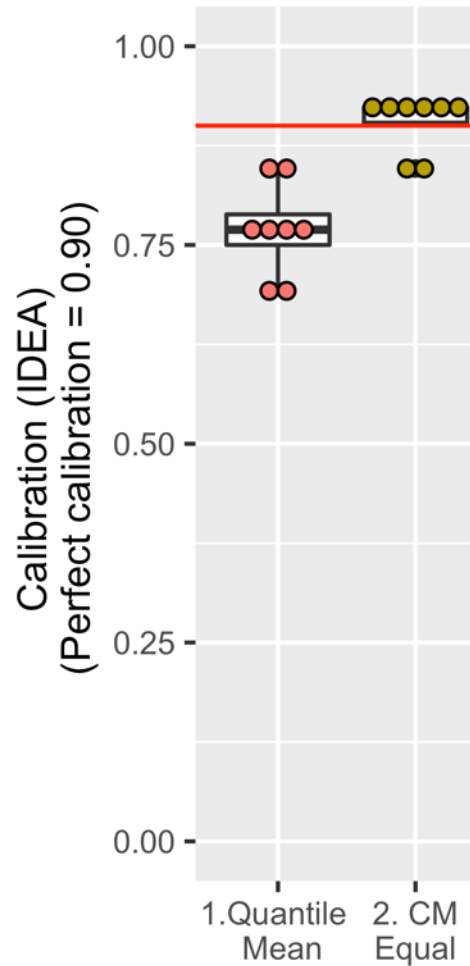
Conclusion

Quantile aggregation = overconfident

Quantile aggregation = informative

A trade-off (i.e. value judgement) is required

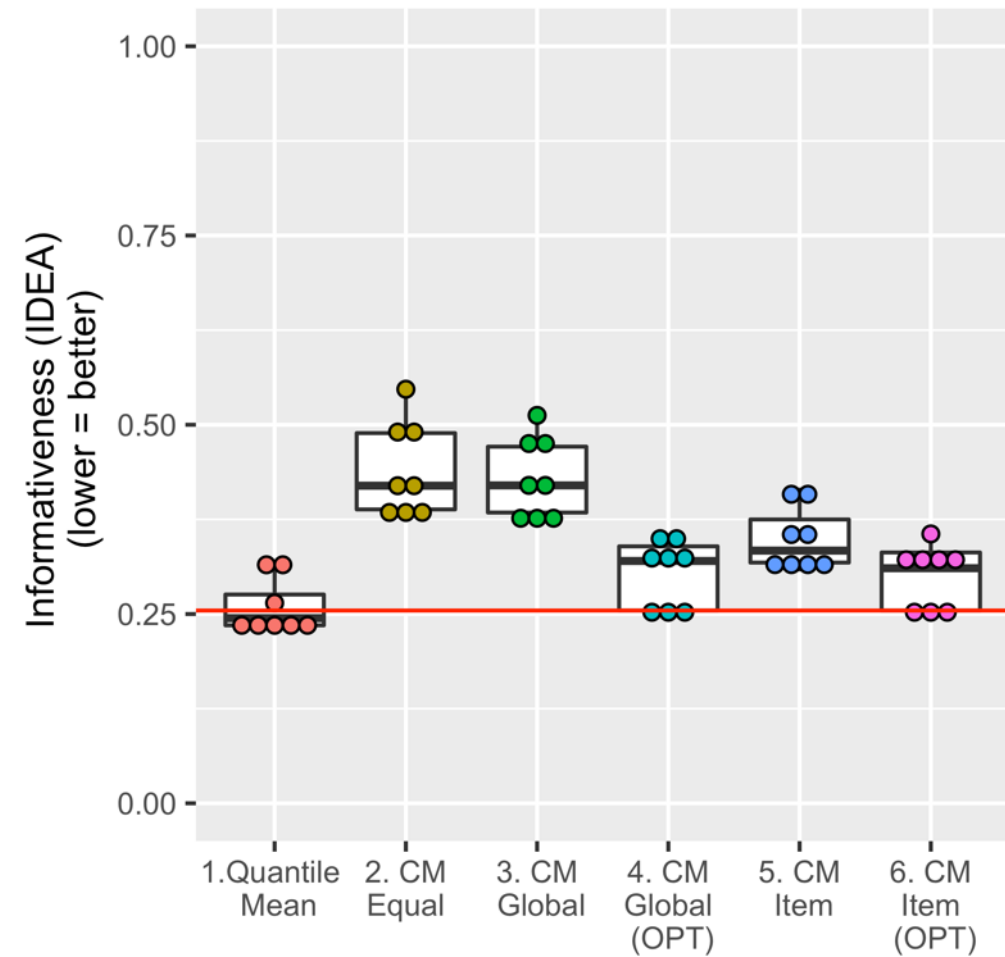# Performance weighting
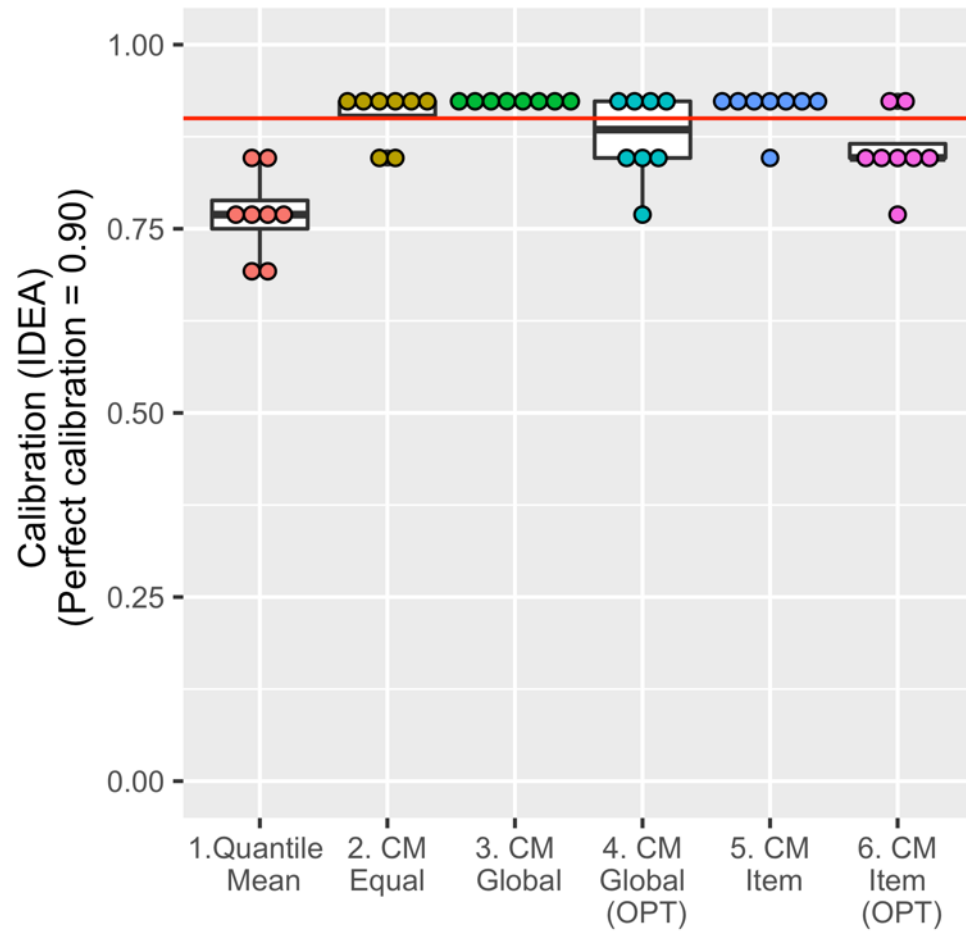
Classical Model: Calibration

Very crudely, it answers questions like "how likely is it that at least 8 out of 10 realizations should fall outside an expert's 90% confidence bands, if each value really had an independent 90% chance of falling inside the bands?"

Expected (0.05, 0.45, 045, 0.05)

Observed (0.1, 0.40, 0.40, 0.1)

SA= 0.83

Colson, A.R. & Cooke, R.M. (2017)

Performance weighting (Classical Model)

