

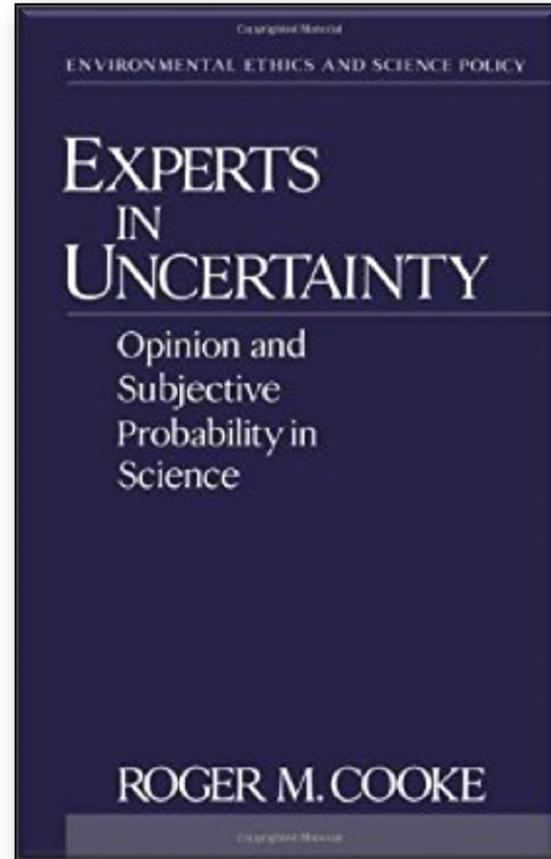
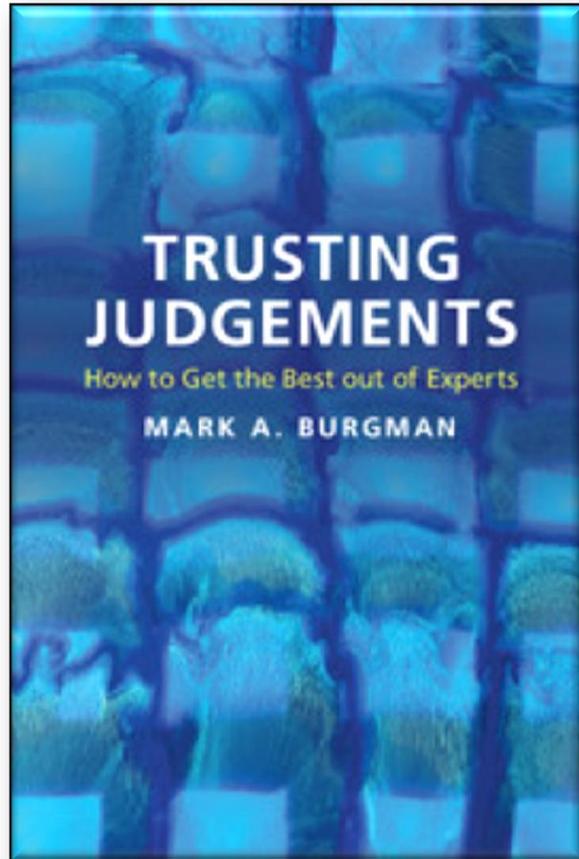
Who to Trust?

Assessing and improving expert judgement within
conservation

Victoria Hemming
PhD Candidate

Centre of Excellence for Biosecurity Risk Analysis
The University of Melbourne
hemmingv@student.unimelb.edu.au

 @v_hemming



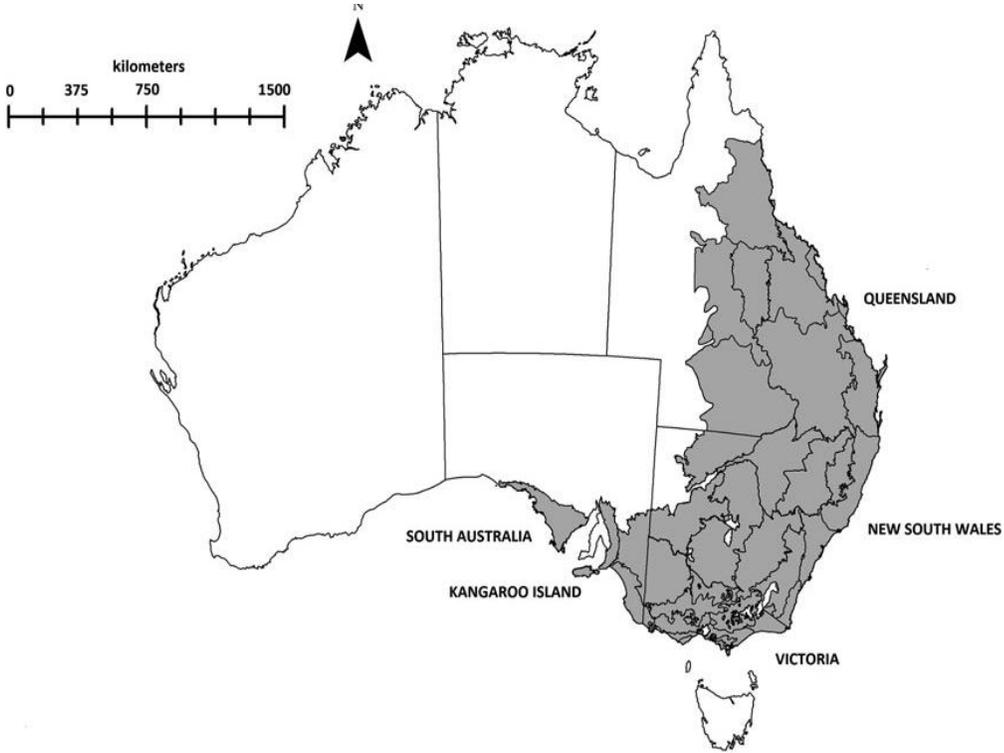
Why Conservation?



Estimating Population Sizes



© Australian Koala Foundation



Cost and Benefit of Management



© Al Hartmann Salt Lake Tribune



© Iki Films Ryan Kohatsu



© MerriCreekManagementCommittee



© Weedsnetwork.com



© Foxnews



Environmental Impact Assessment



© TheAge

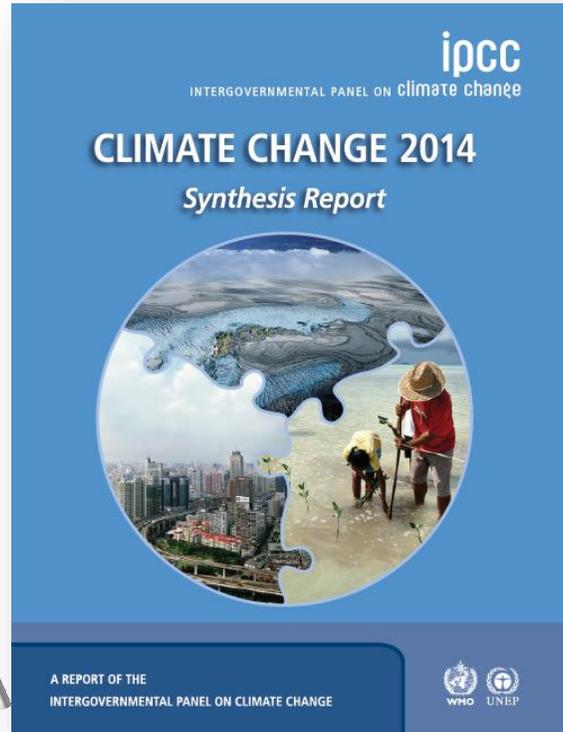
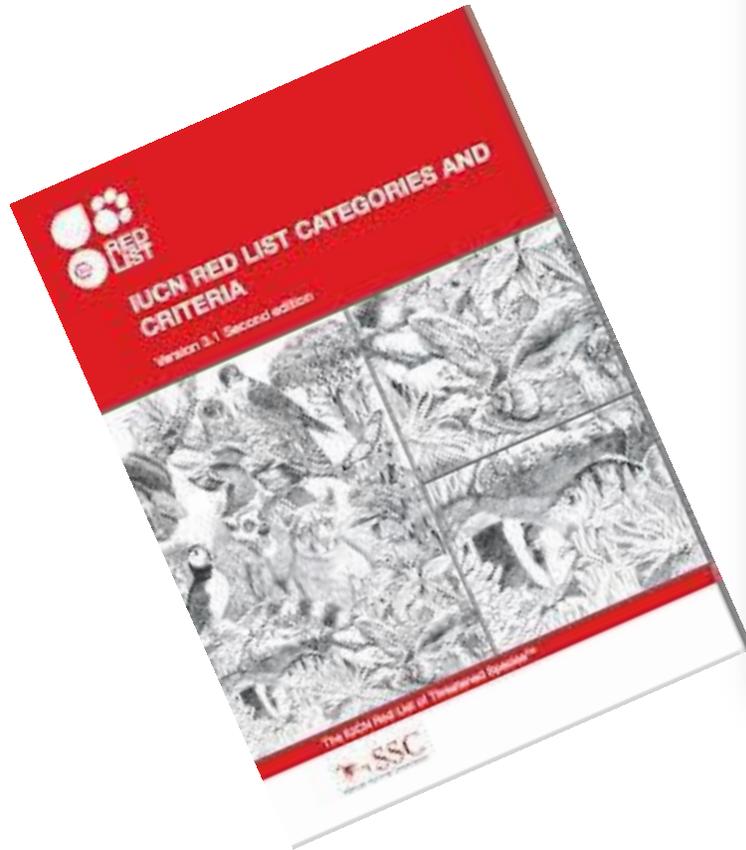


© Craig Abraham-TheAge

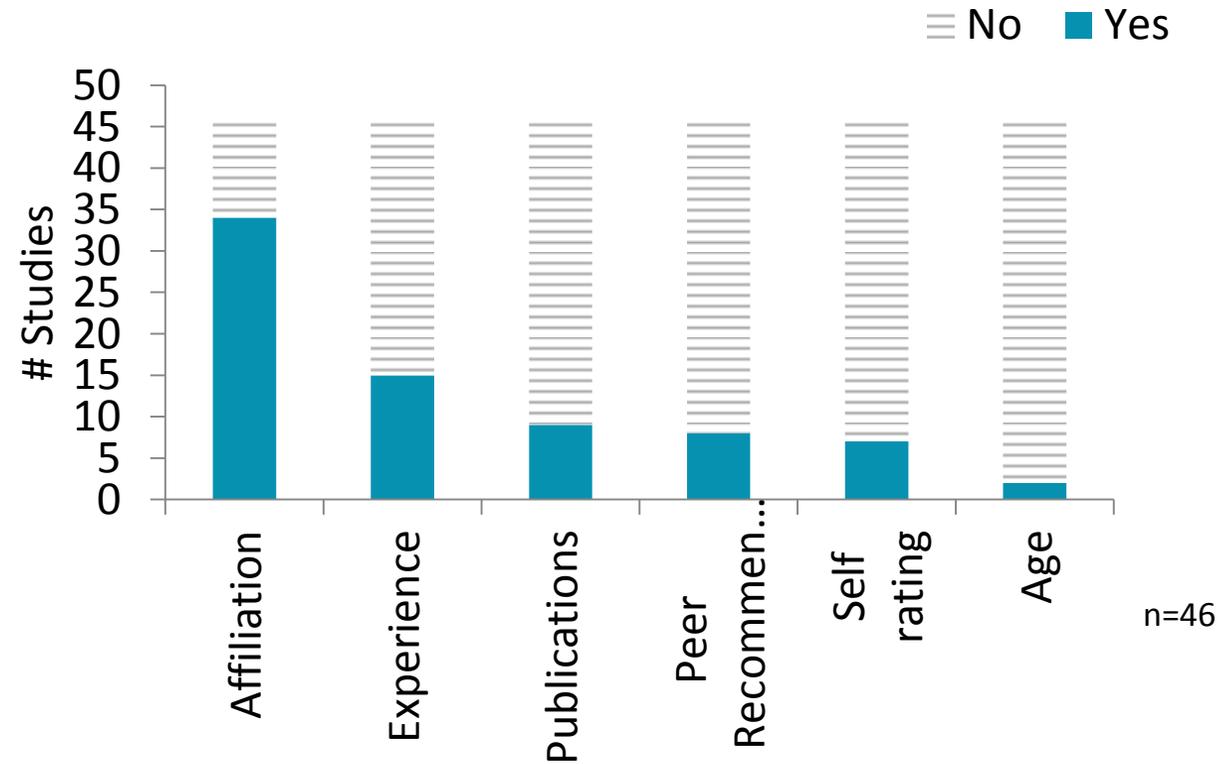


© Framepool

Global Environmental Policy



The experts?



Hemming, V (Draft) "The Reproducibility Crisis of Expert Judgement in Conservation".

Expert Judgement within Conservation

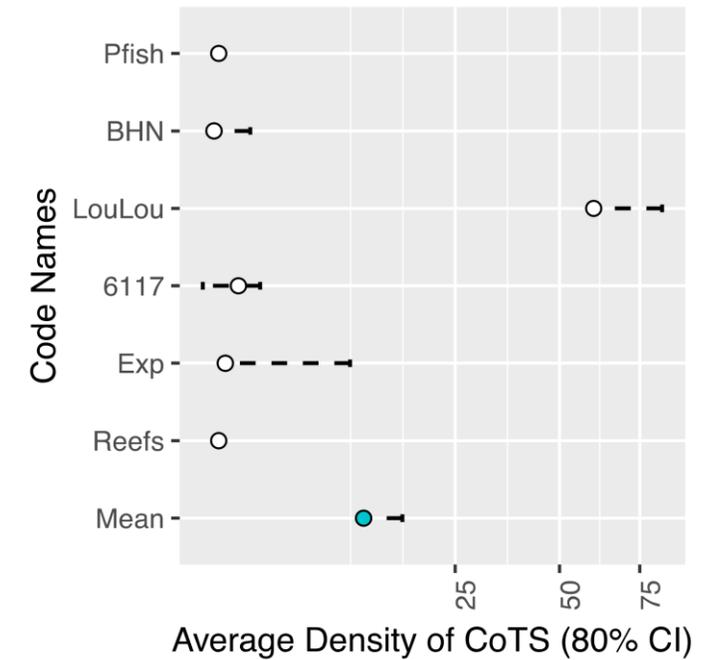
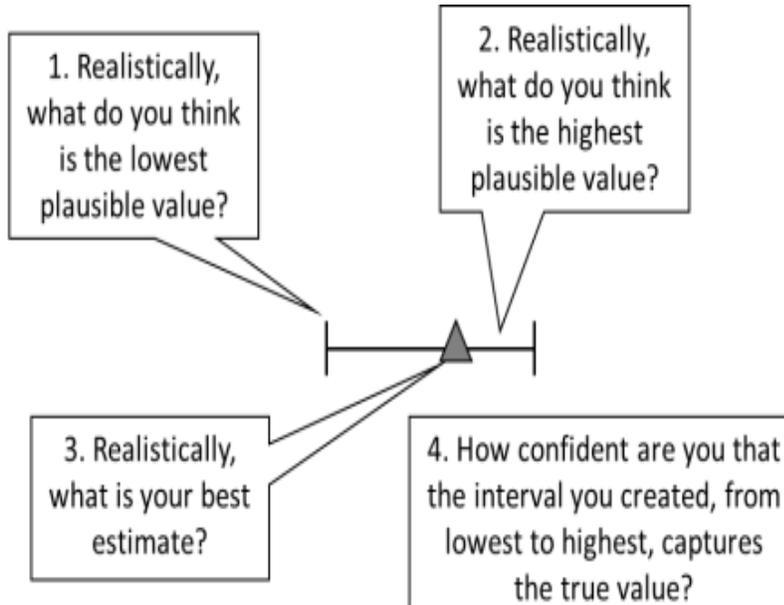


"At last we've reached a consensus!"

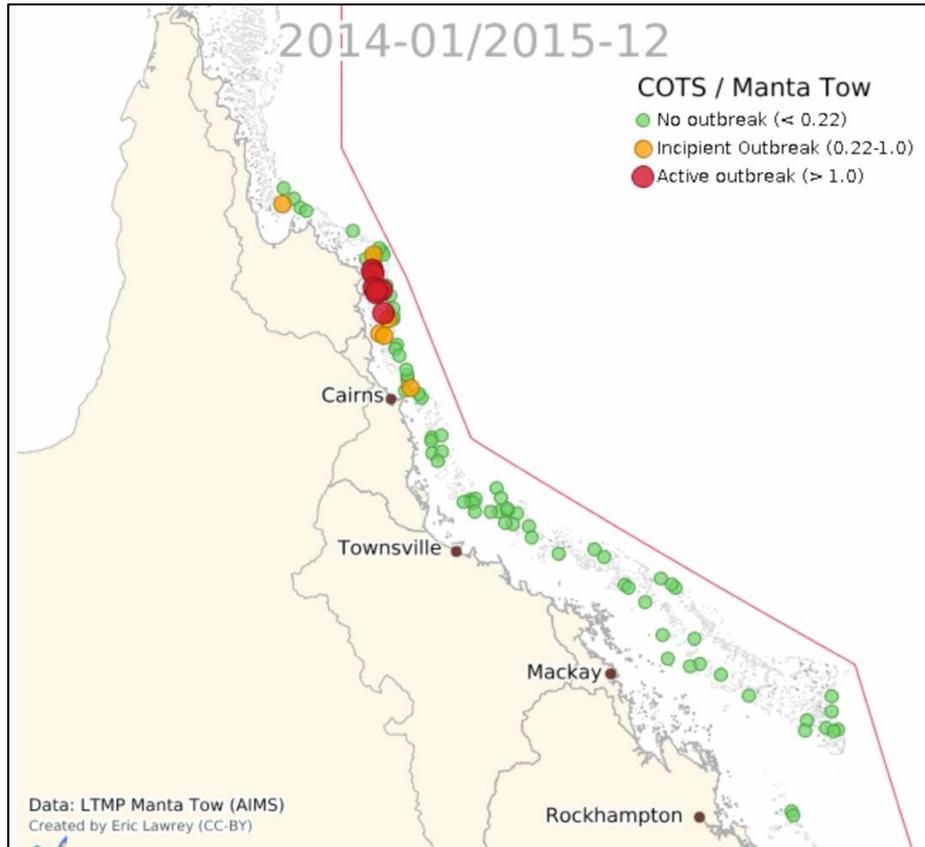
Expert Judgement within Conservation

		Increasing consequence >>>> Years				
		1- Insignificant 0.00-0.083	2-Minor 0.084-1	3-Moderate 1.01-3.00	4- Major 3.01-5.00	5-Severe 5.01-10.00
Increasing likelihood >>	5-Almost Certain 0.95-1.00	High	High	Extreme	Extreme	Extreme
	4-Likely 0.71-0.95	Moderate	High	High	Extreme	Extreme
	3- Moderate 0.31-0.71	Low	Moderate	High	Extreme	Extreme
	2- Unlikely 0.051-0.30	Low	Low	Moderate	High	Extreme
	1- Rare 0.00-0.05	Low	Low	Moderate	High	High

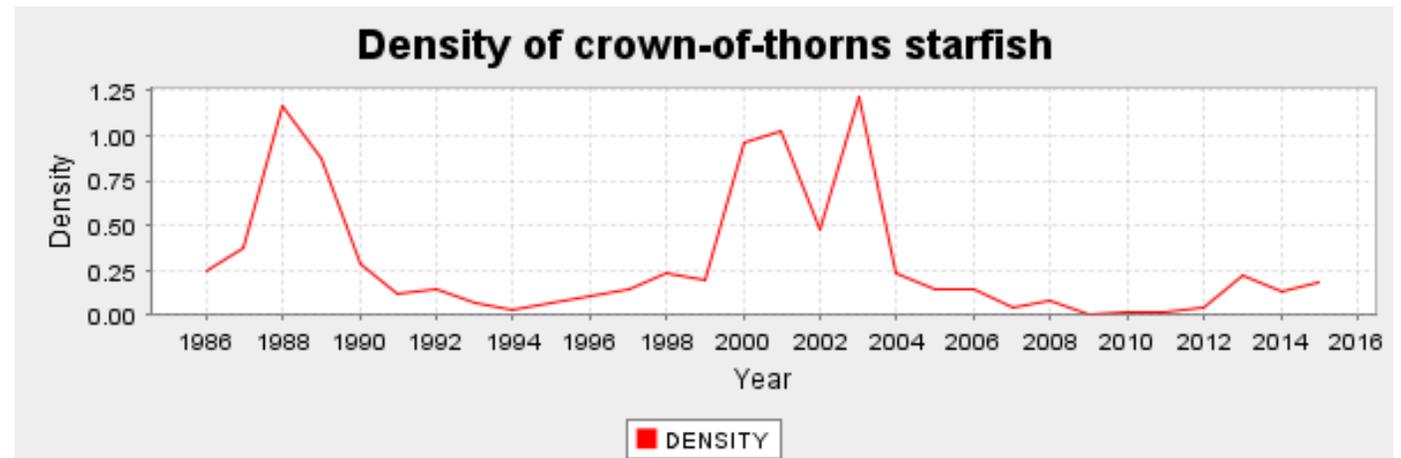
The IDEA Protocol



Case Study: Crown of Thorns Starfish on the Great Barrier Reef



The Great Barrier Reef, Australia



Source: Australian Institute of Marine Science

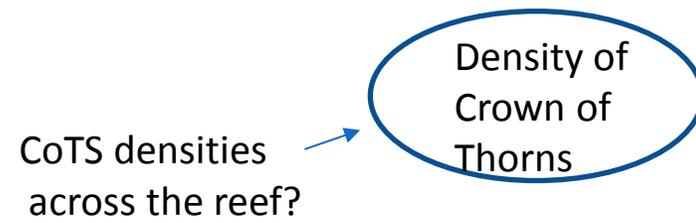
Defining Good Test Questions

- A minimum of 10 questions
- Things experts would need to know to answer the questions of interest
- Ideally domain predictions

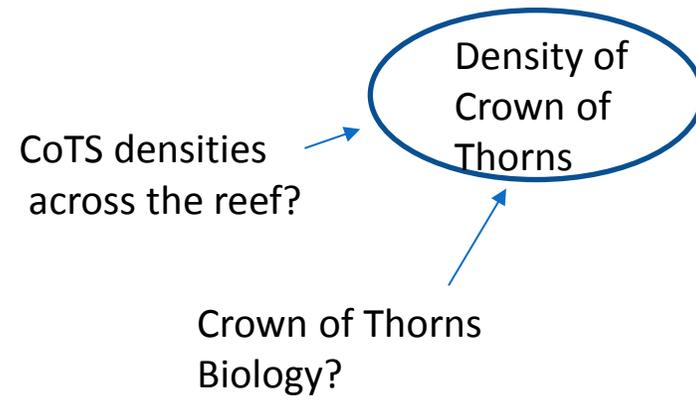
What are Relevant Questions?