Conclusion

Performance weights:

well calibrated +

informative

Unwary
practitioners
should still take
care…

Assumptions have to be made

1. Realistically what do you think is the **lowest** plausible value?

2. Realistically what do you think is the **highest** plausible value?

3. Realistically what is your best estimate?

4. How confident are you that the interval you created from lowest to highest, captures the true value?

Expert B

	<5	5-50	50-95	>95	SA	CA
	1	9	0	0	0.003	0.90
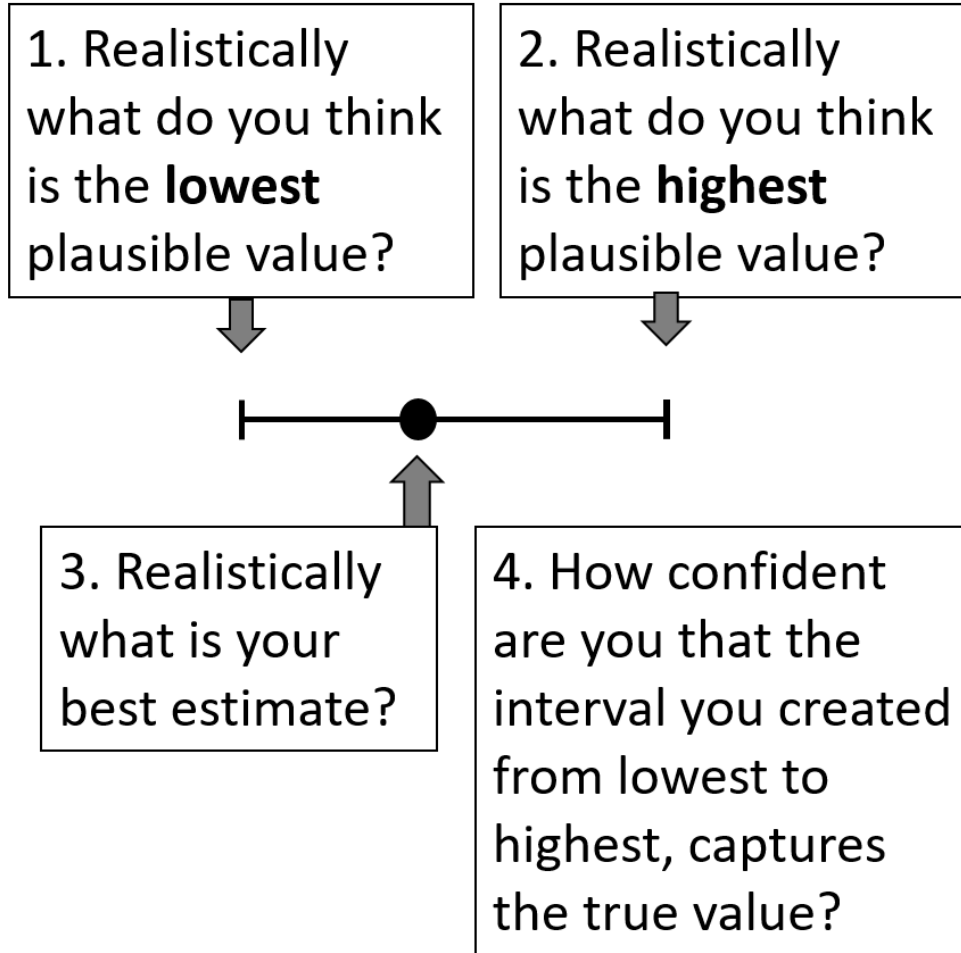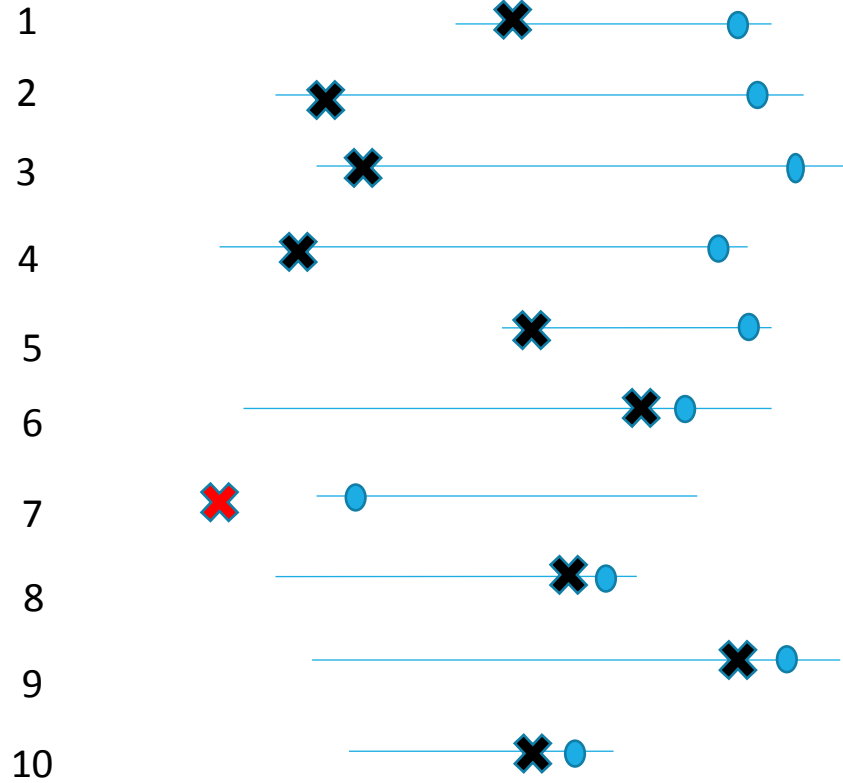
Overconfidence interval judgements:

Observed : 0.90
Expected: 0.90

Overconfidence CM:

Observed: s(0.01, 0.09, 0.00, 0.00)
Expected: p(0.05, 0.45, 0.45, 0.05)

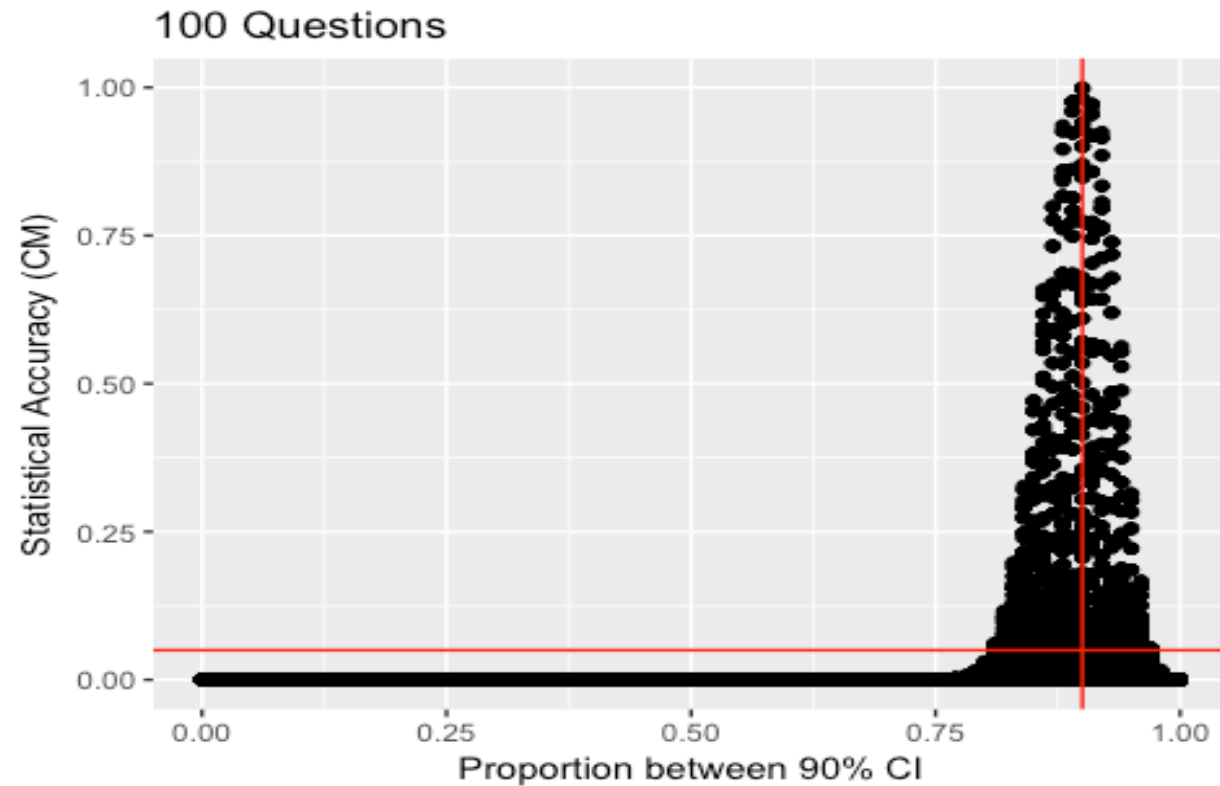# Classical Model: Calibration

Very crudely, it answers questions like "how likely is it that at least 8 out of 10 realizations should fall outside an expert's 90% confidence bands, if each value really had an independent 90% chance of falling inside the bands?"