Density of
Crown of
Thorns

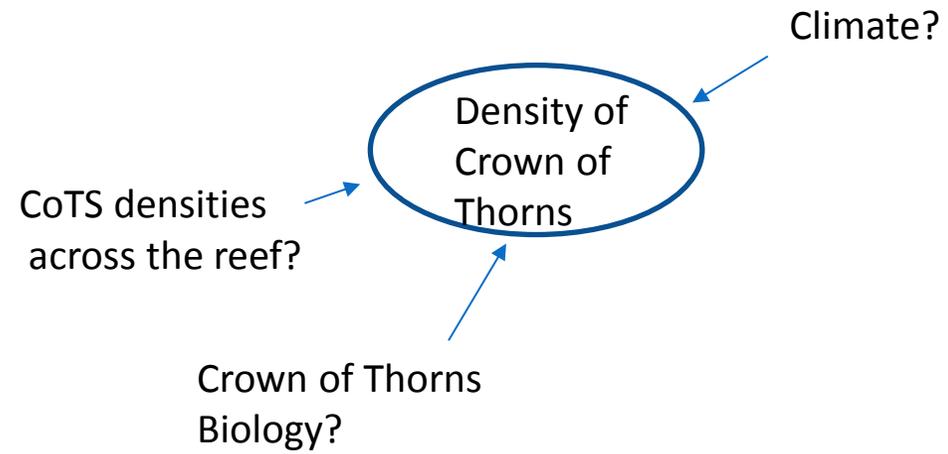
What are Relevant Questions?



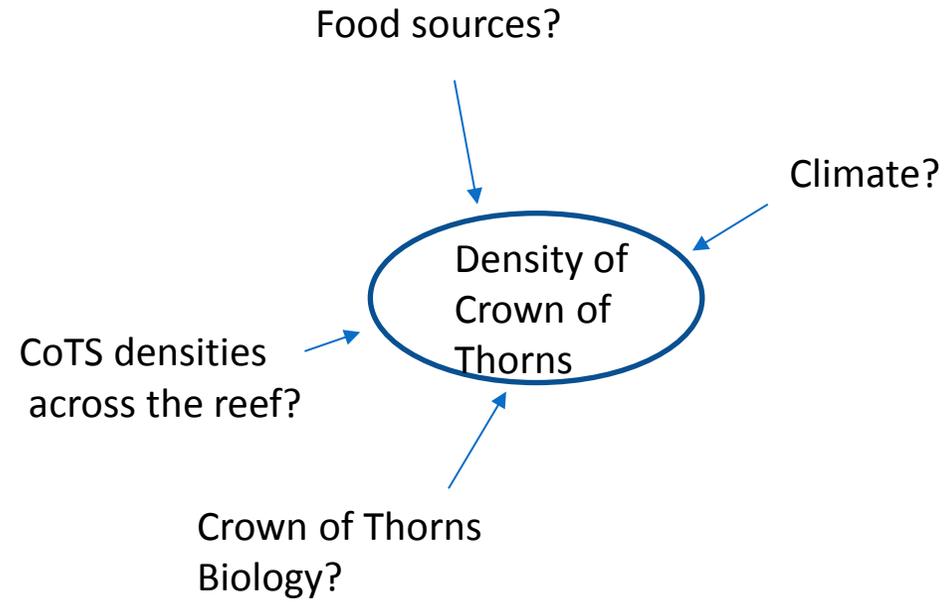
What are Relevant Questions?



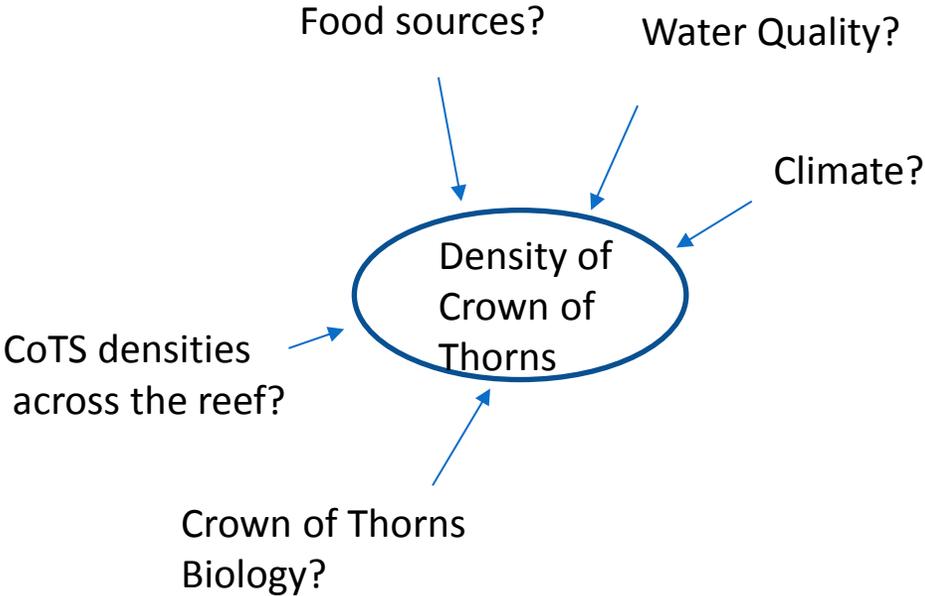
What are Relevant Questions?



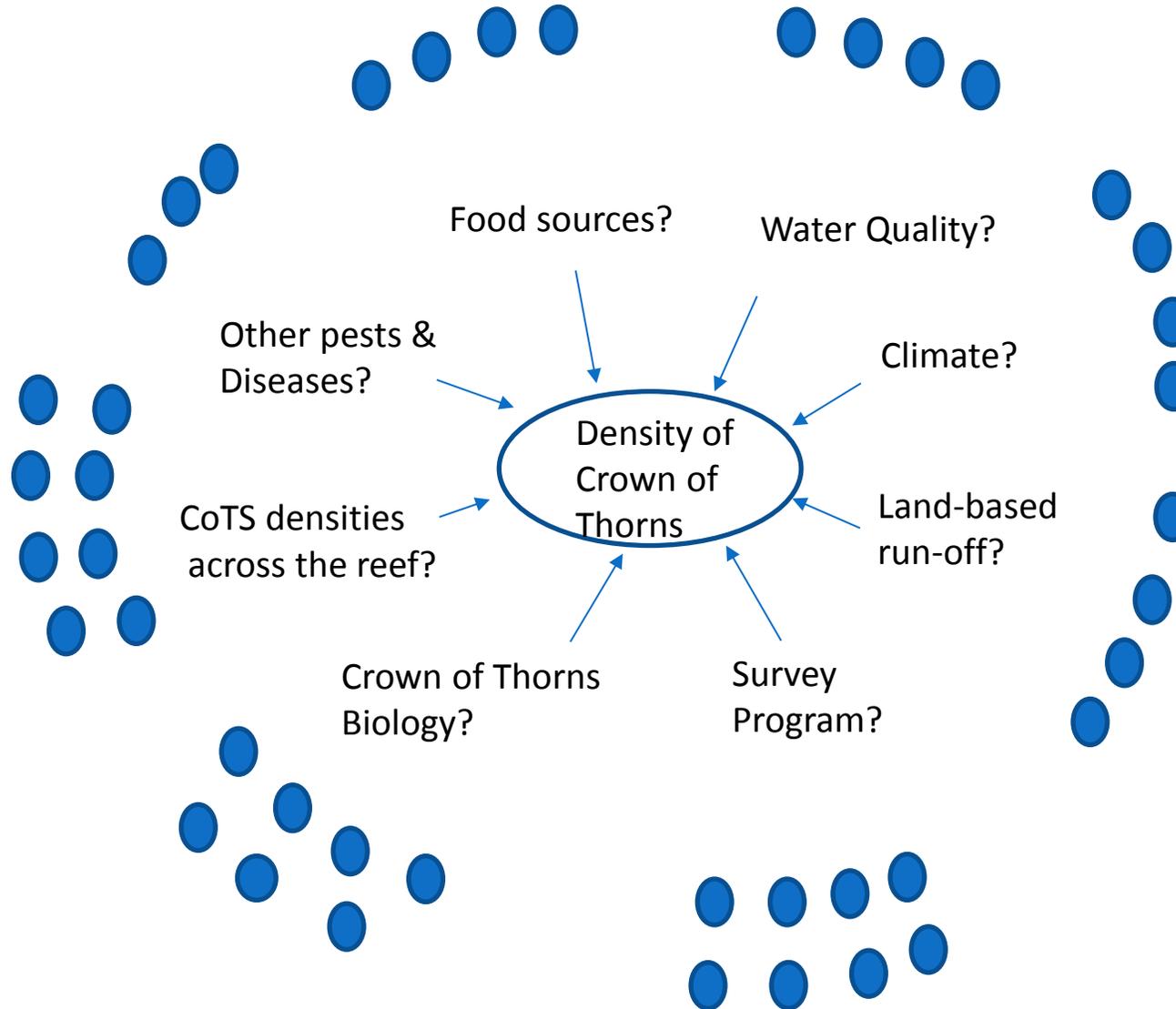
What are Relevant Questions?



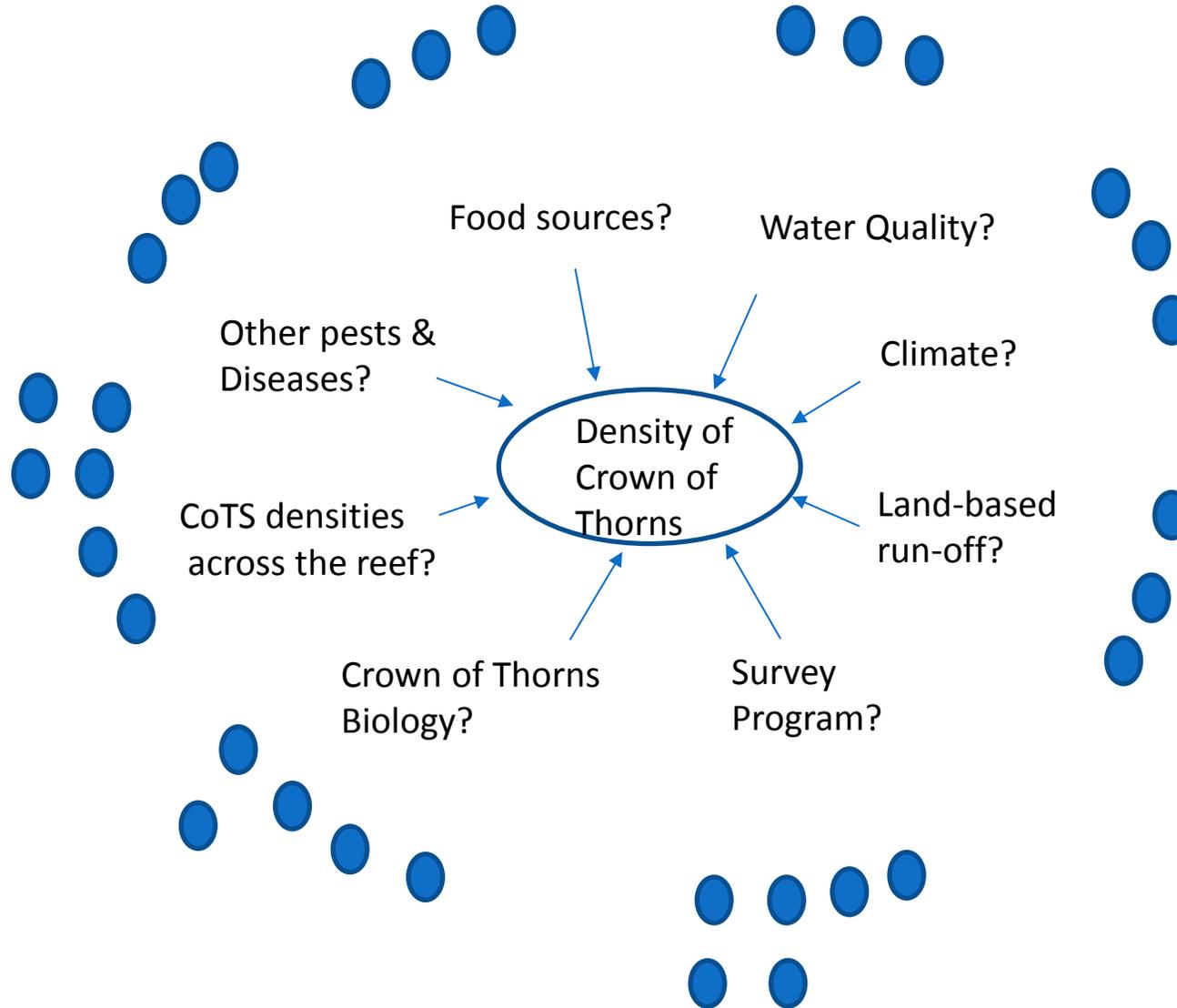
What are Relevant Questions?



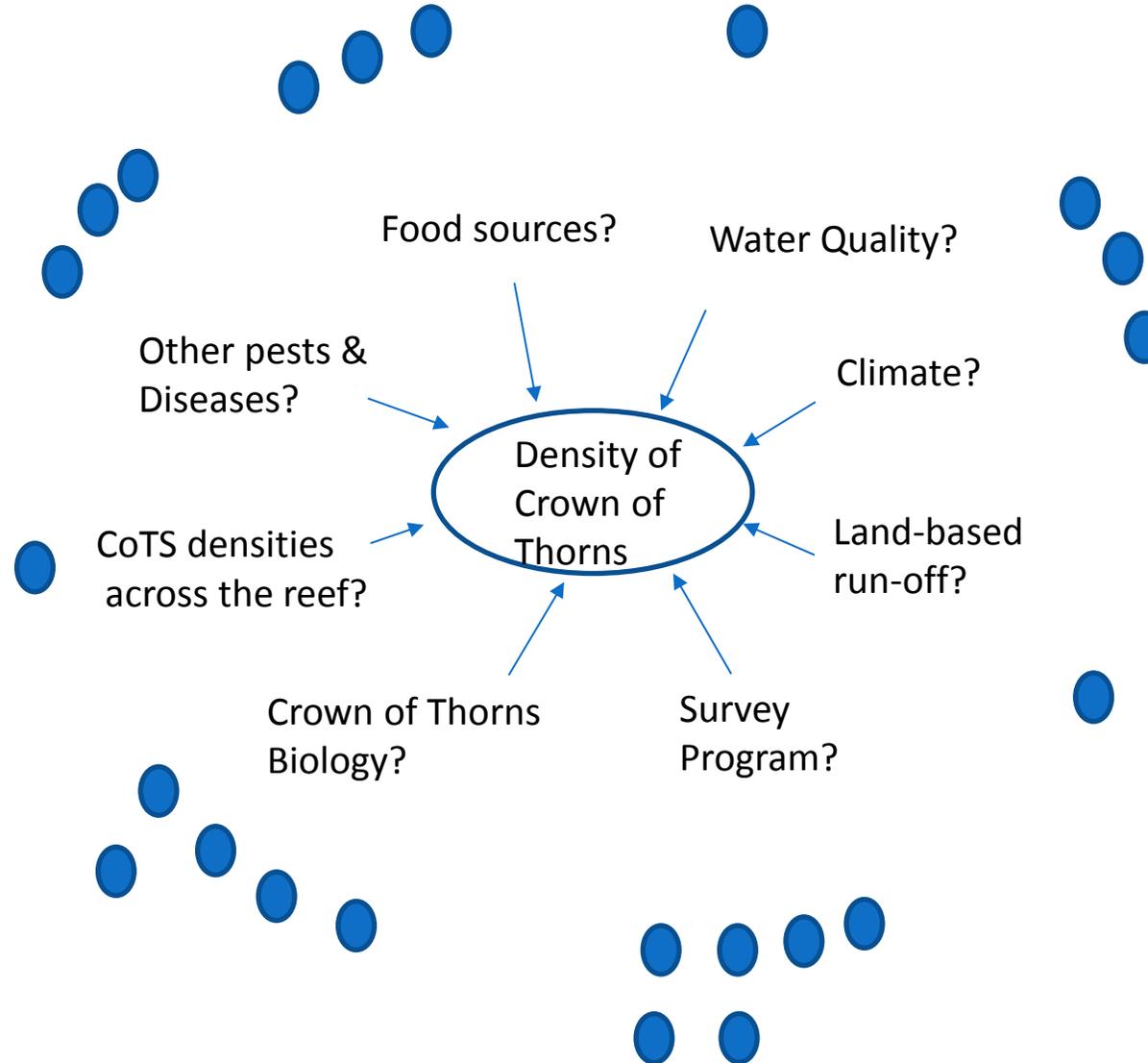
What are Relevant Questions?



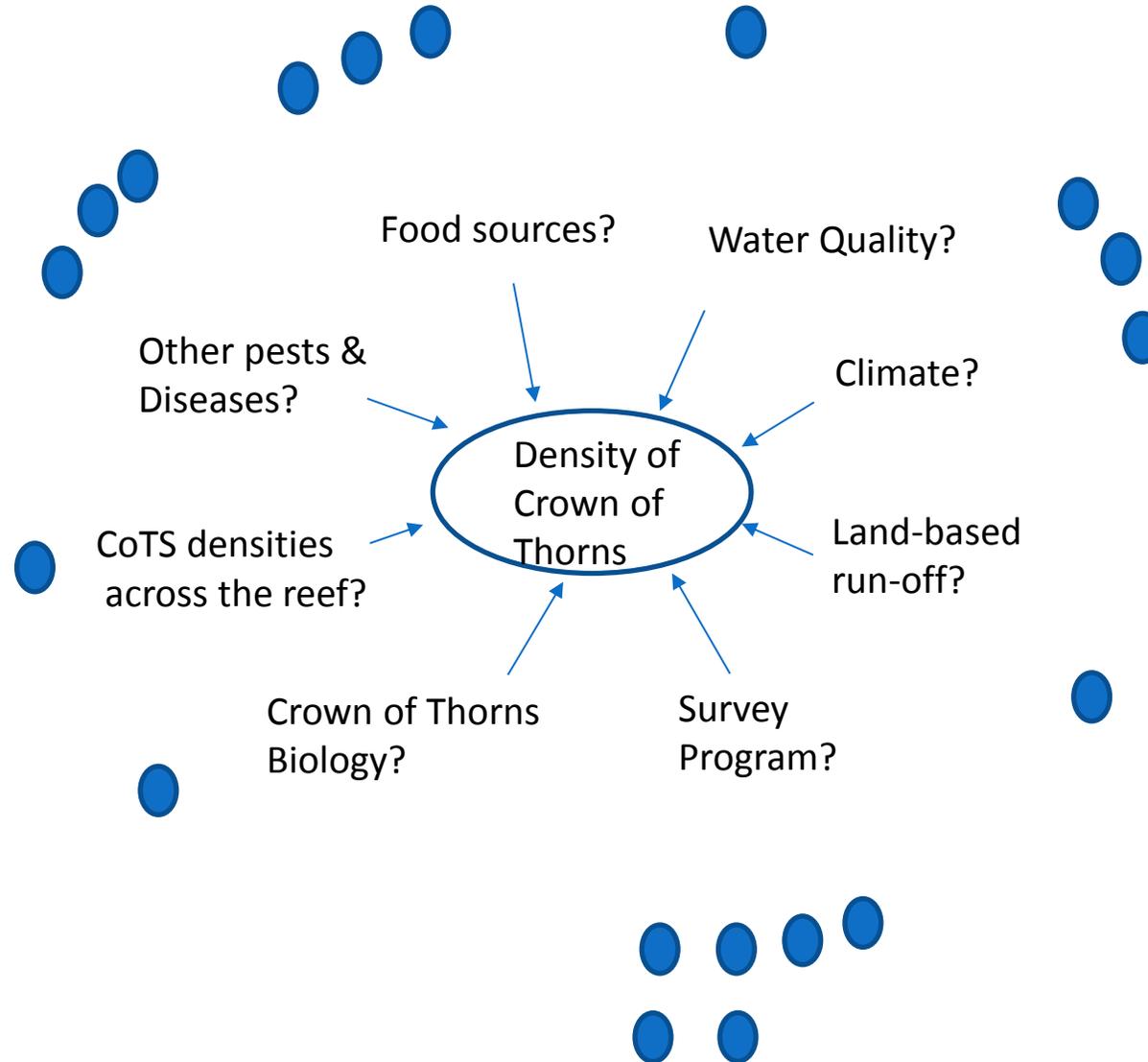
What are Relevant Questions?



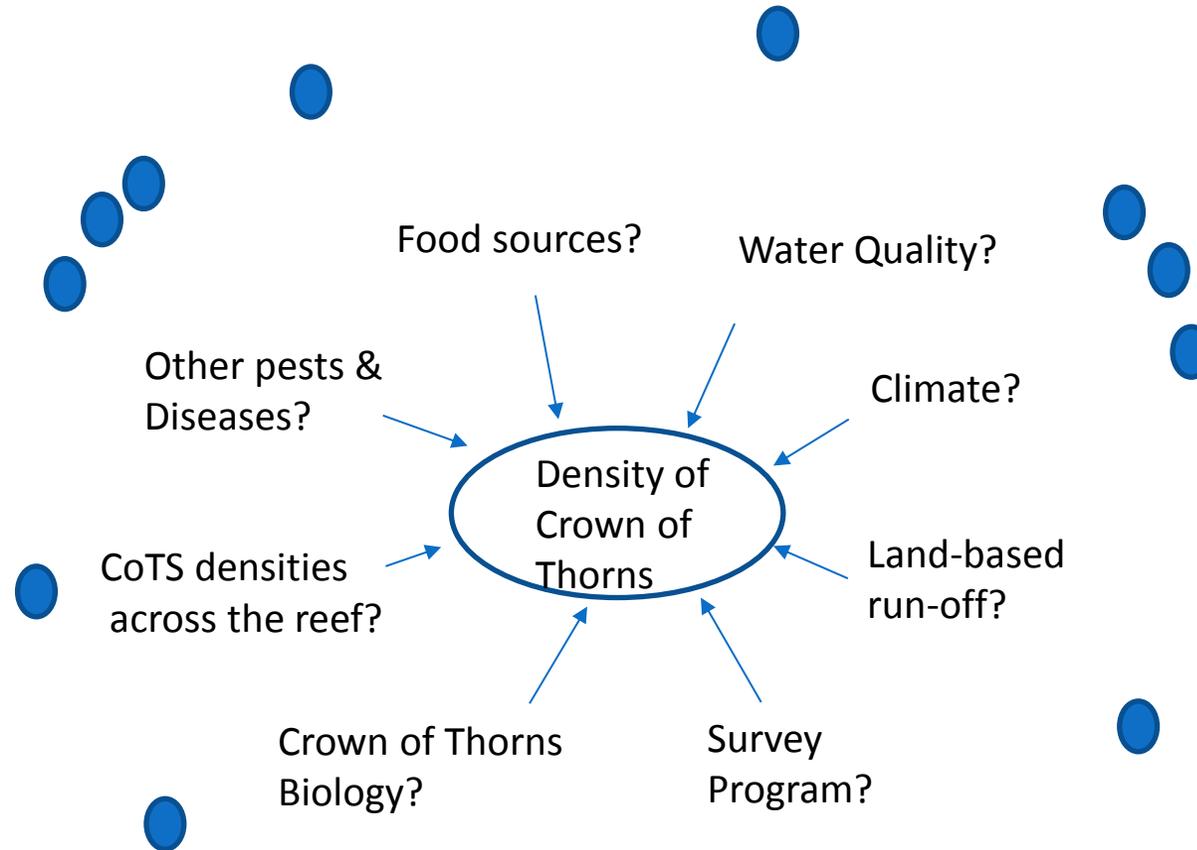
What are Relevant Questions?



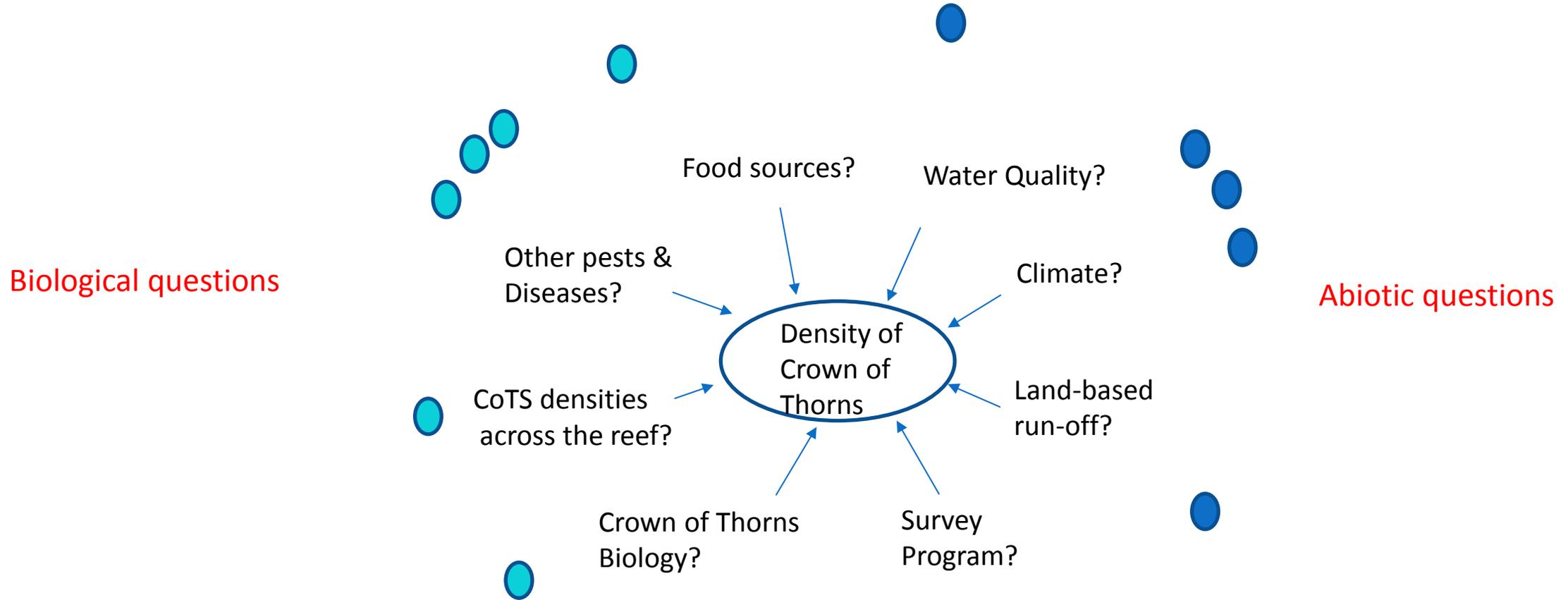
What are Relevant Questions?



What are Relevant Questions?



What are Relevant Questions?





How far can the questions be from the questions of interest?

- 7 Biotic Questions
- 7 Abiotic Questions
- 7 Geopolitical Questions

Great Barrier Reef Elicitation

*Images not my own

Biotic: Bleaching, Crown of Thorns, Invasive Species, Disease, Threatened Species, Predators, Culling



Abiotic: River discharge, El-Nino, Wind speed, Turbidity, Water Temperature, Chlorophyll, Air Temperature



Geopolitical: Zika Virus, Twitter Price, Gold, Space Launches, Refugees, Coal, Brexit.



How far can the questions be from the questions of interest?

- 7 Abiotic Questions
 - 7 Geopolitical Questions
- 
- 7 Biotic Questions

How far can the questions be from the questions of interest?



- 7 Biotic Questions
- 7 Abiotic Questions
- 7 Geopolitical Questions

An example of output from current protocol: Crown of Thorns

Component	Worst 10%	Best 10%	Most likely	Confidence
Crown-of-thorns starfish	Very poor	Very good	Poor	Medium

Current method (Ward 2014)

21 Clear Questions



"The Great Barrier Reef Intelligence Game, 2016"
Victoria Hemming, Mark Burgman, Terry Walthe, Anca Hanes,
School of Biosciences, University of Melbourne



Question 1 Density of Crown of Thorns Starfish (*Acanthaster planci*)

*"What will be the average density of Crown of Thorns Starfish (*Acanthaster planci*) detected per 2 minute manta-tow at Rib Reef, in the Townsville region, as surveyed by the Australian Institute of Marine Science (AIMS) as part of the Long-term Monitoring Program between 1 March, 2016 and 30 June, 2016 (inclusive)?"*

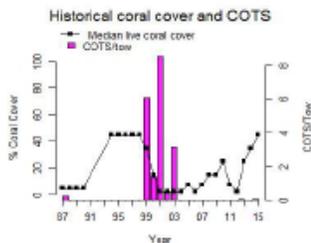
Clarification: Crown of Thorns Starfish (*Acanthaster planci*) (CoTS) are found at numerous coral reef ecosystems, including the Great Barrier Reef. They consume hard corals, and are the focus of manta-tow surveys undertaken by AIMS as part of the Long Term Monitoring Program (LTMP).

This question relates specifically to the density of CoTS per two minute manta-tow that will be detected by AIMS during surveys at Rib Reef between 1 March 2016 and 30 June 2016 (inclusive). Rib Reef is located in the Townsville region of the Great Barrier Reef (GBR) (Appendix A). The average density per 2 minute manta tow, is a standard metric used to compare between reefs and years. The average density of CoTS per 2 minute manta tow refers to the total number of CoTS that are detected by AIMS during manta-tow surveys, divided by the total number of manta-tow surveys undertaken at Rib Reef. We will accept survey results for Rib Reef recorded between 1 March, 2016 and 30 June 2016 (inclusive). If the survey does not occur, or occurs outside of this period, the question will be void. As with all monitoring data, it is important to note that this question relates specifically to the number of CoTS detected and reported, not necessarily the actual number of CoTS present at Rib Reef.

Resolution: The question will be resolved when the report for the Townsville section for the 2015/2016 monitoring period is published online by AIMS (see latest surveys in useful links).

Additional Information:

- Historical density of CoTS per 2 minute manta tow at Rib Reef recorded by the AIMS LTMP



Useful Links:

- Rib Reef <http://data.aims.gov.au/v2/regions/ribreef/ribreef-180025>
- Latest surveys <http://www.aims.gov.au/docs/research/monitoring/reef/latest-surveys.html>
- Survey methods <http://www.aims.gov.au/documents/20001/2015M4E-410b-480E-9d07-4152931c3d7f>
- Map of LTMP regions [Appendix A](#)

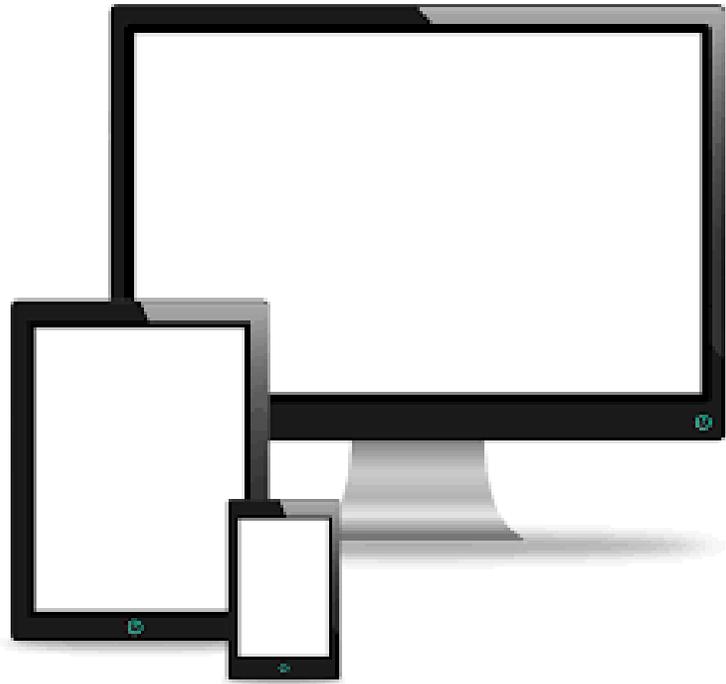
*"What will be the average density of Crown of Thorns Starfish (*Acanthaster planci*) detected per 2 minute manta-tow at Rib Reef, in the Townsville region, as surveyed by the Australian Institute of Marine Science (AIMS) as part of the Long-term Monitoring Program between 1 March, 2016 and 30 June, 2016 (inclusive)?"*

76 experts

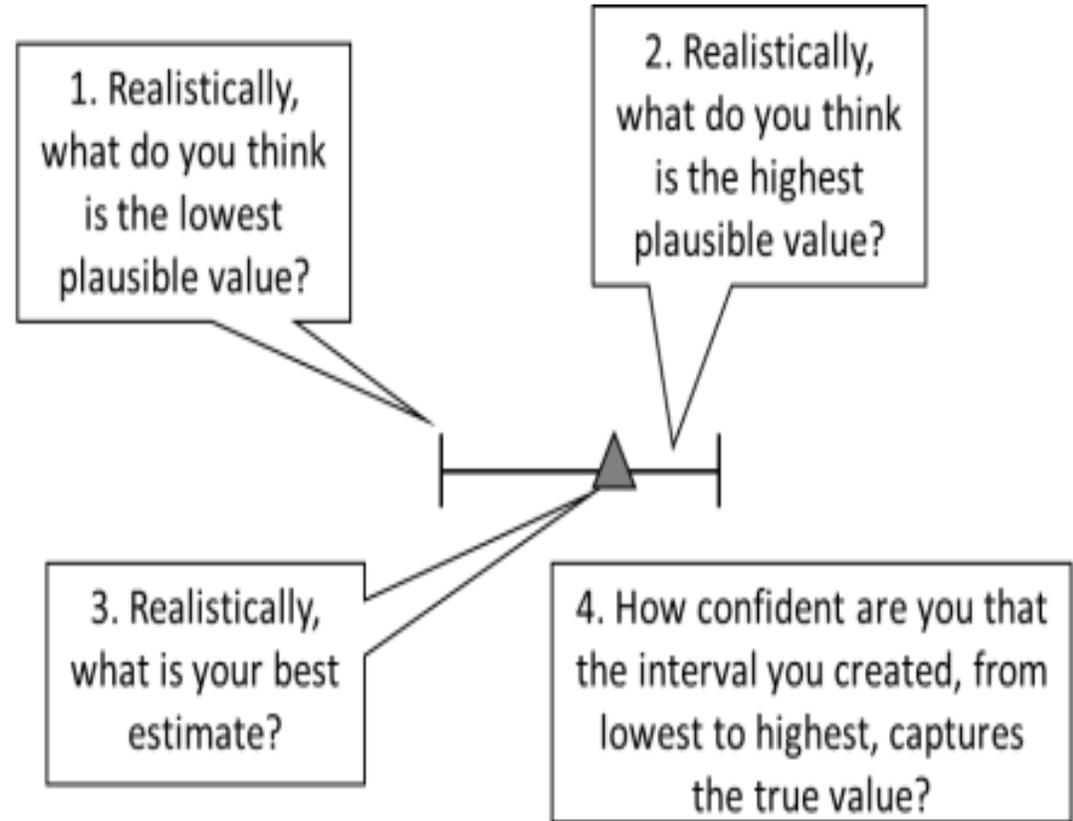


Divided into
8 groups





Remote elicitation

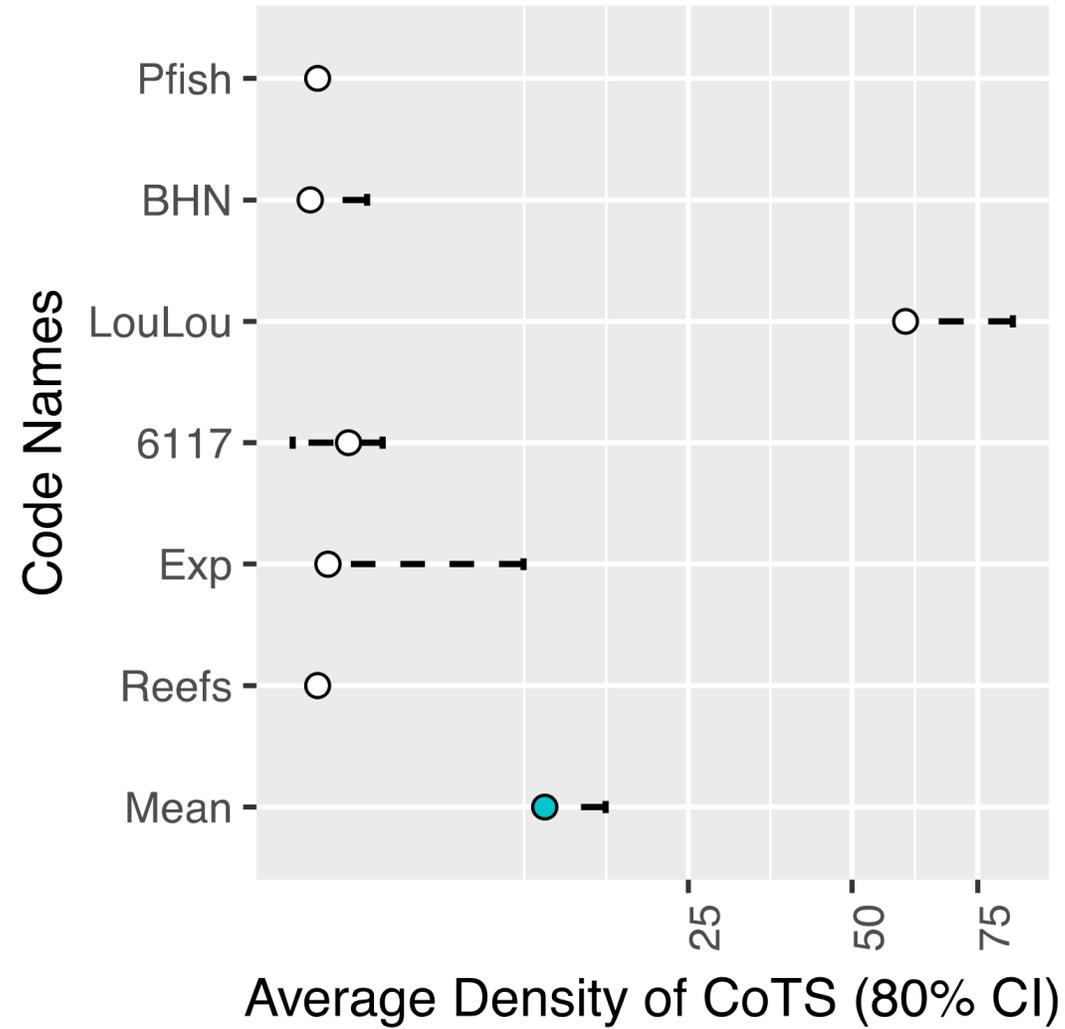


4-step elicitation

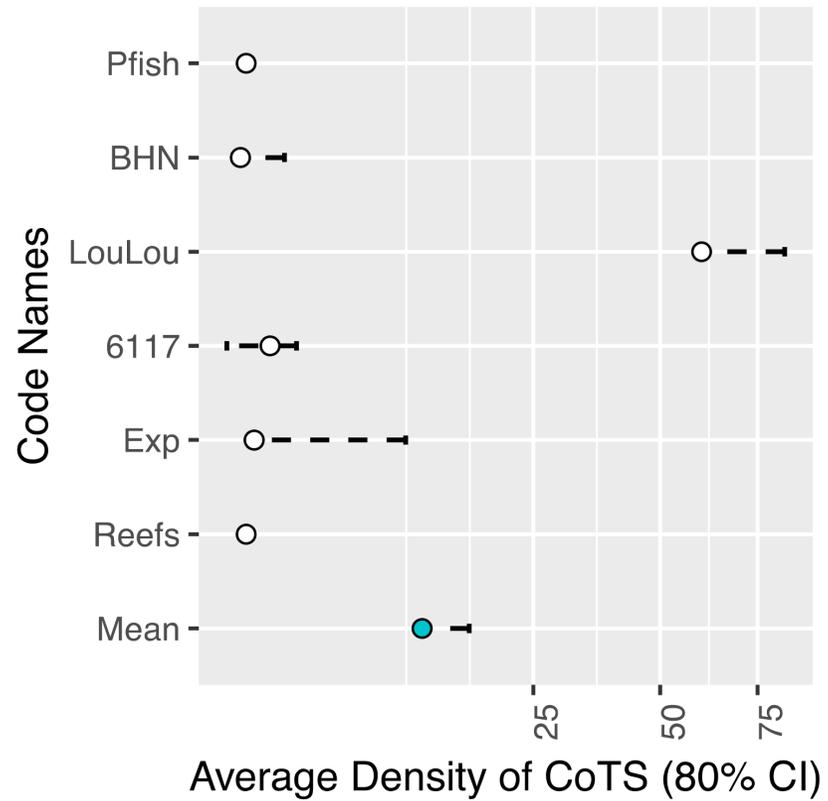
Round 1: Remote Elicitation (10 days)



Feedback



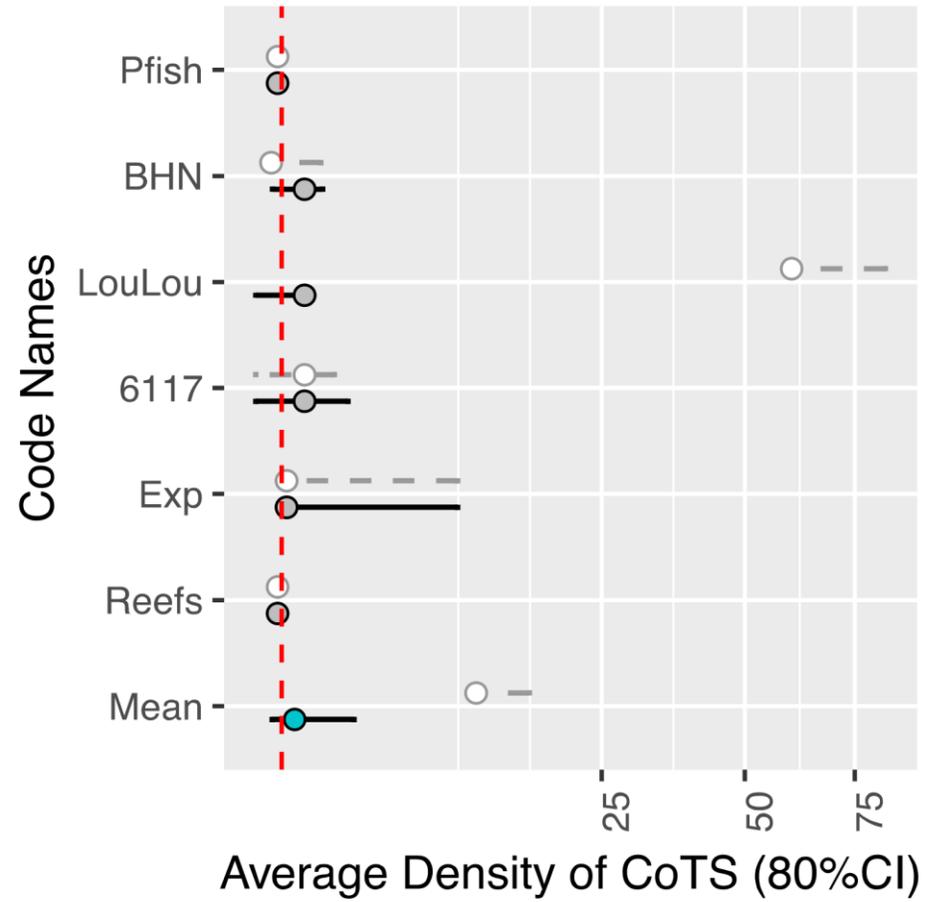
Discussion: 10 days



Round 2: Revised anonymous estimate (7 days)



Final Results

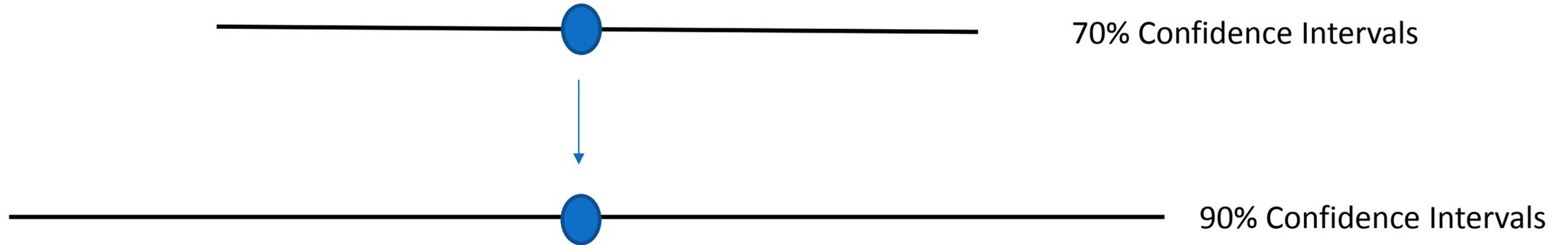


An example of current protocol output: Crown of Thorns

Component	Worst 10%	Best 10%	Most likely	Confidence
Crown-of-thorns starfish	Very poor	Very good	Poor	Medium