Expected (0.05, 0.45, 045, 0.05)
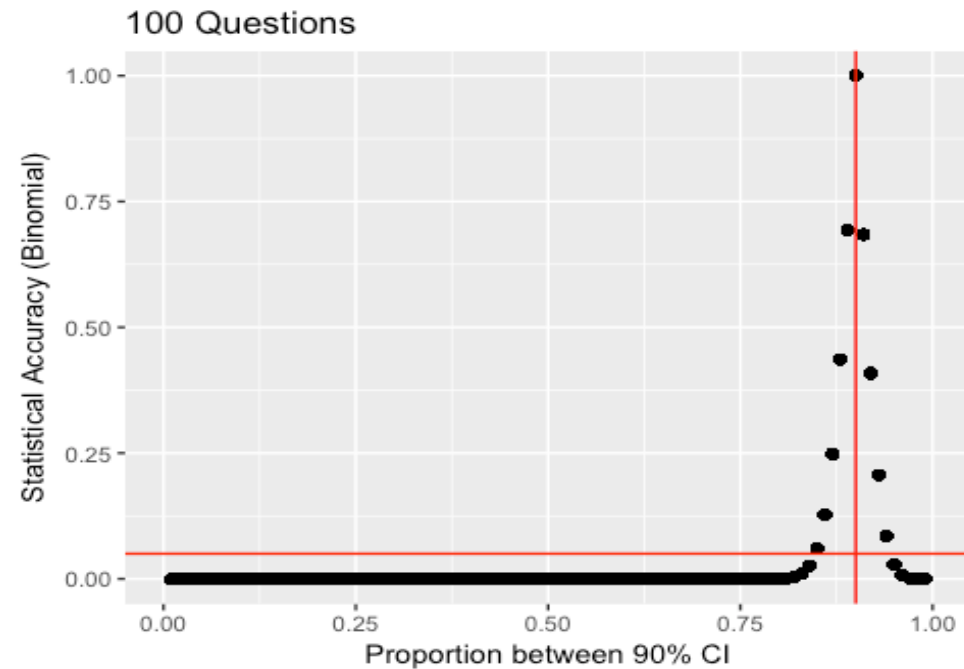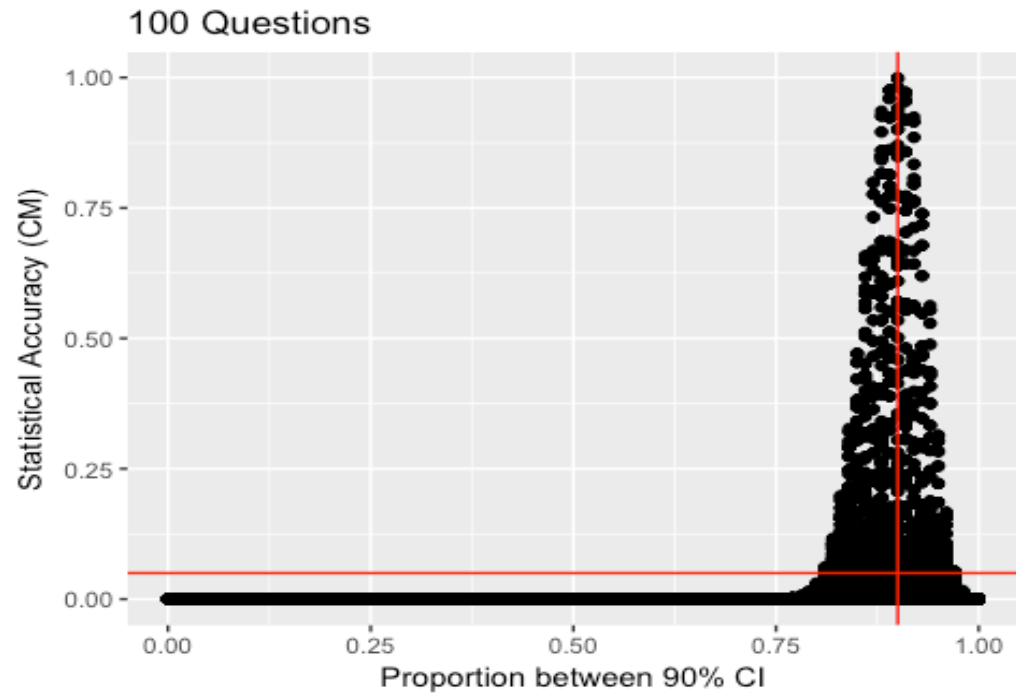
Observed (0.1, 0.40, 0.40, 0.1)

SA= 0.83

Colson, A.R. & Cooke, R.M. (2017)

Differences in calibration

100 Questions

Hemming, V., Hanea, A.M., Burgman, M.A. (in revision).