Current method (Ward 2014)

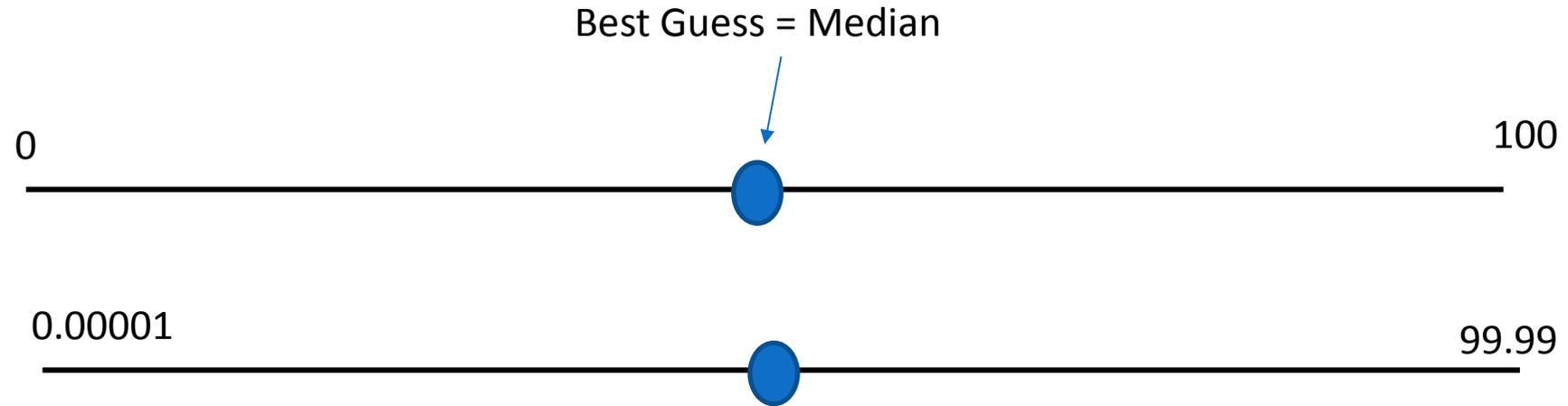
Preparing for Excalibur



Preparing for Excalibur

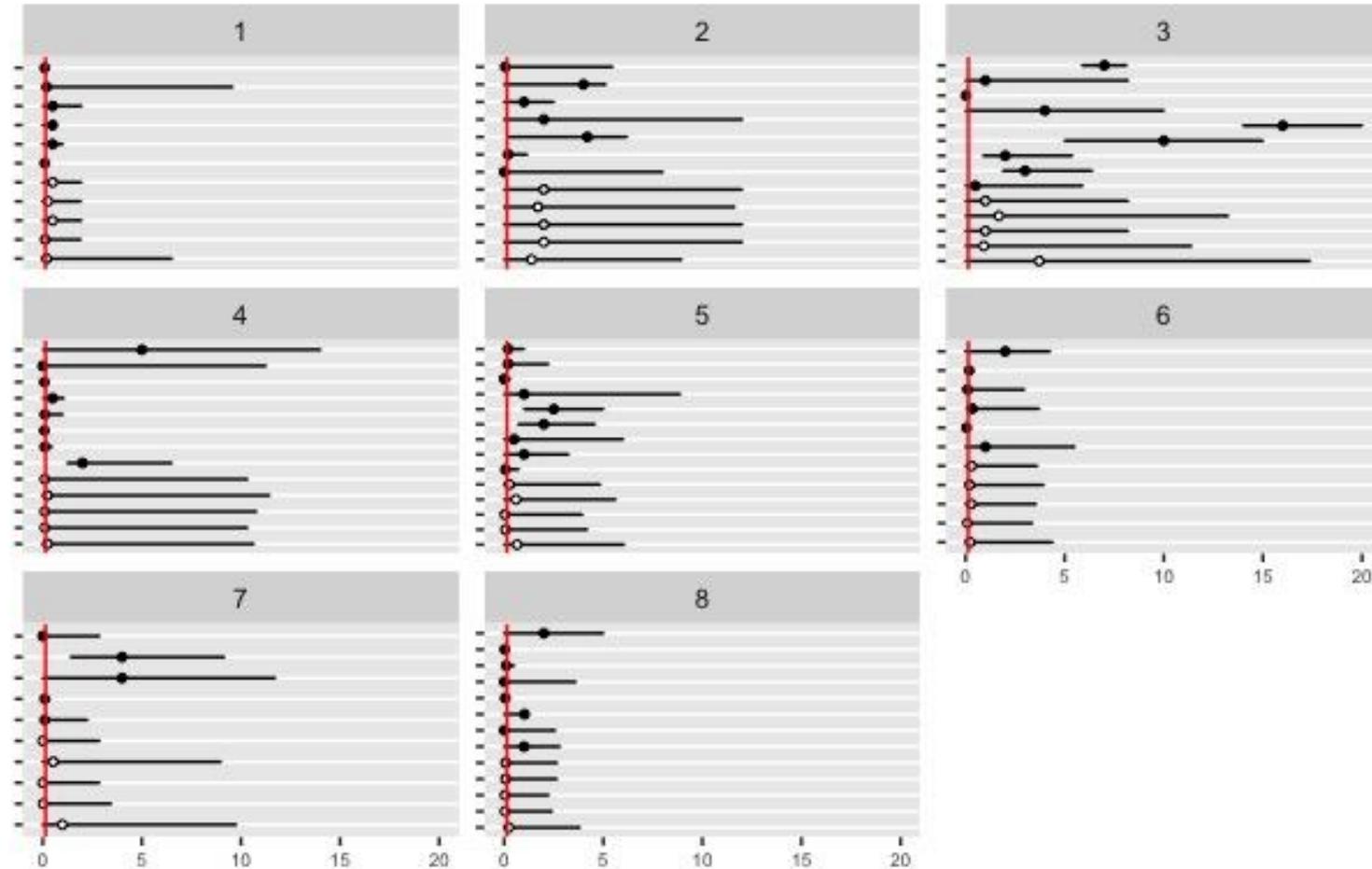


Preparing for Excalibur



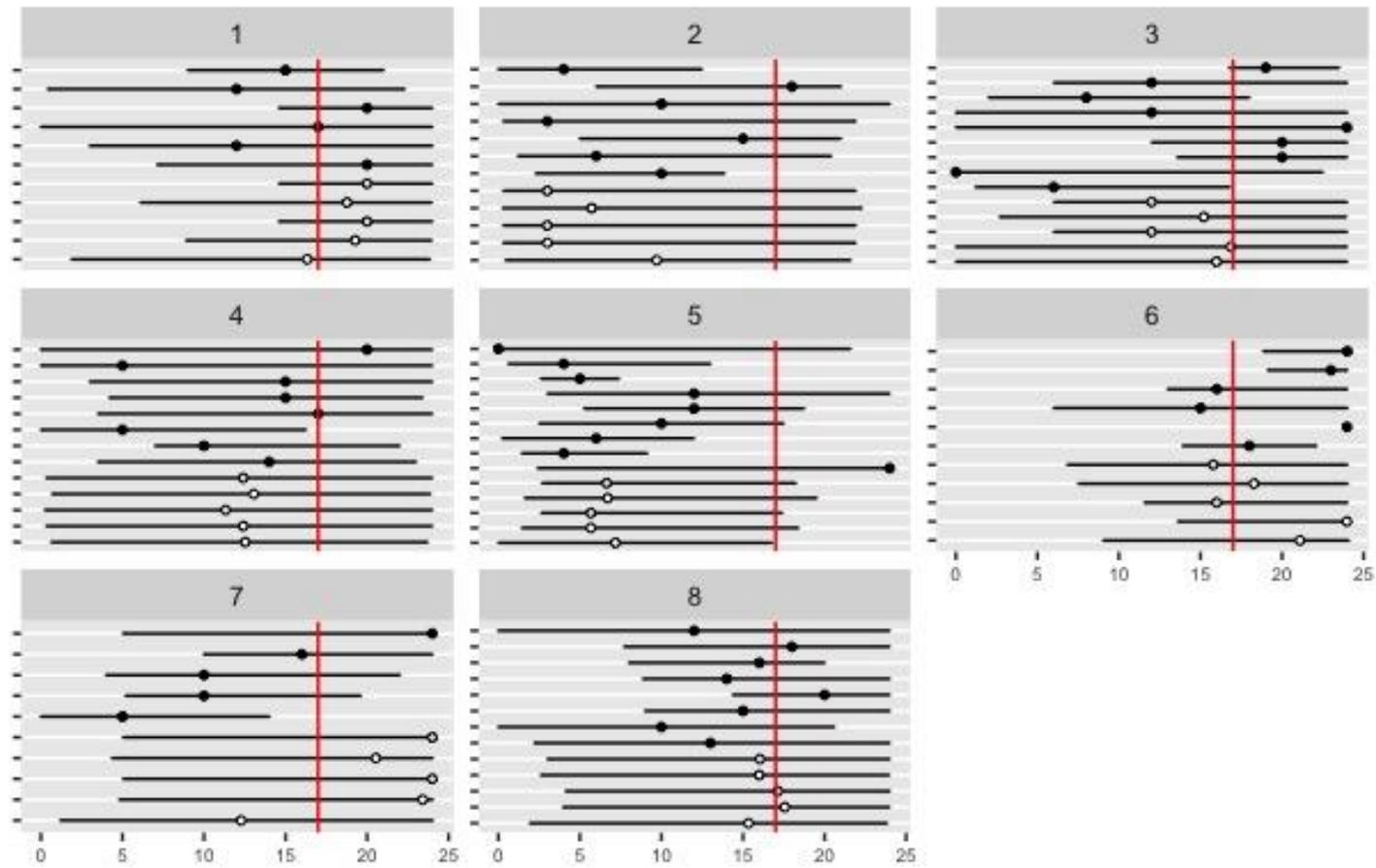
Results: Abiotic and Biotic Questions

Average density of CoTS at Rib Reef



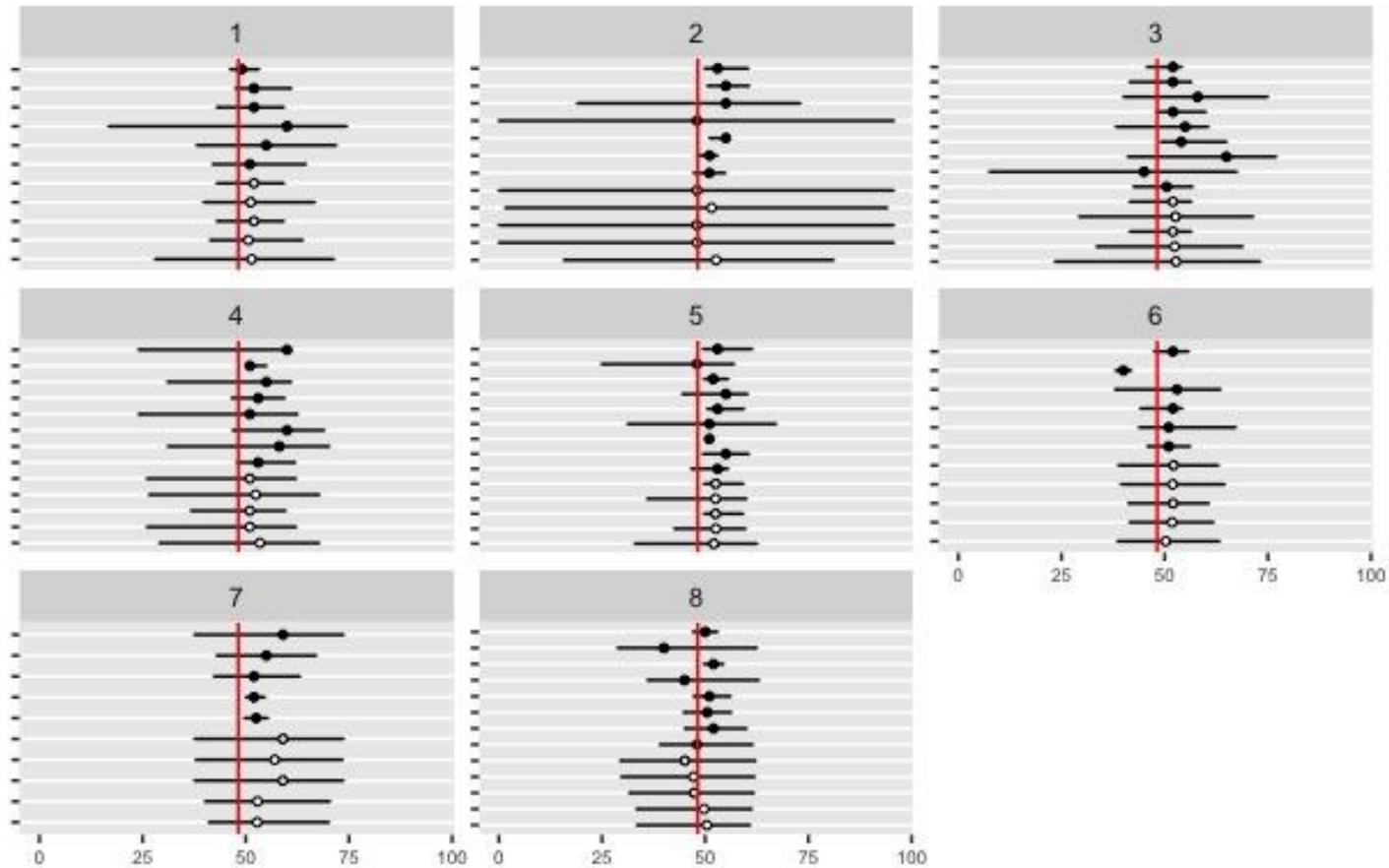
Results: Abiotic and Biotic Questions

Number of reefs (out of 24) with at least 1% bleaching of hard coral

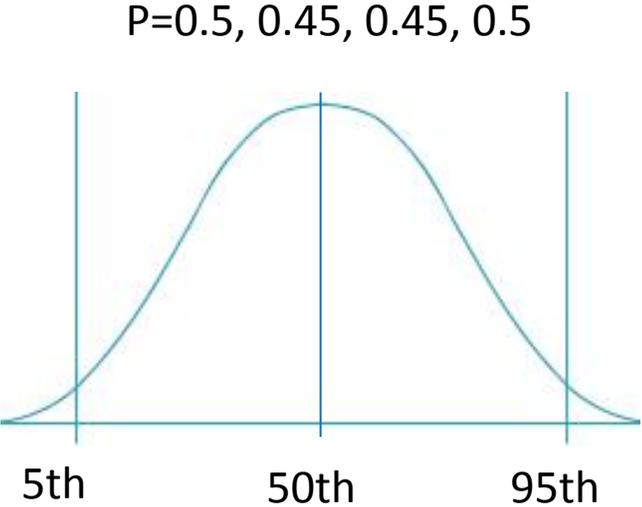


Results: Abiotic and Biotic Questions

Votes in favour of UK REMAINING in the EU



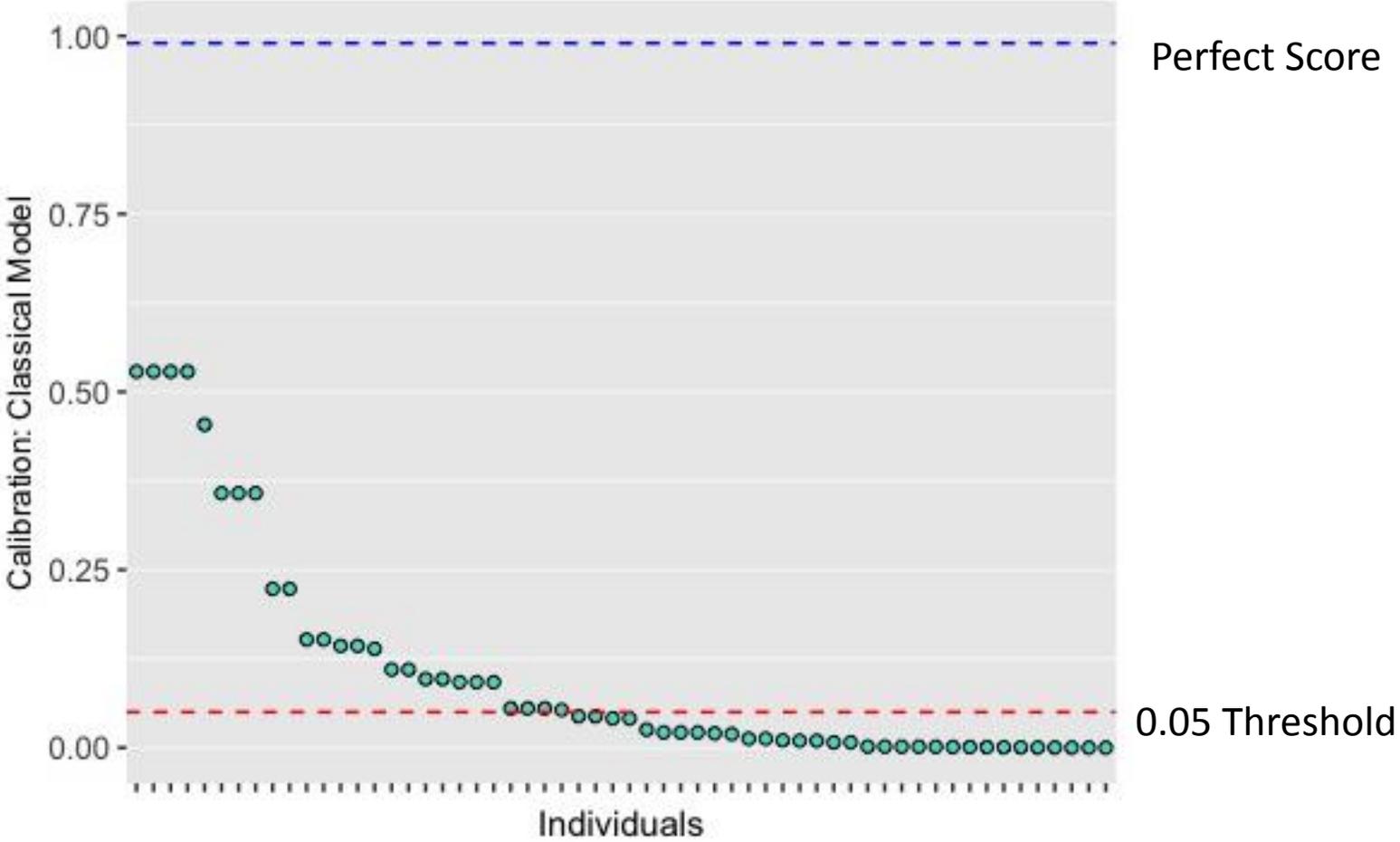
Statistical Accuracy: Classical Model (13 Questions)



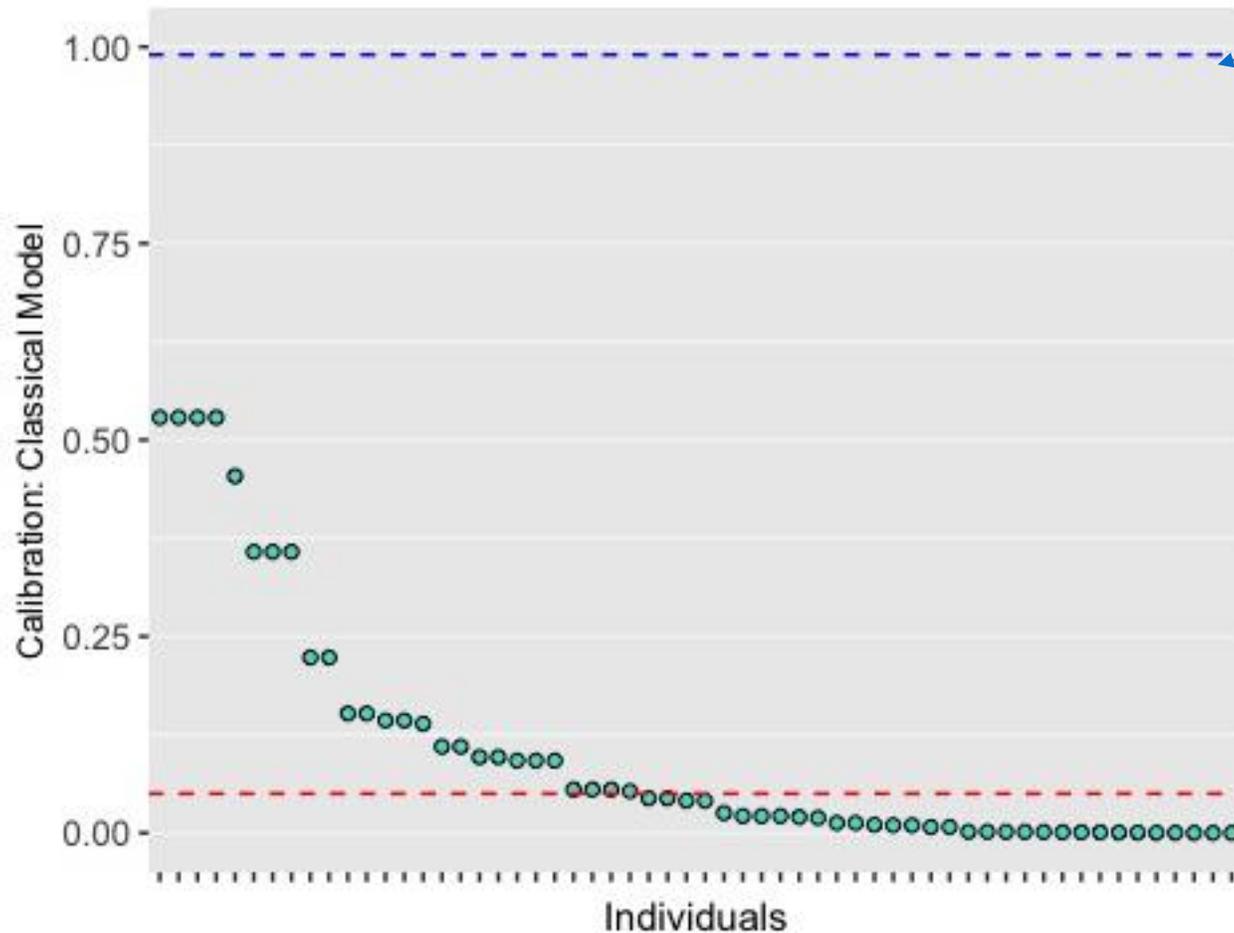
Cooke's Calibration = 0			
<5th	5th- 50th	50th-95th	>95th
X X X X X X X X X X X X X			
13	0	0	0

Cooke's Calibration = 0.928			
<5th	5th- 50th	50th-95th	>95th
X	X X X X X X X X	X X X X X	X
1	6	5	1

Individual Performance: Statistical Accuracy

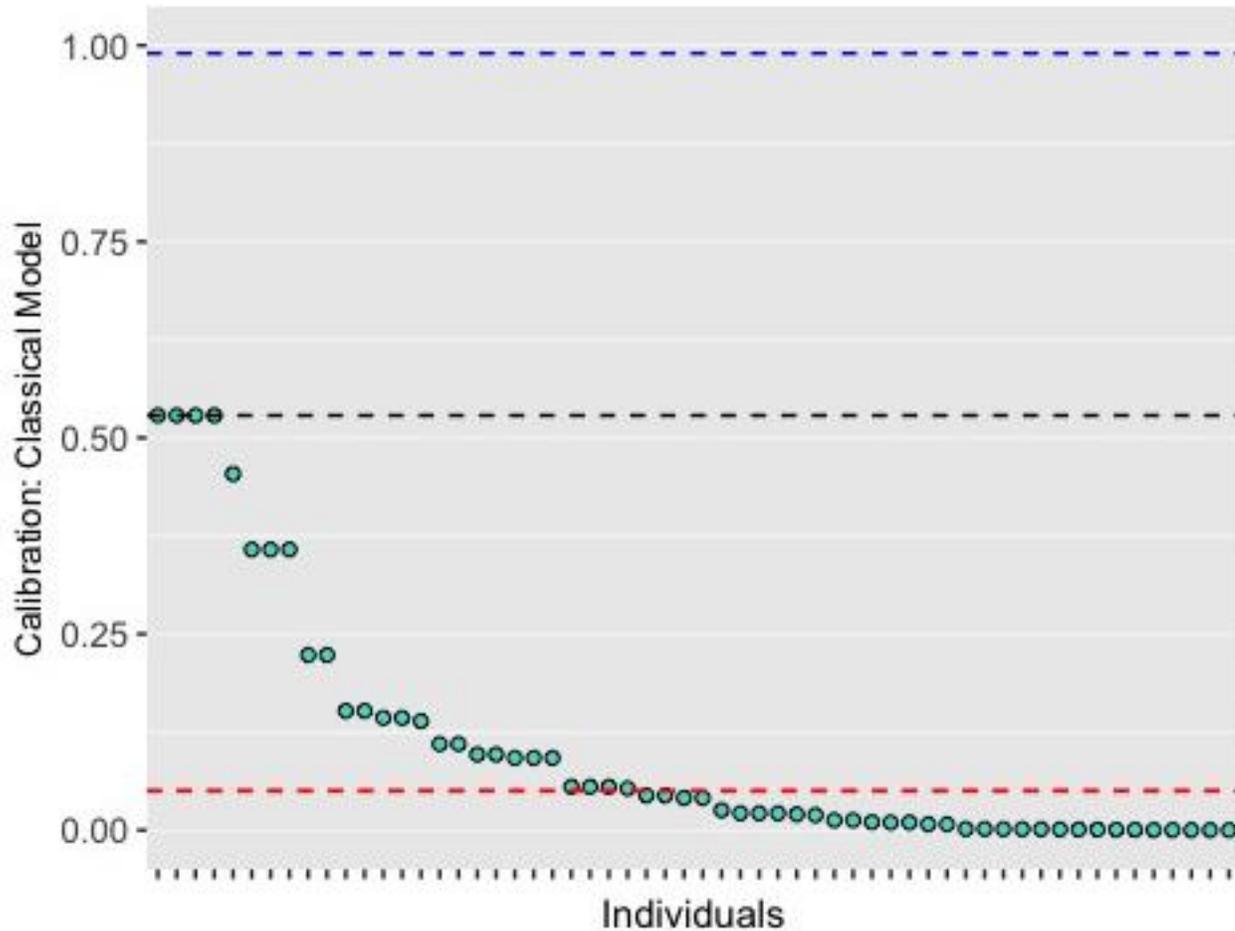


Individual Performance: Statistical Accuracy



<5 th	5 th - 50 th	50 th -95 th	>95 th
	X X X X X X X	X X X X X	
X			X
1	6	5	1

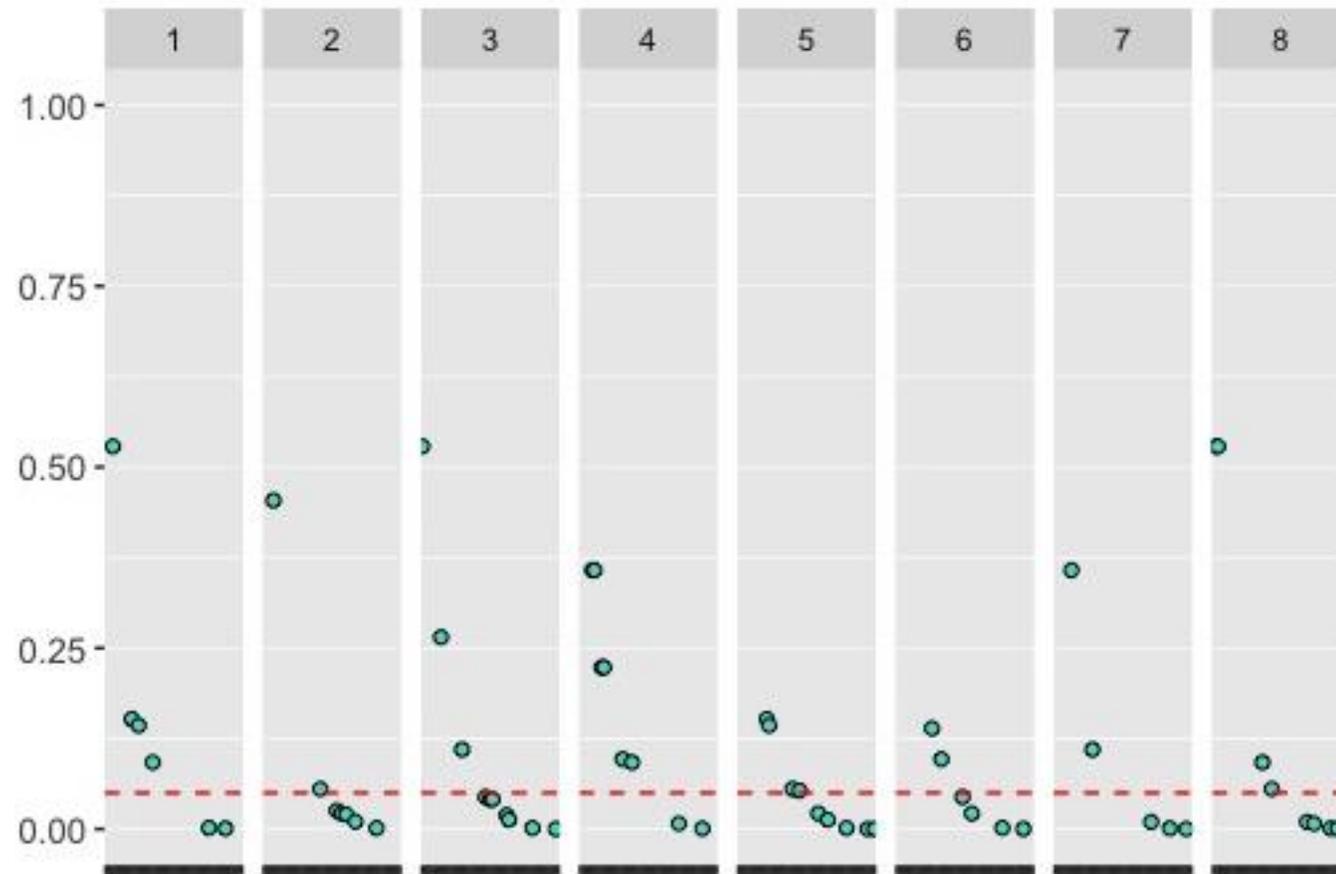
Individual Performance: Statistical Accuracy



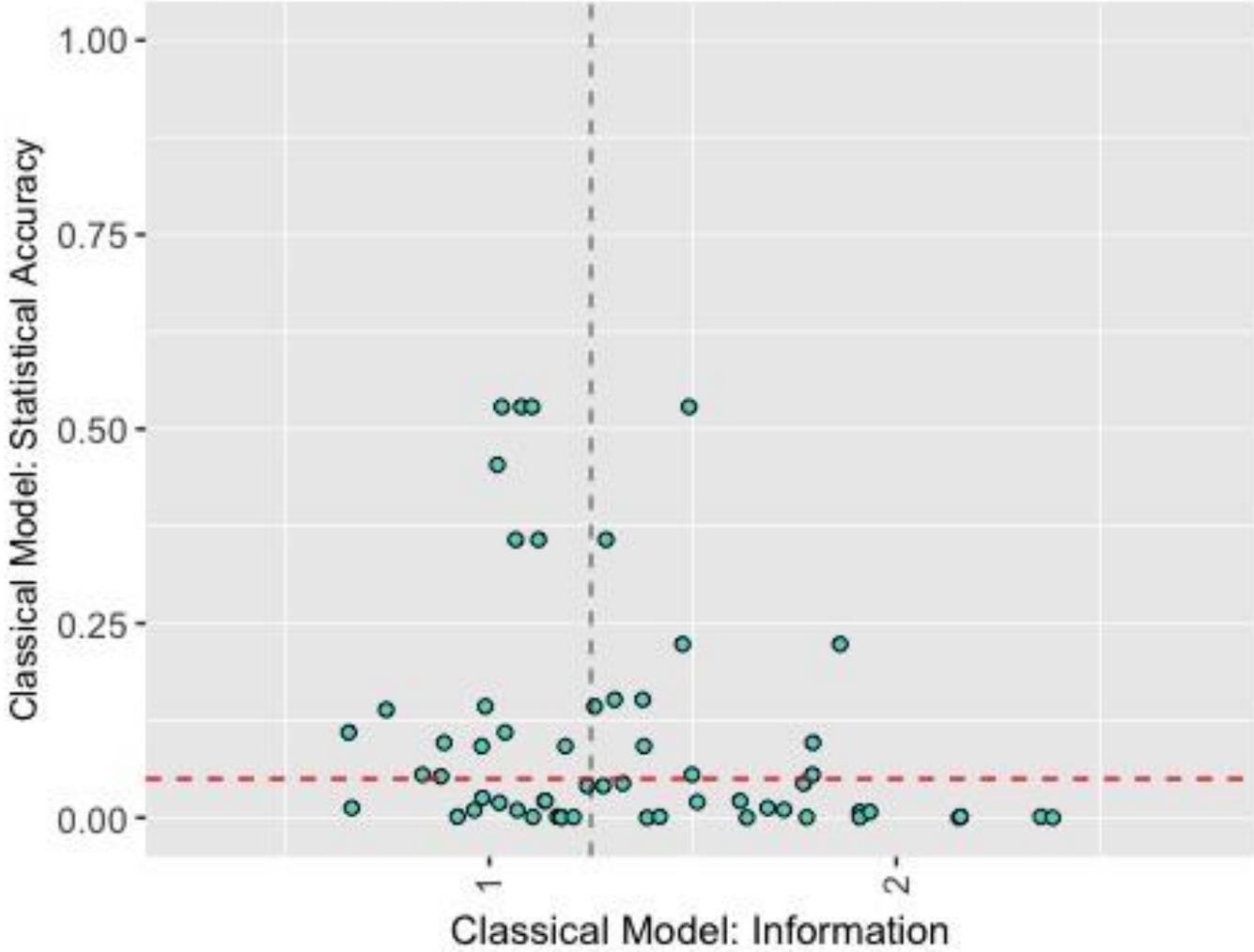
<5 th	5 th - 50 th	50 th -95 th	>95 th
X	X X X X X X	X X X X X	X
1	6	5	1

<5 th	5 th - 50 th	50 th -95 th	>95 th
X X	X X X X X	X X X X X	X
2	5	5	1

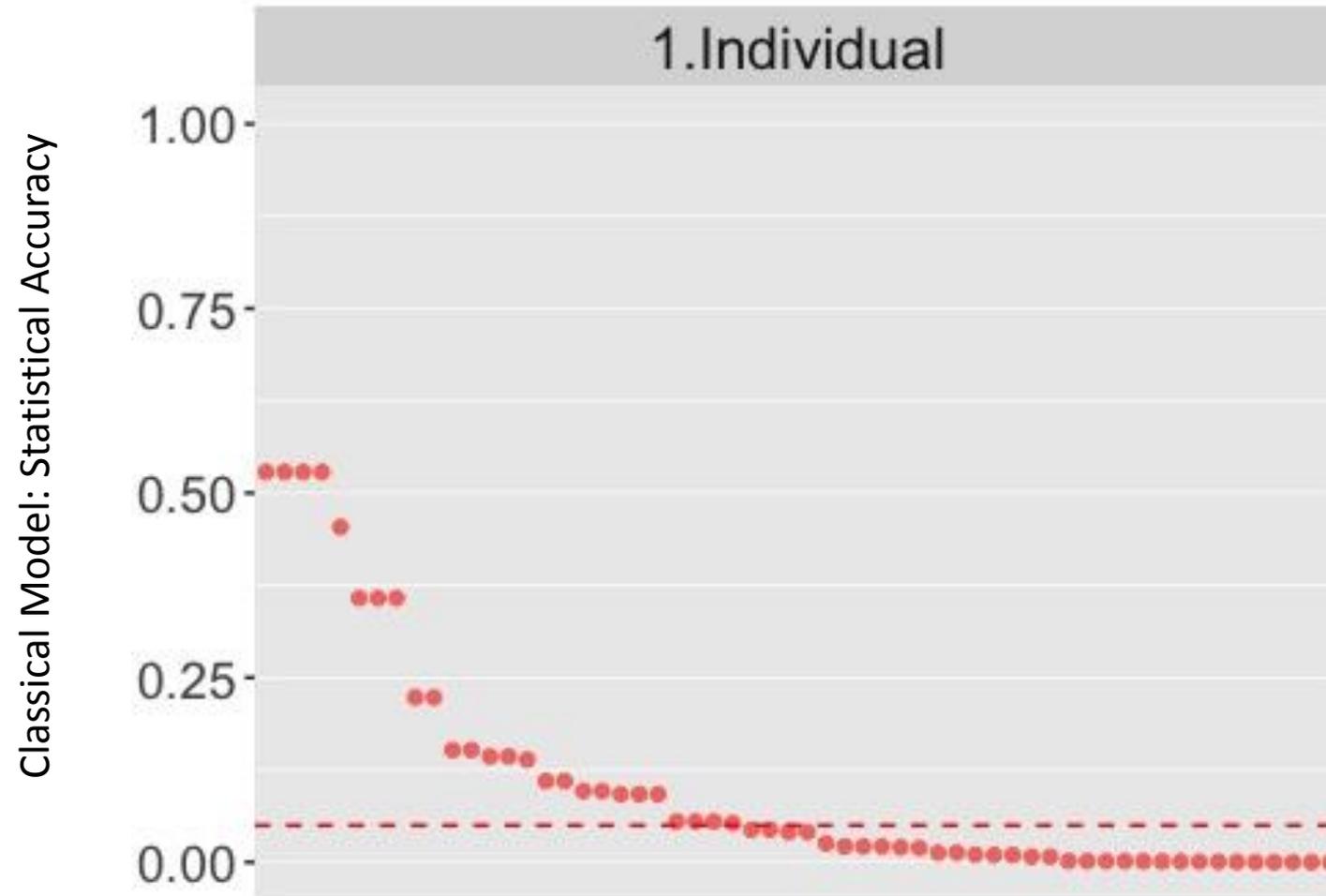
Calibration of Individuals per group



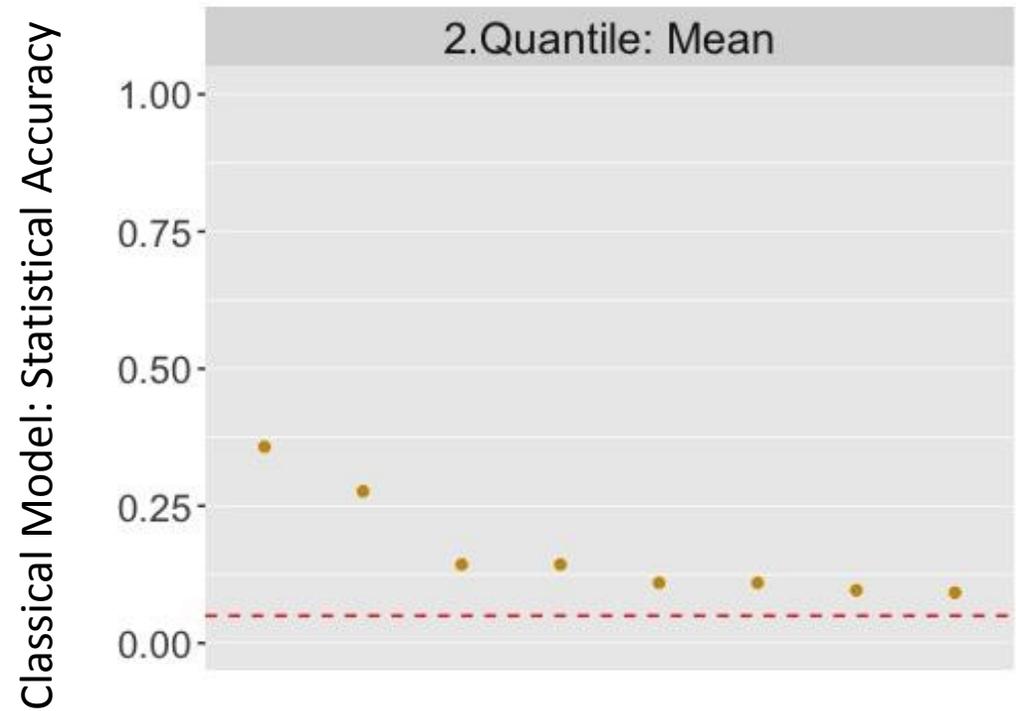
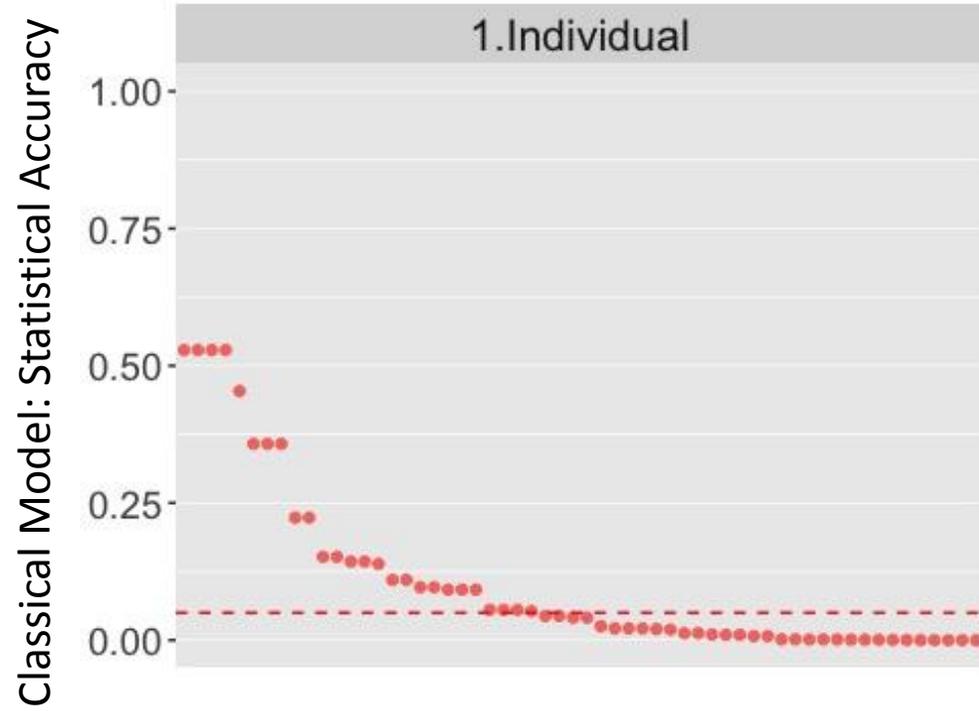
Information vs Statistical Accuracy



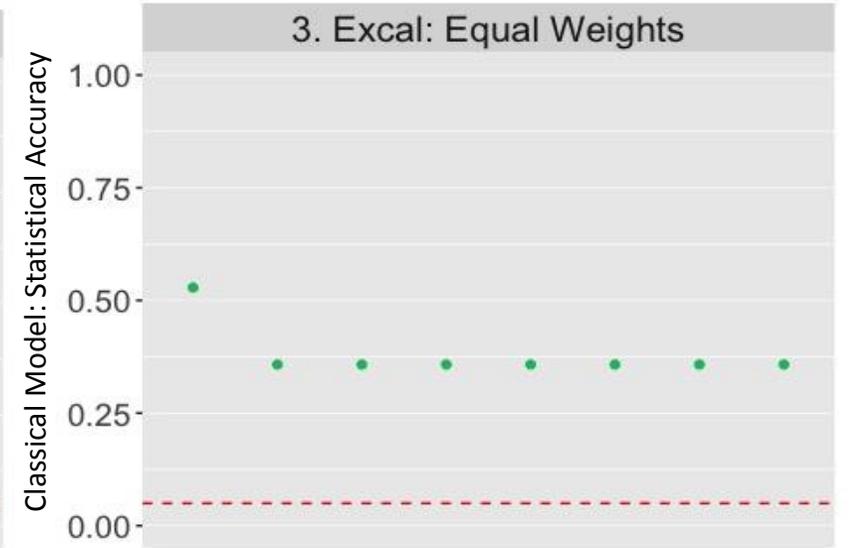
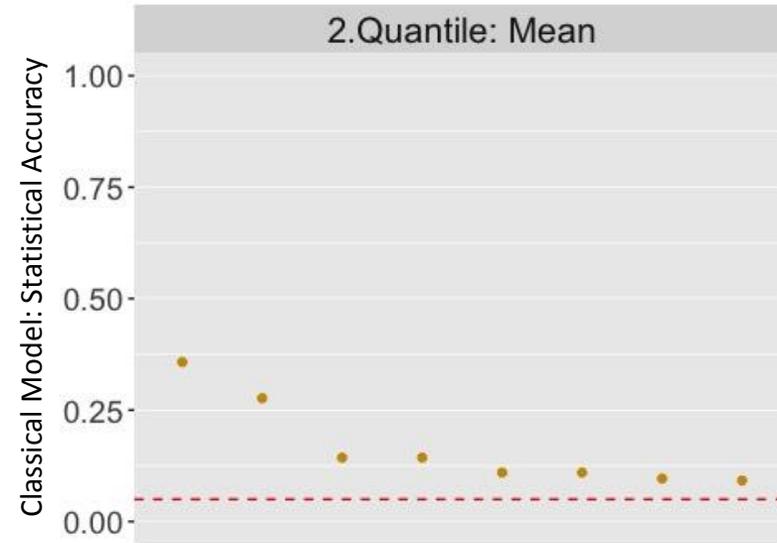
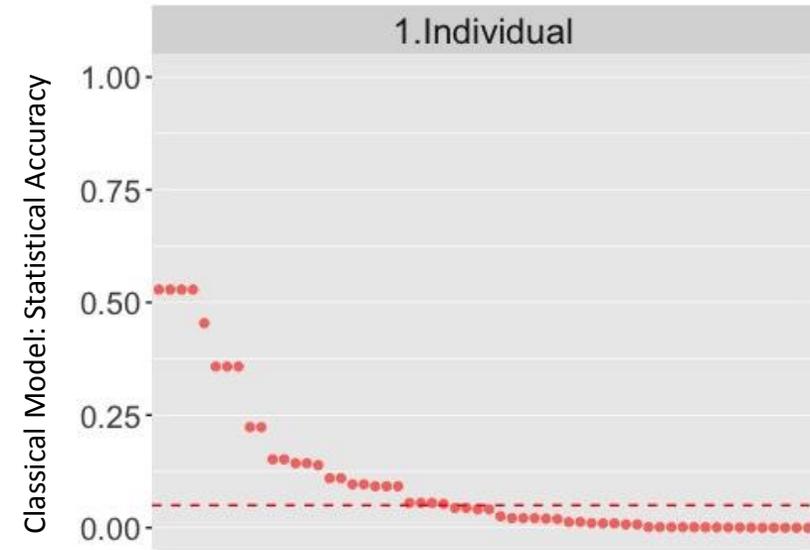
Which Aggregation to Trust?