Possible ways forward

Should we average probabilities of interval judgements?

Create a scoring rule for a Binomial

Hemming, V., Hanea, A.M., Burgman, M.A. (in revision).

Making value judgements in scoring rules explicit
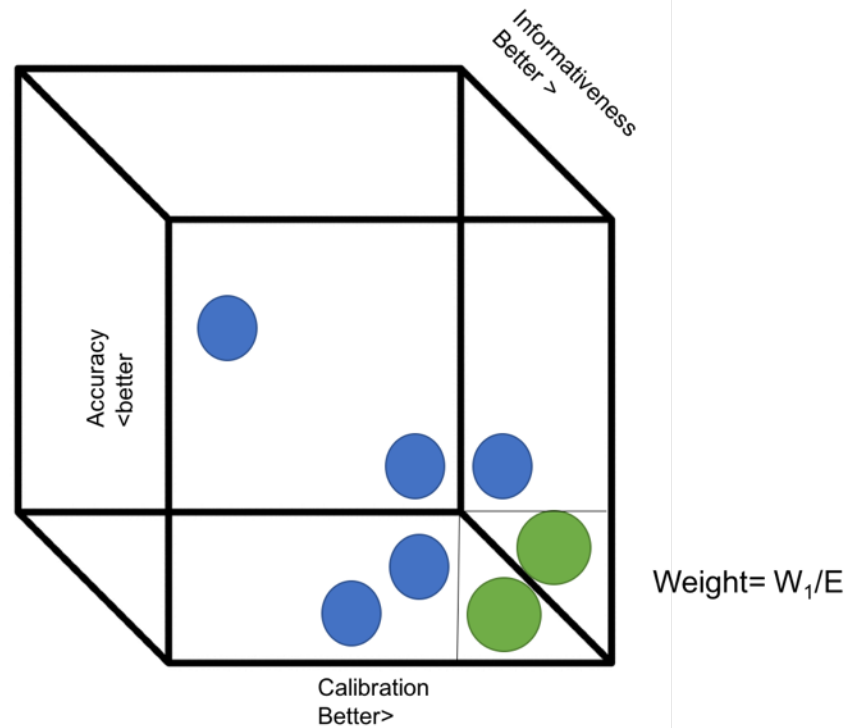
Acknowledgements

Prof. Mark Burgman, Dr. Anca Hanea, Dr. Terry Walshe,
A/Prof Fiona Fidler

The experts / novices that let me test them

Victoria Hemming
Centre of Environmental and Economic Research
The University of Melbourne
hemmingv@student.unimelb.edu.au
https://hemmingresearch.wordpress.com

@v_hemming

References

Burgman MA. Trusting judgements: How to get the best out of experts. Cambridge, United Kingdom: Cambridge University Press; 2015. 203 p.

Colson, A.R. & Cooke, R.M. (2017) Cross validation for the classical model of structured expert judgment. *Reliability Engineering & System Safety,* 163, 109-120.

Hemming, V., Burgman, M.A., Hanea, A.M., McBride, M.F. & Wintle, B.C. (2018a) A practical guide to structured expert elicitation using the IDEA protocol. Methods in Ecology and Evolution, 9, 169-181.

Hemming, V., Walshe, T.V., Hanea, A.M., Fidler, F. & Burgman, M.A. (2018b) Eliciting improved quantitative judgements using the IDEA protocol: A case study in natural resource management. PLoS One, 13, e0198468.

Hemming, V., Hanea, A.M., Burgman, M.A. (in revision) Improving the crowd with a Classical approach to the IDEA protocol.

Lichtendahl Jr, K.C., Grushka-Cockayne, Y. & Winkler, R.L. (2013) Is it better to average probabilities or quantiles? *Management Science,* 59, 1594-1611.