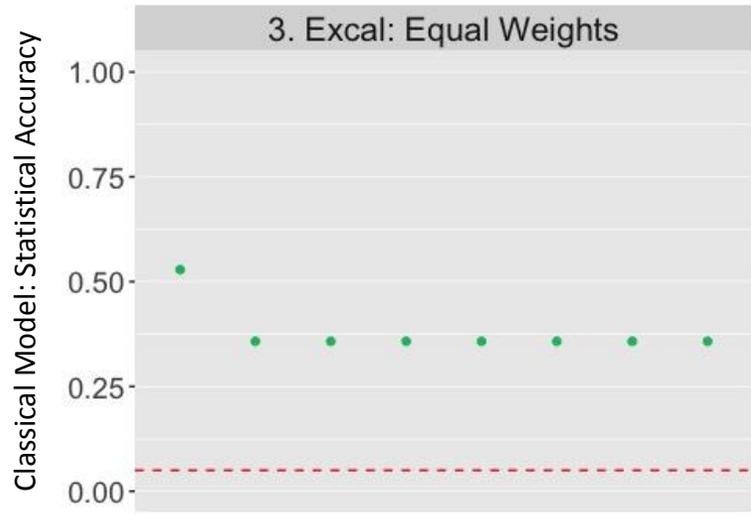
Which Aggregation to Trust?



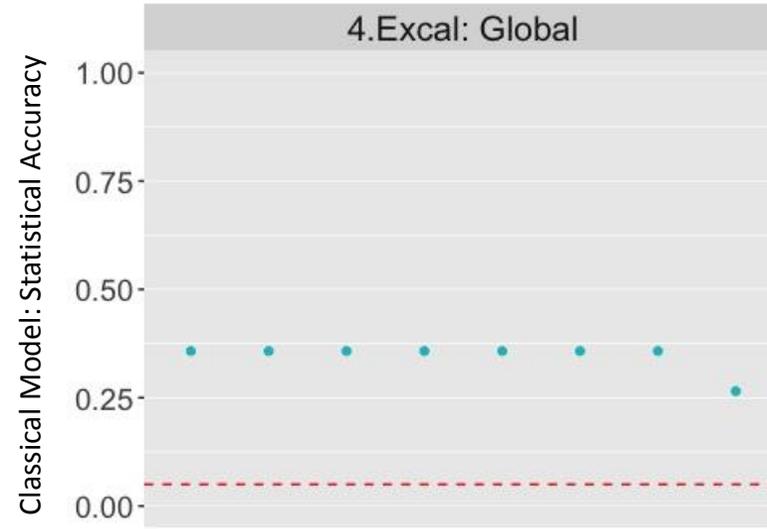
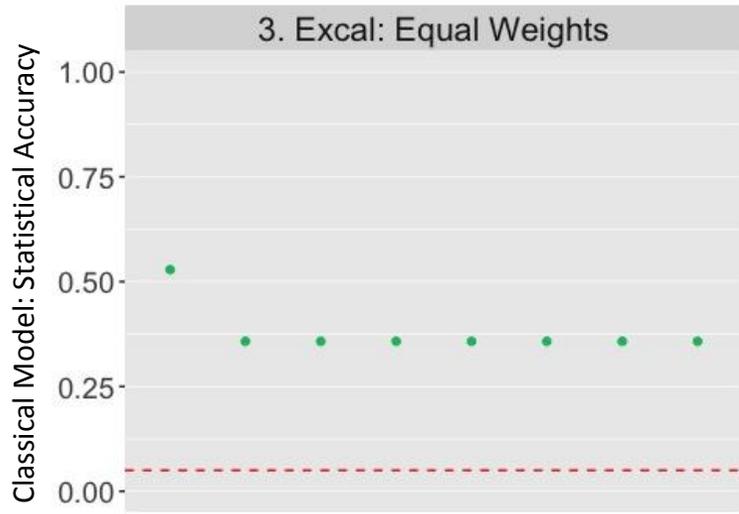
Which Aggregation to Trust?



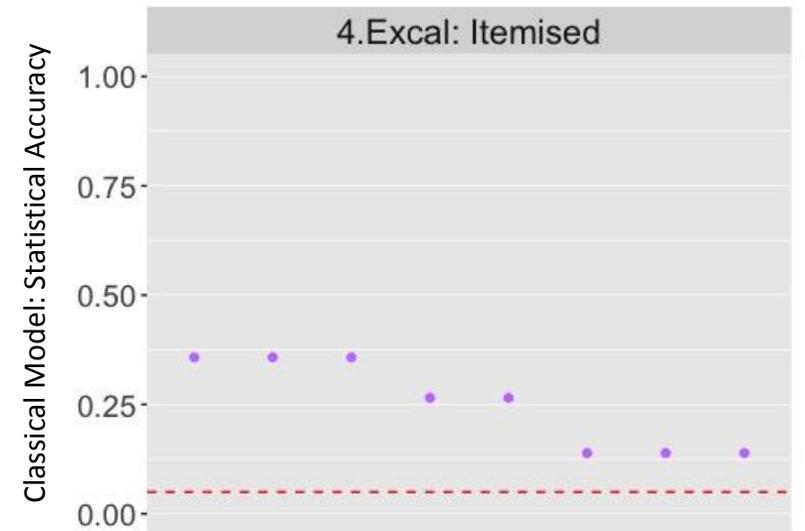
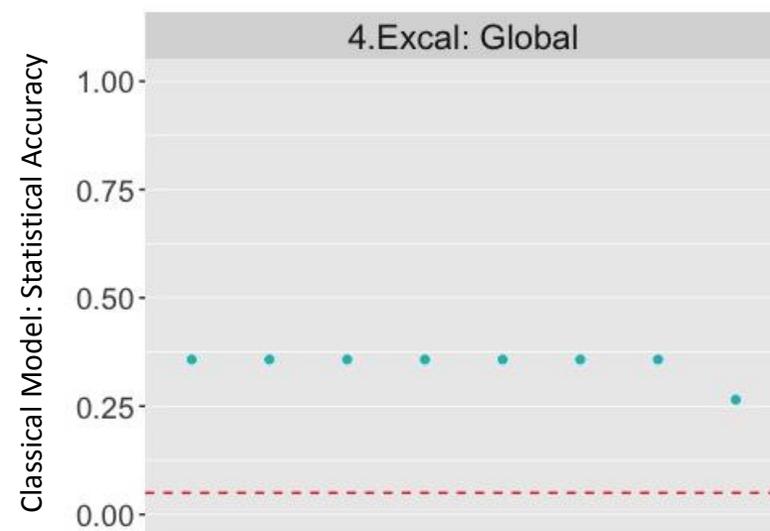
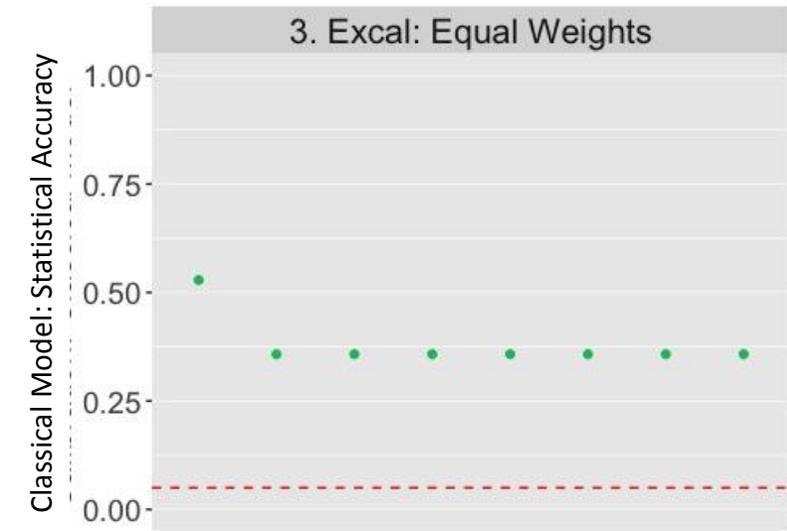
Which Aggregation to Trust?



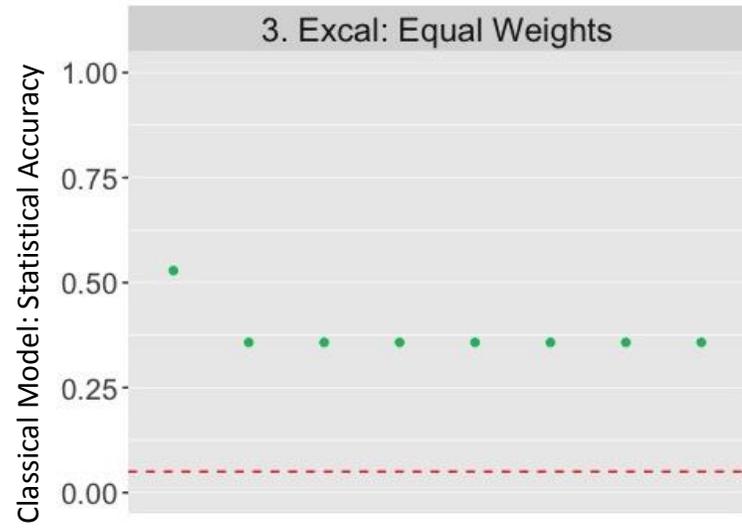
Which Aggregation to Trust?



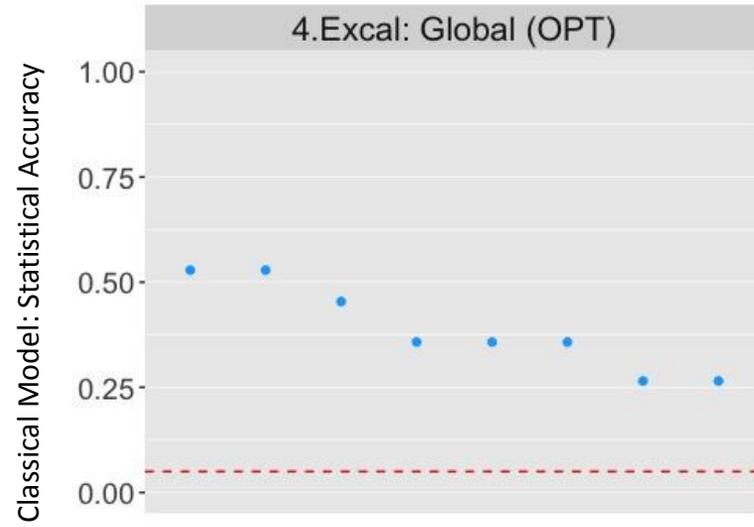
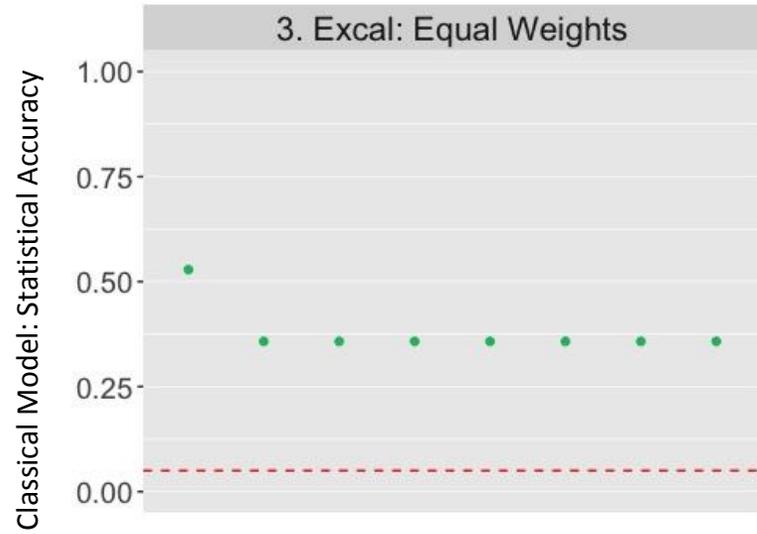
Which Aggregation to Trust?



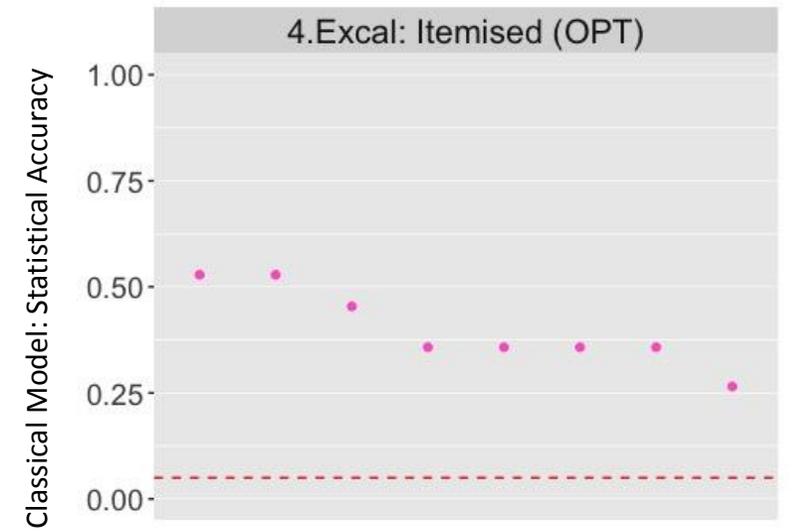
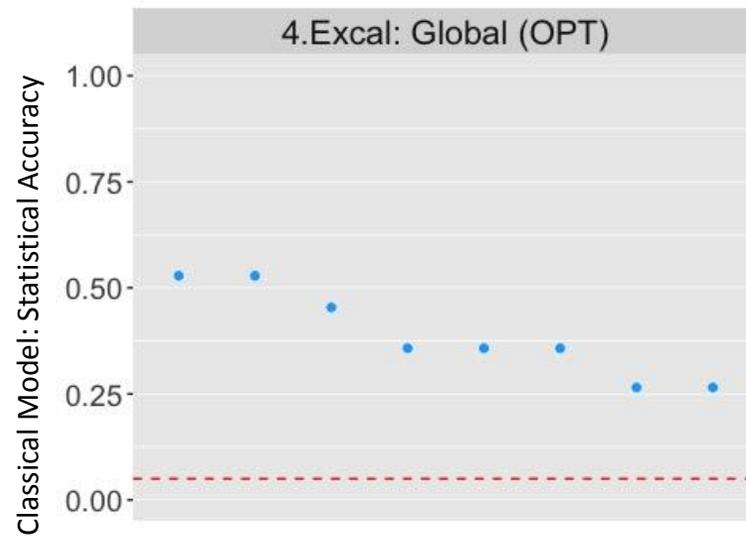
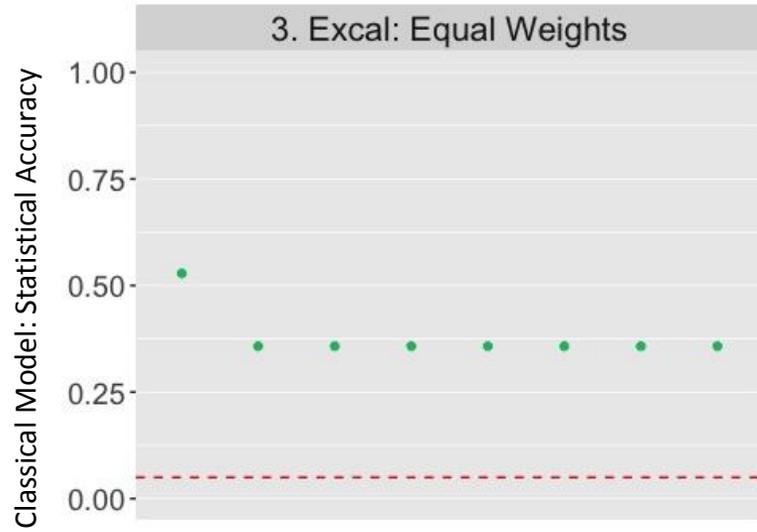
Which Aggregation to Trust?



Which Aggregation to Trust?

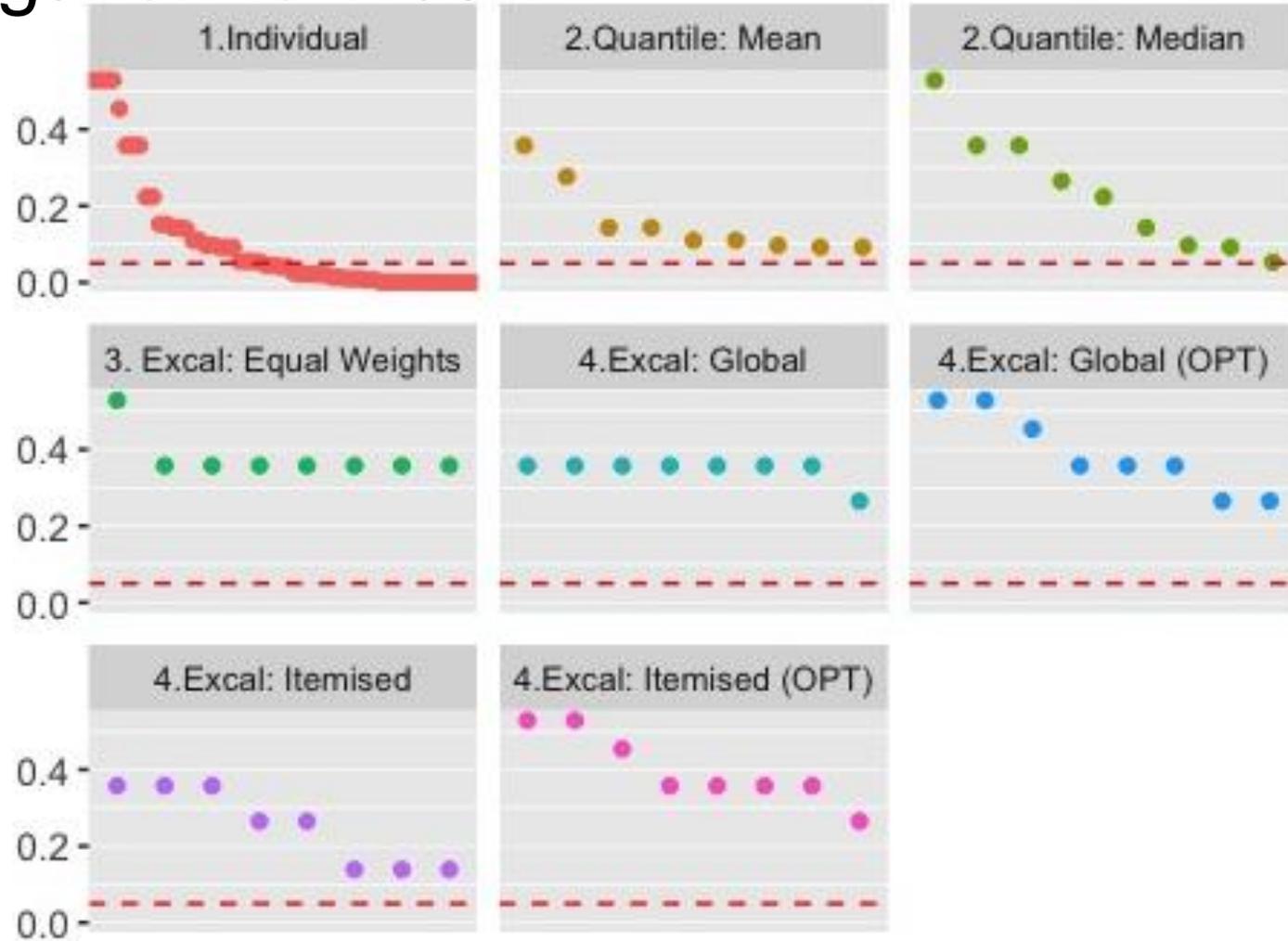


Which Aggregation to Trust?

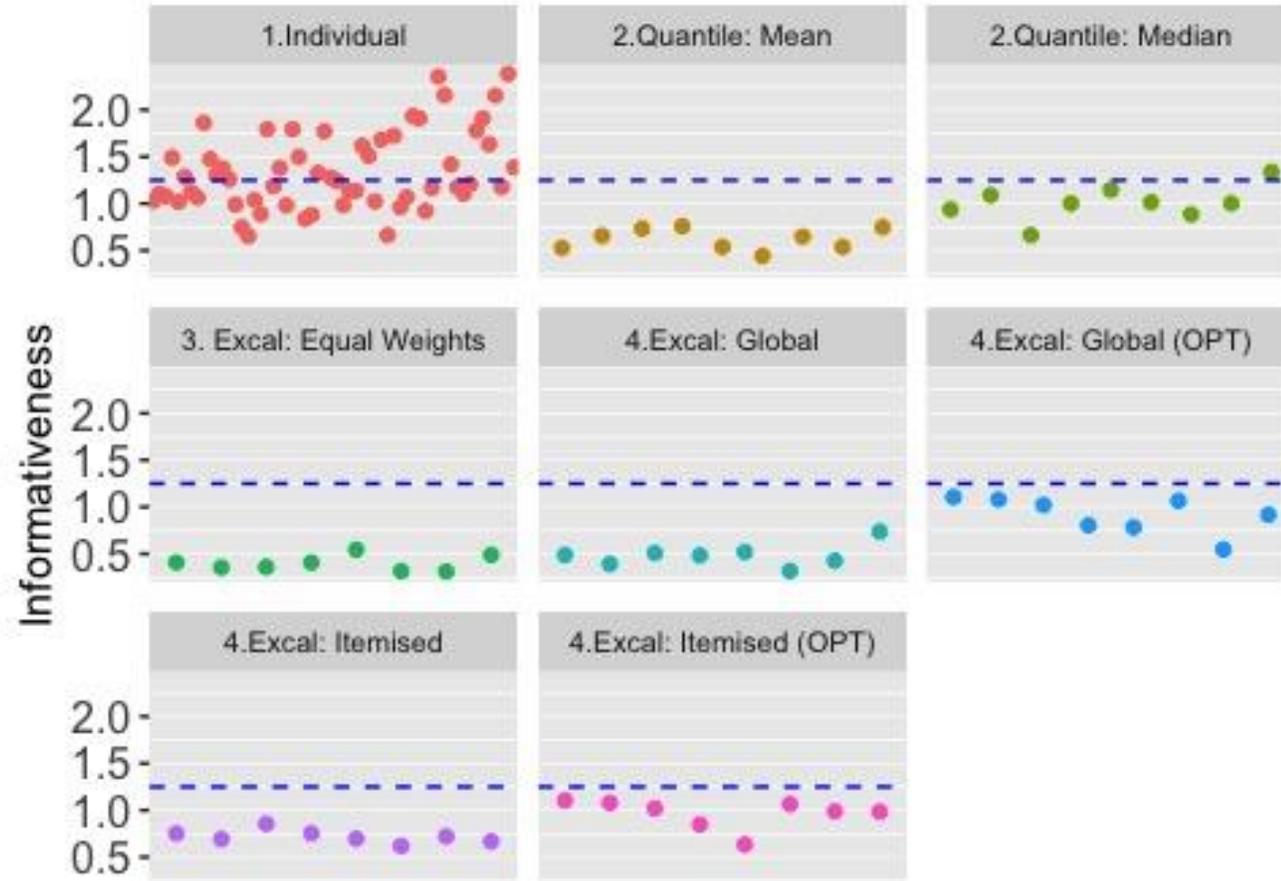


Which Aggregation to Trust?

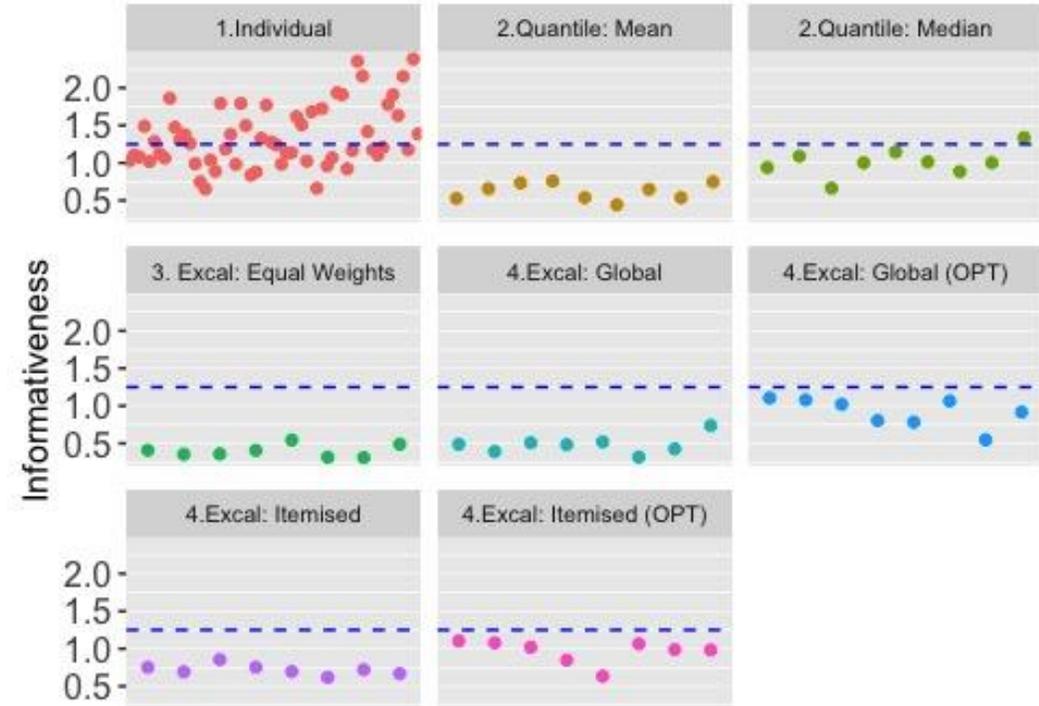
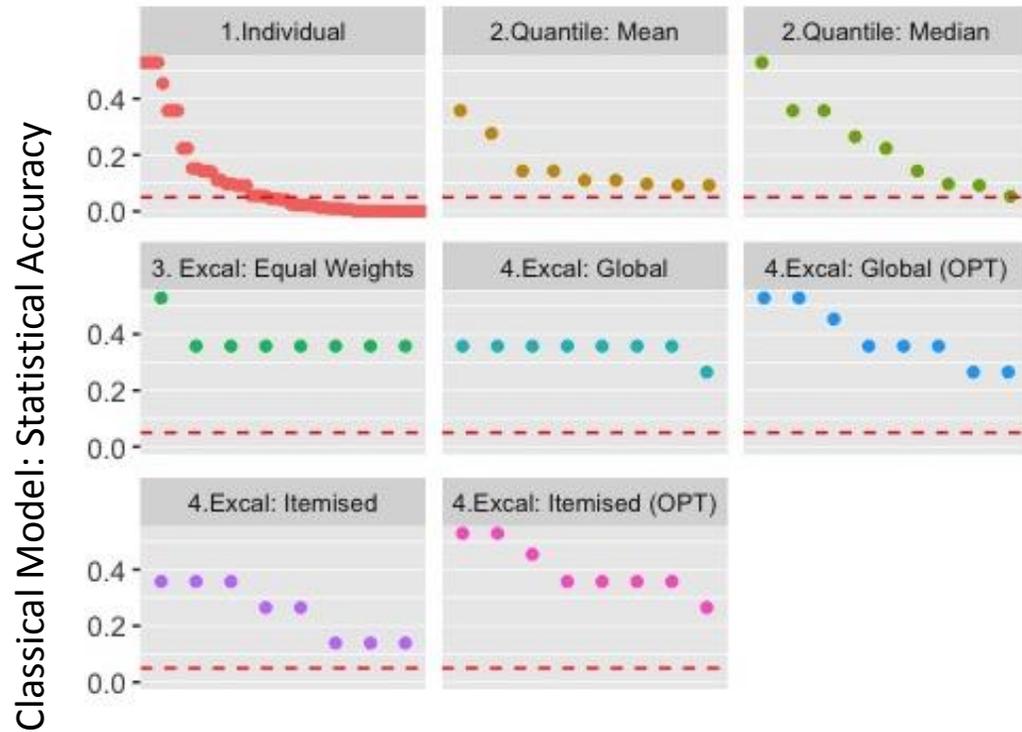
Classical Model: Statistical Accuracy



Classical Model Informativeness



Which Aggregation to Trust?

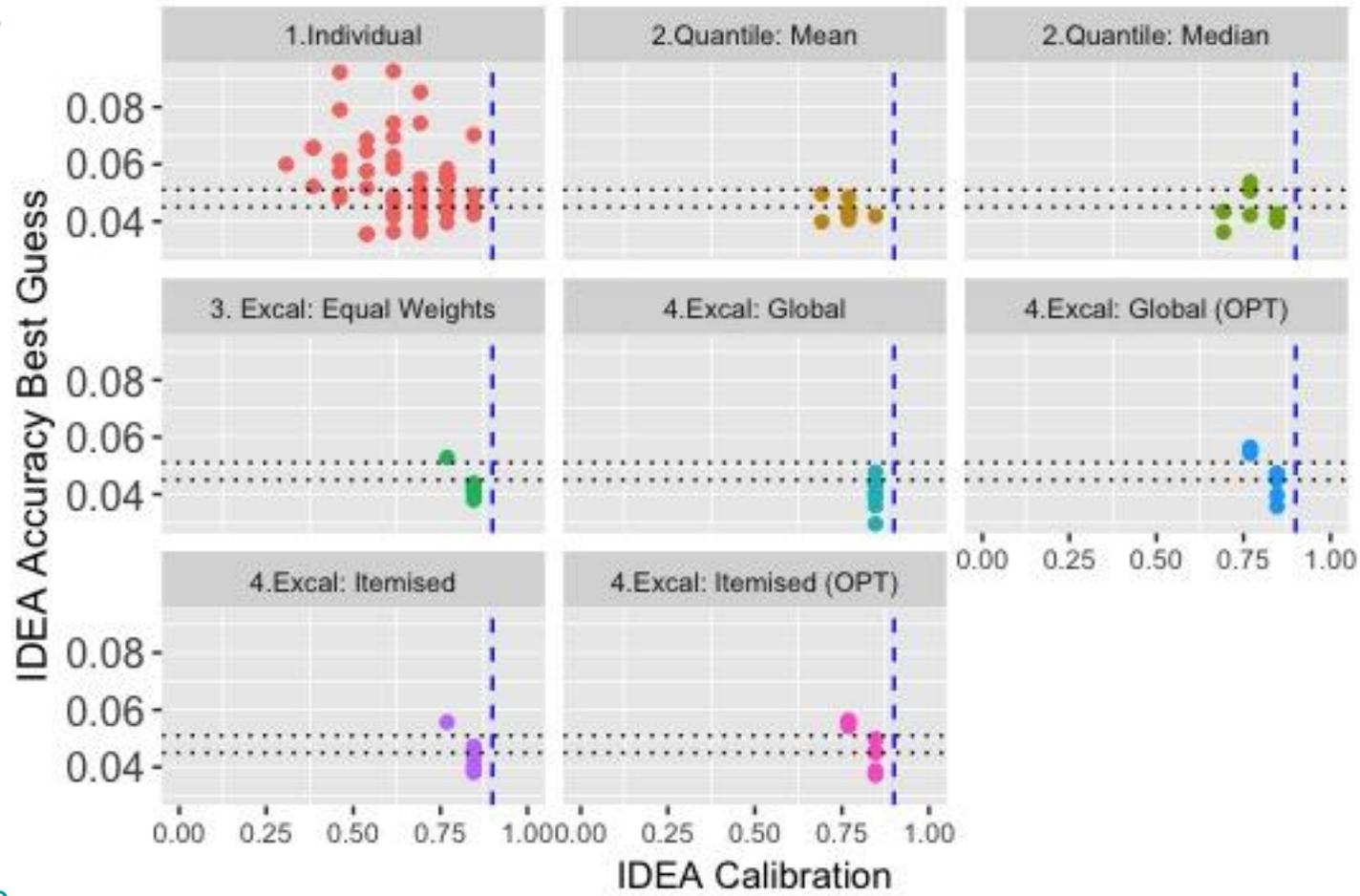


IDEA Scoring

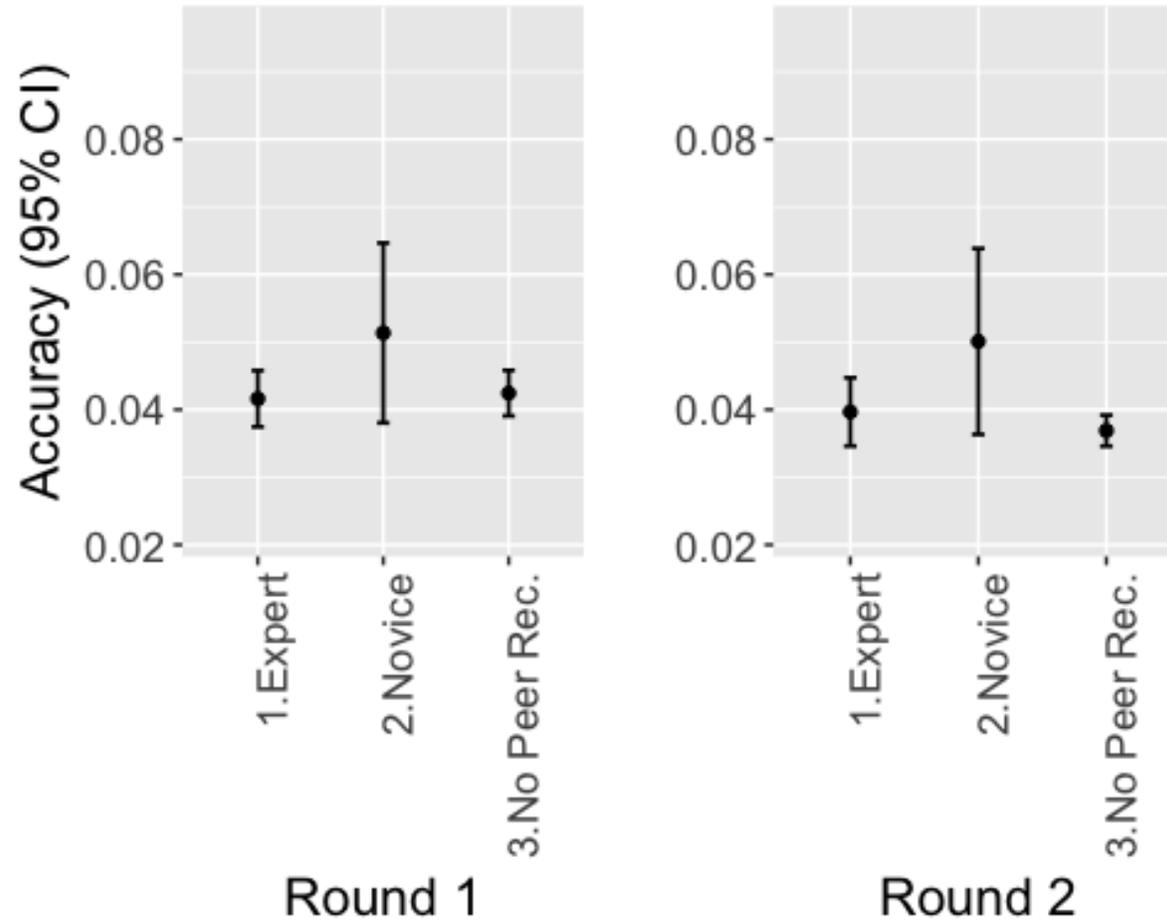
Less Accurate



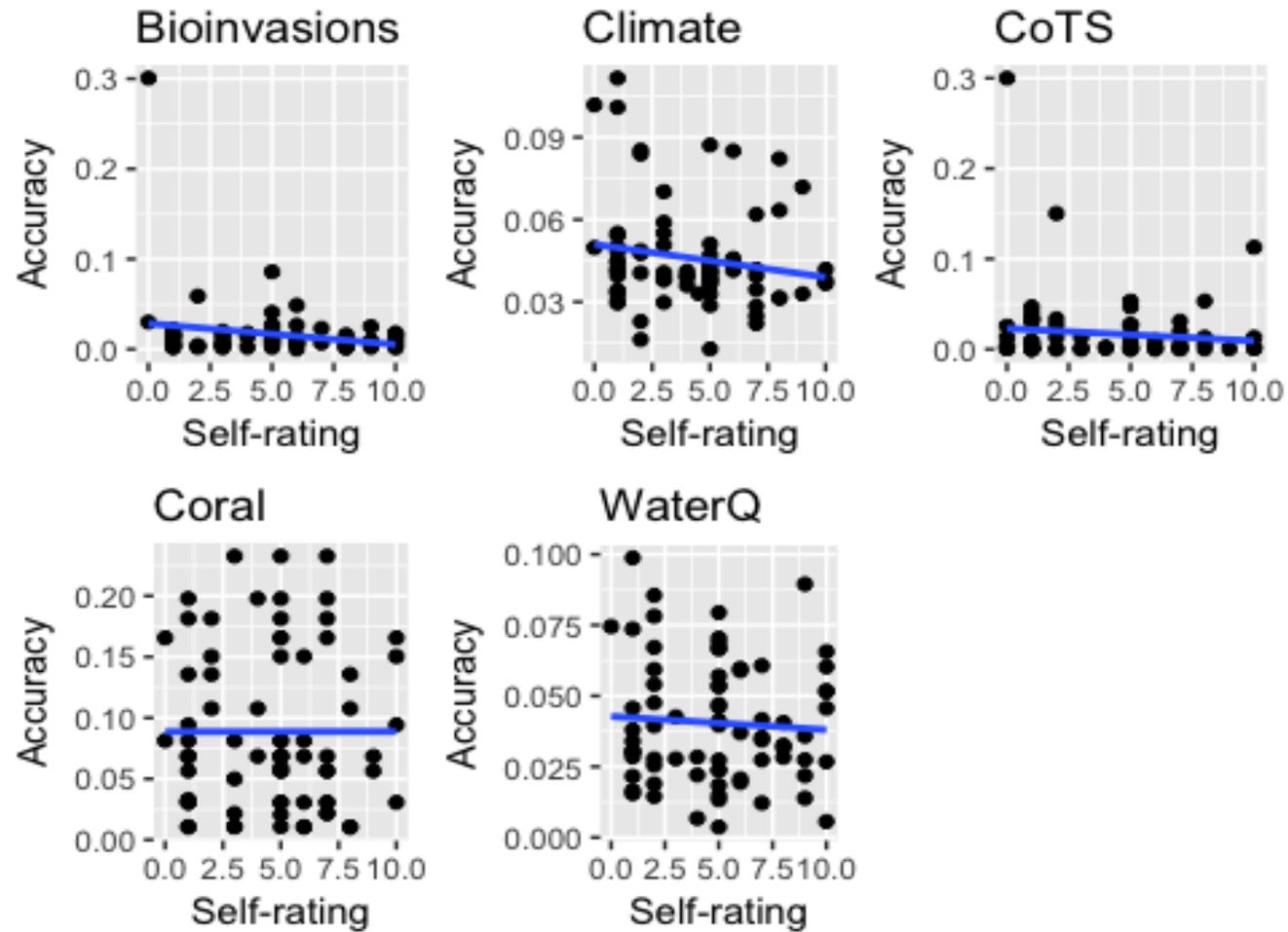
More Accurate



Experts vs Novices

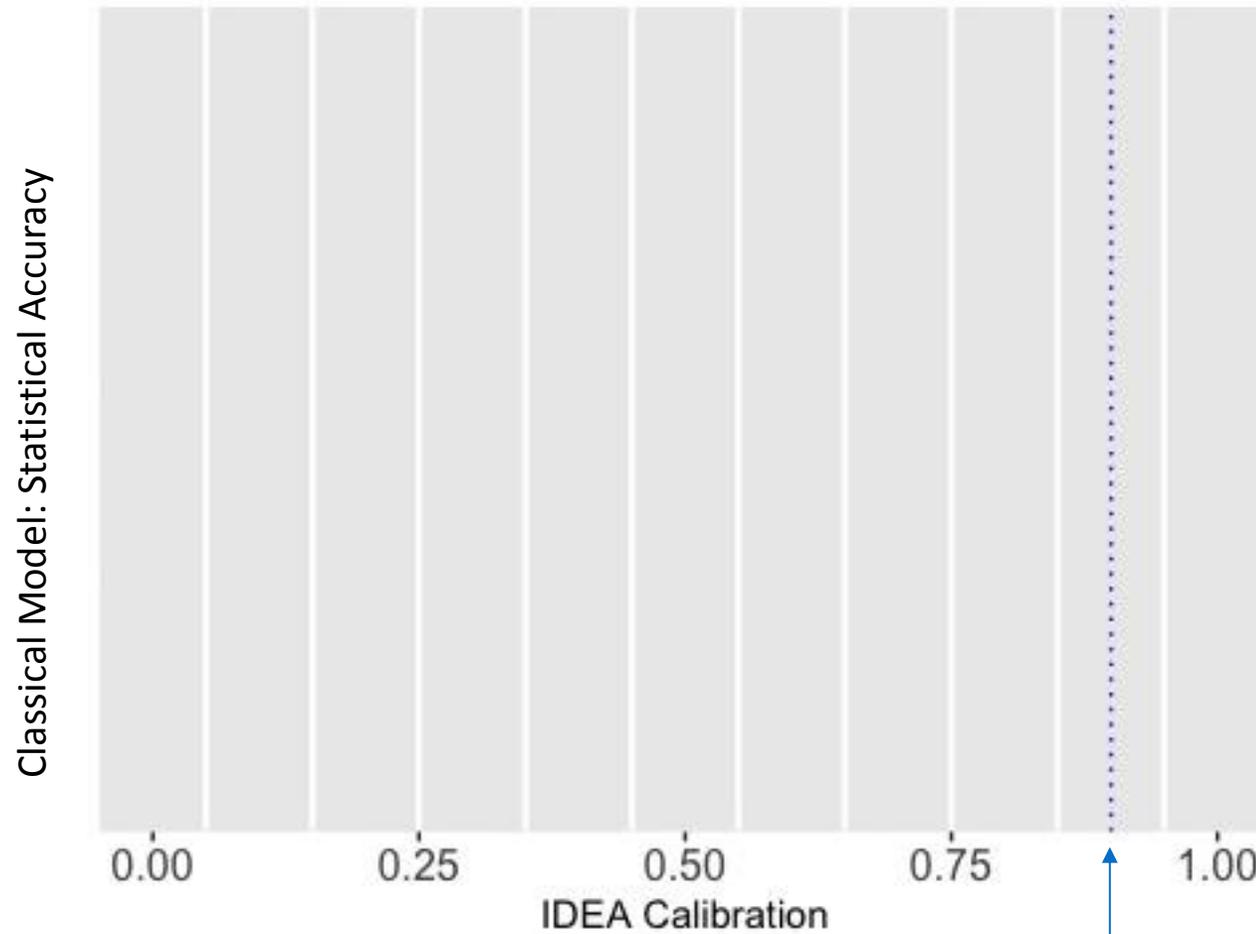


Self-rating and Accuracy?



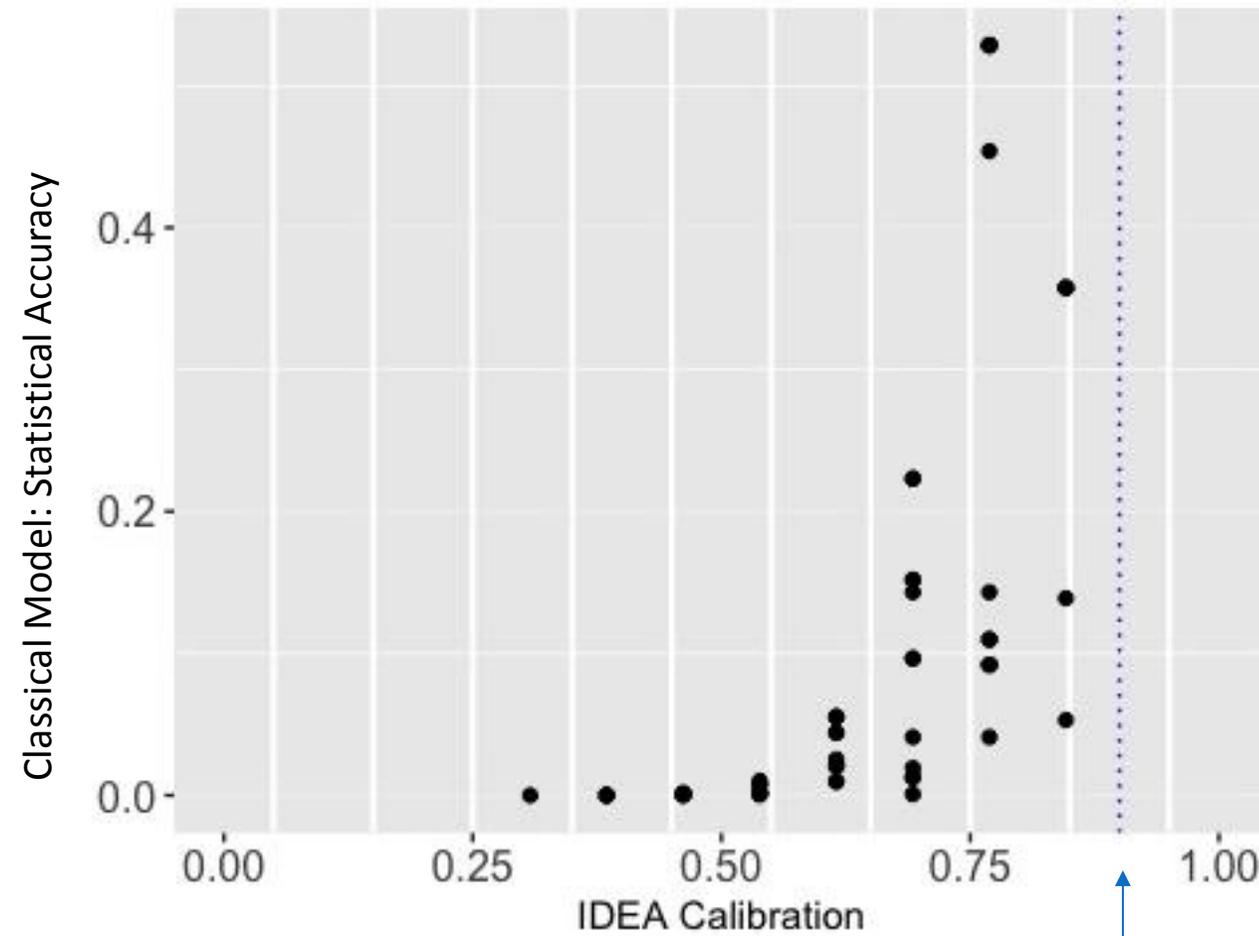
$R^2 < 0.02$,
 $P > 0.05$

Classical Model vs IDEA Scoring



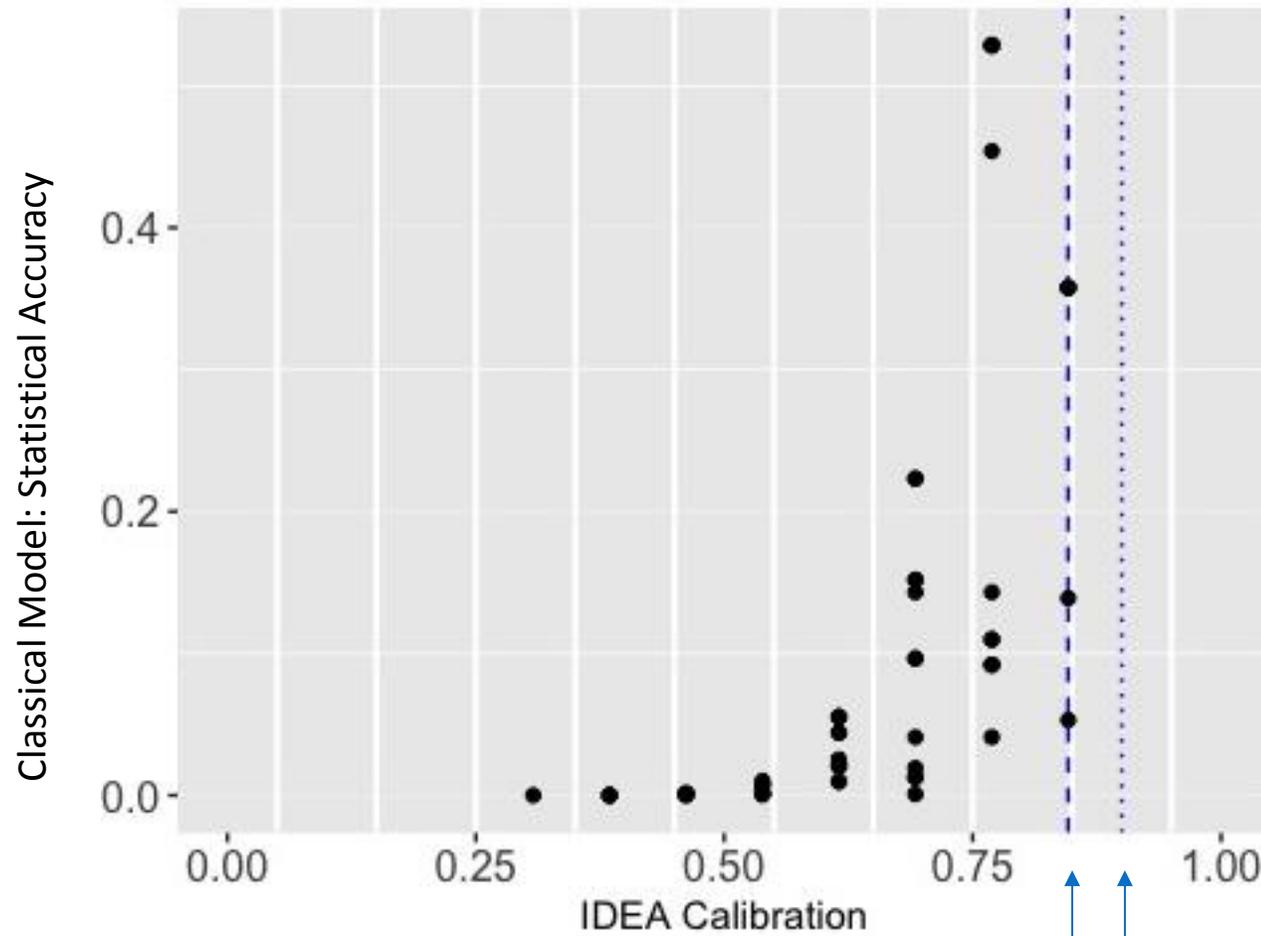
90%= Perfect Calibration (IDEA protocol)

Classical Model vs IDEA Scoring



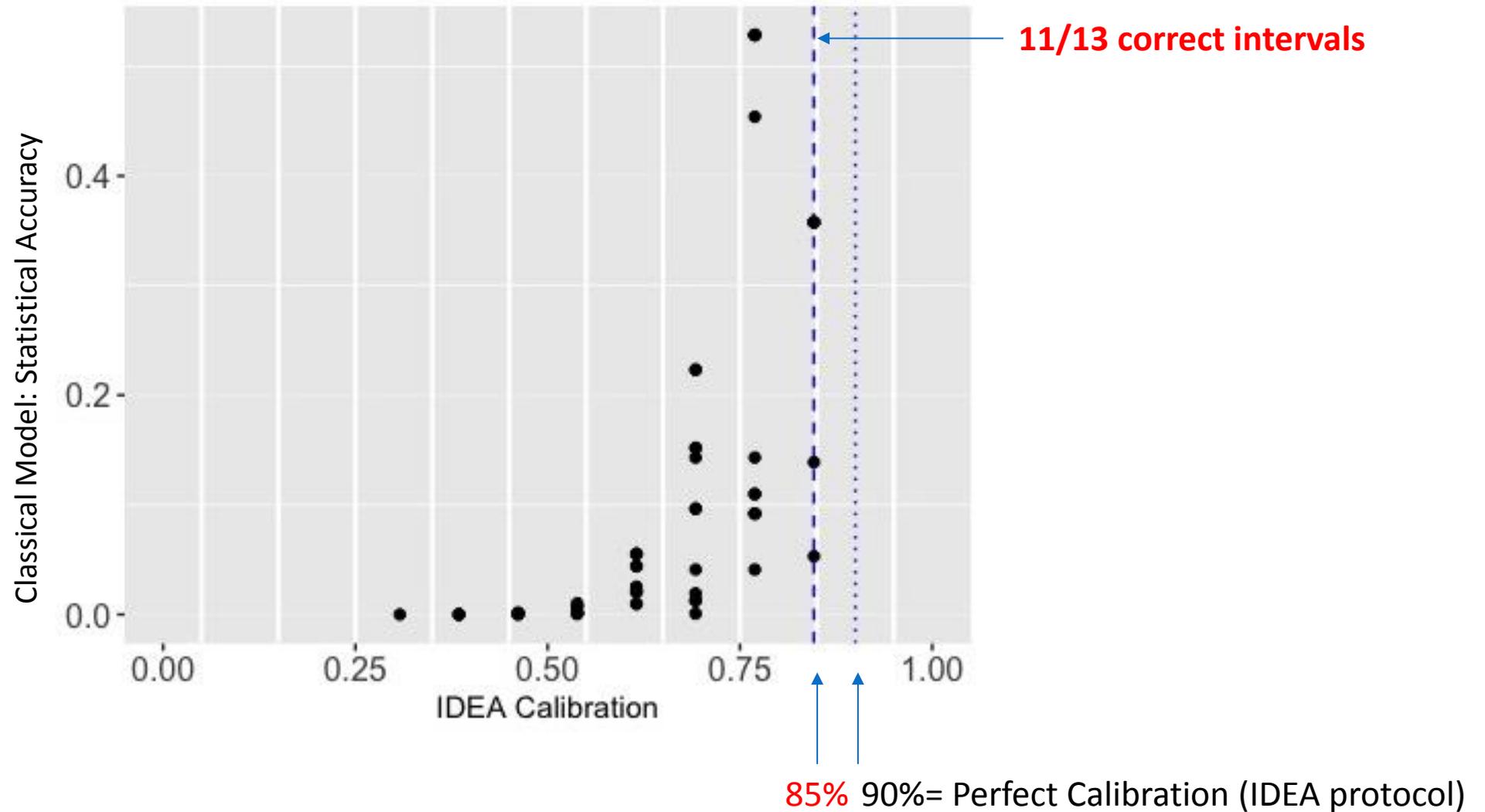
90%= Perfect Calibration (IDEA protocol)

Classical Model vs IDEA Scoring

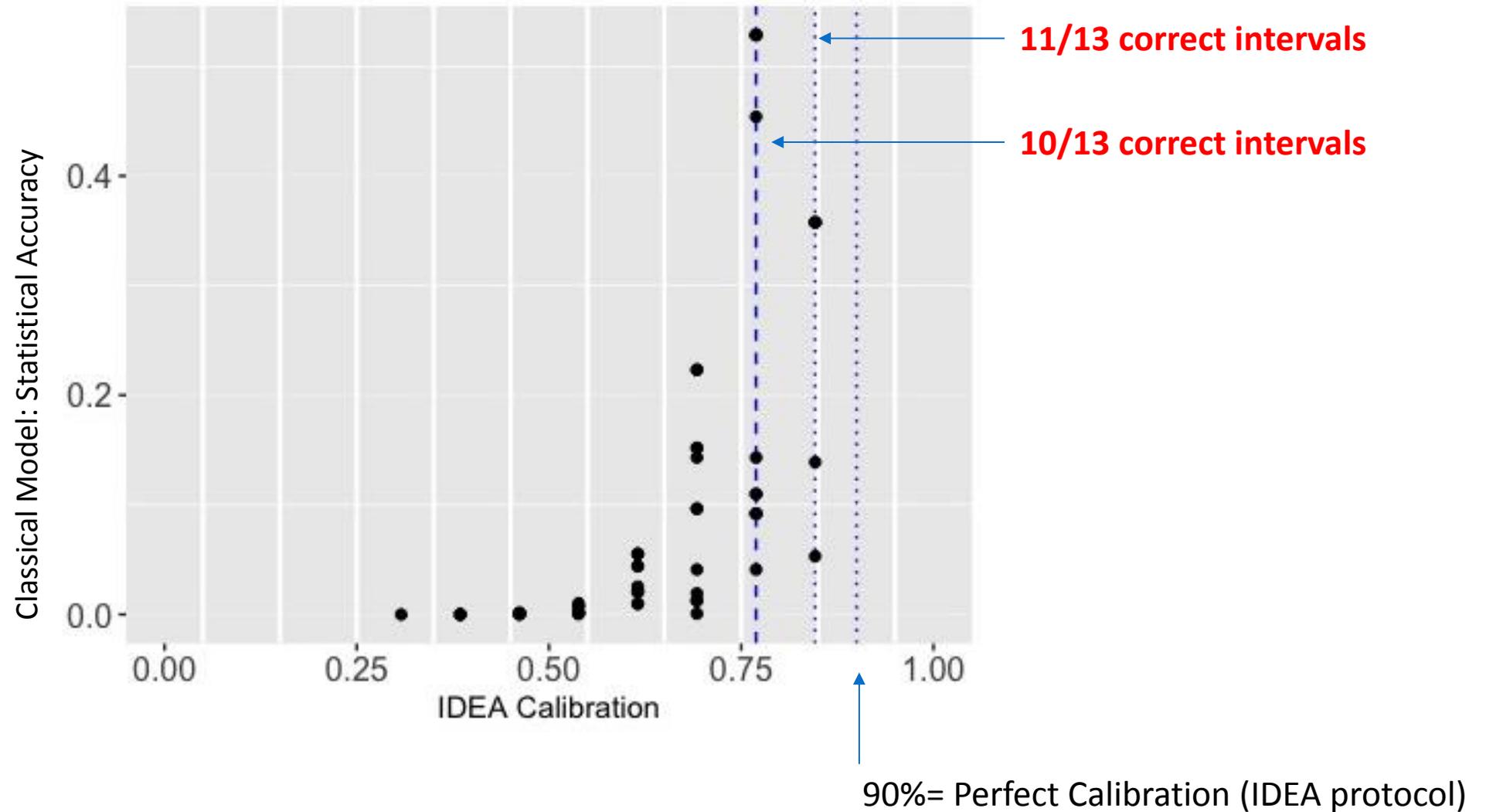


85% 90%= Perfect Calibration (IDEA protocol)

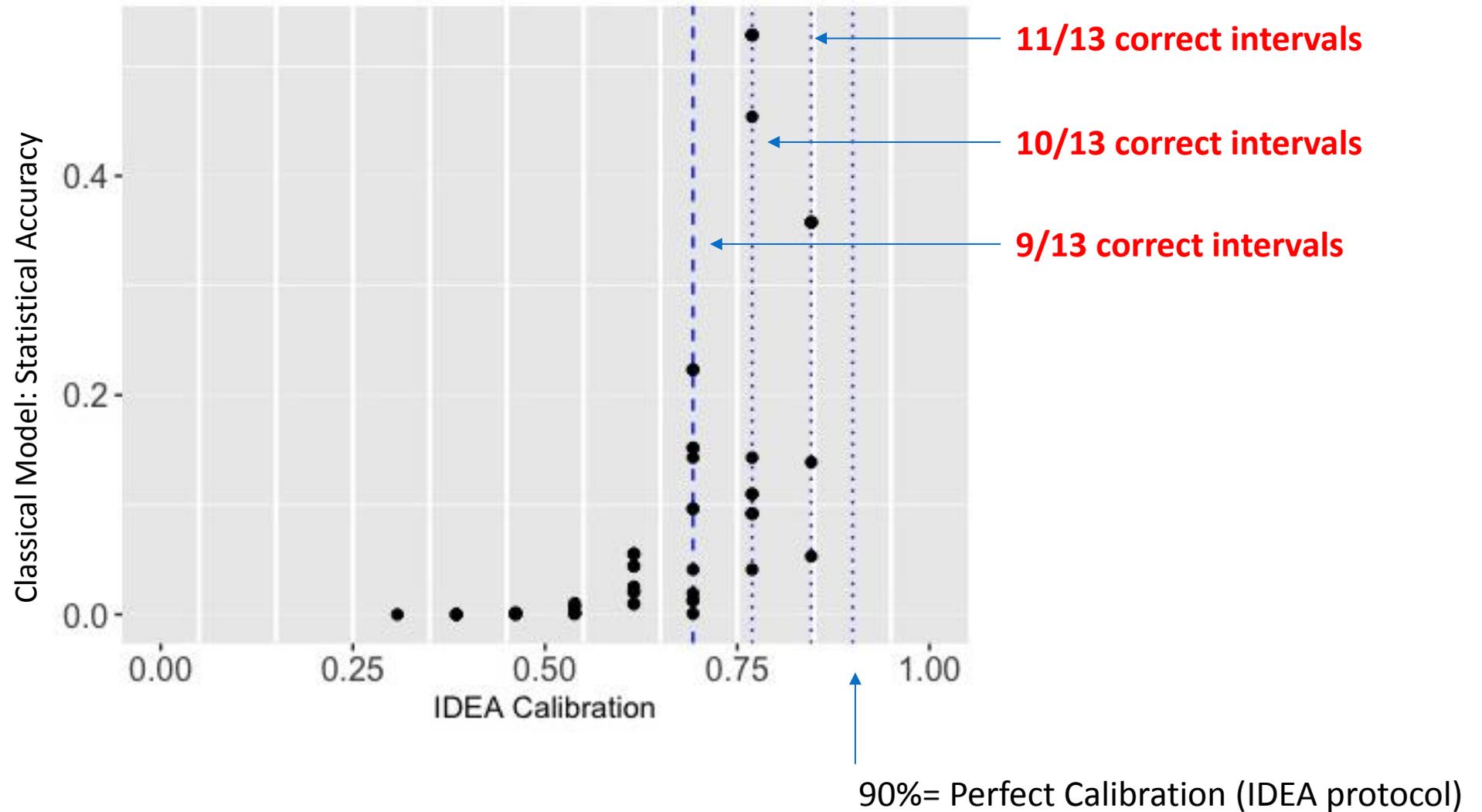
Classical Model vs IDEA Scoring



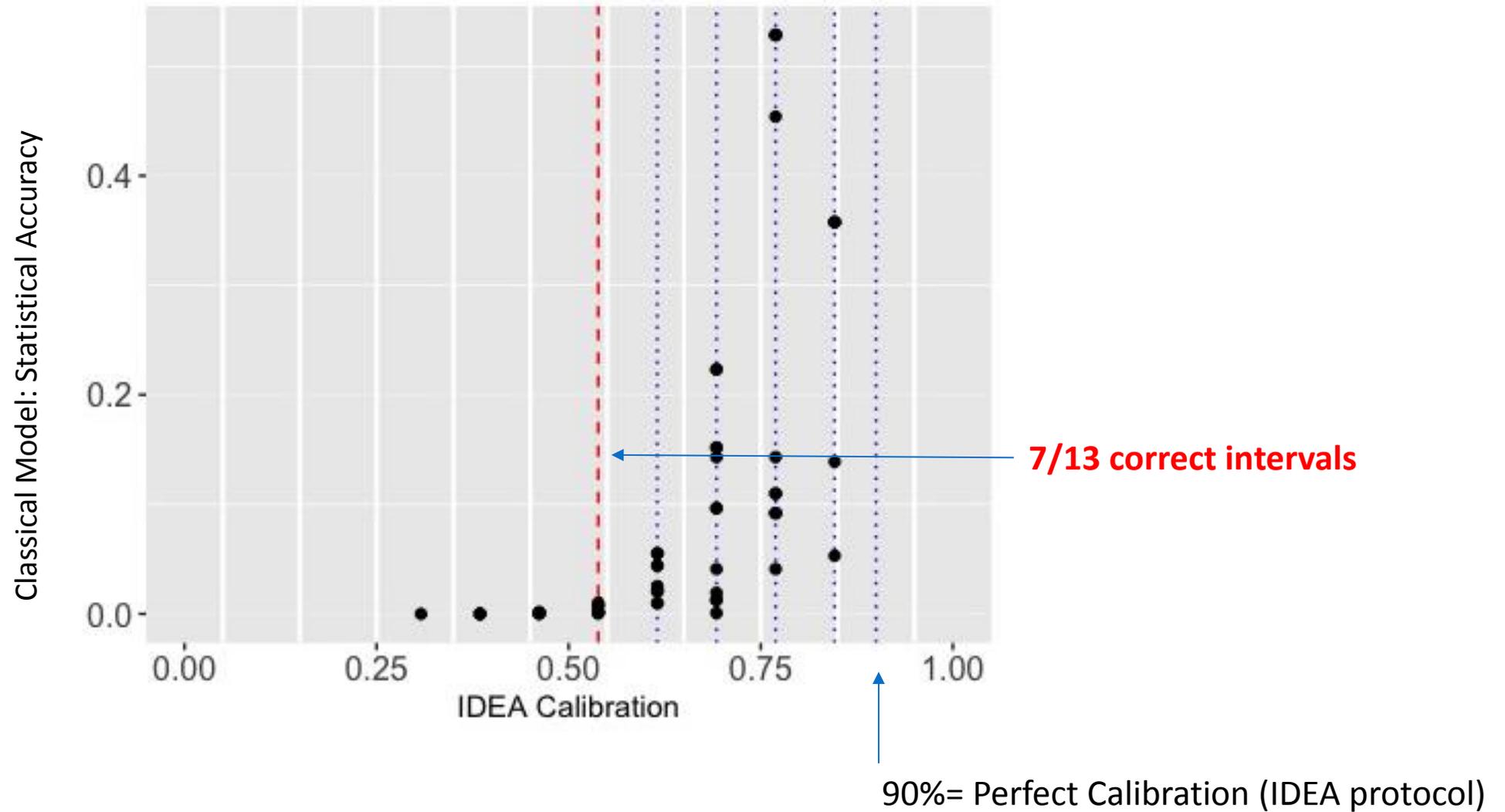
Classical Model vs IDEA Scoring



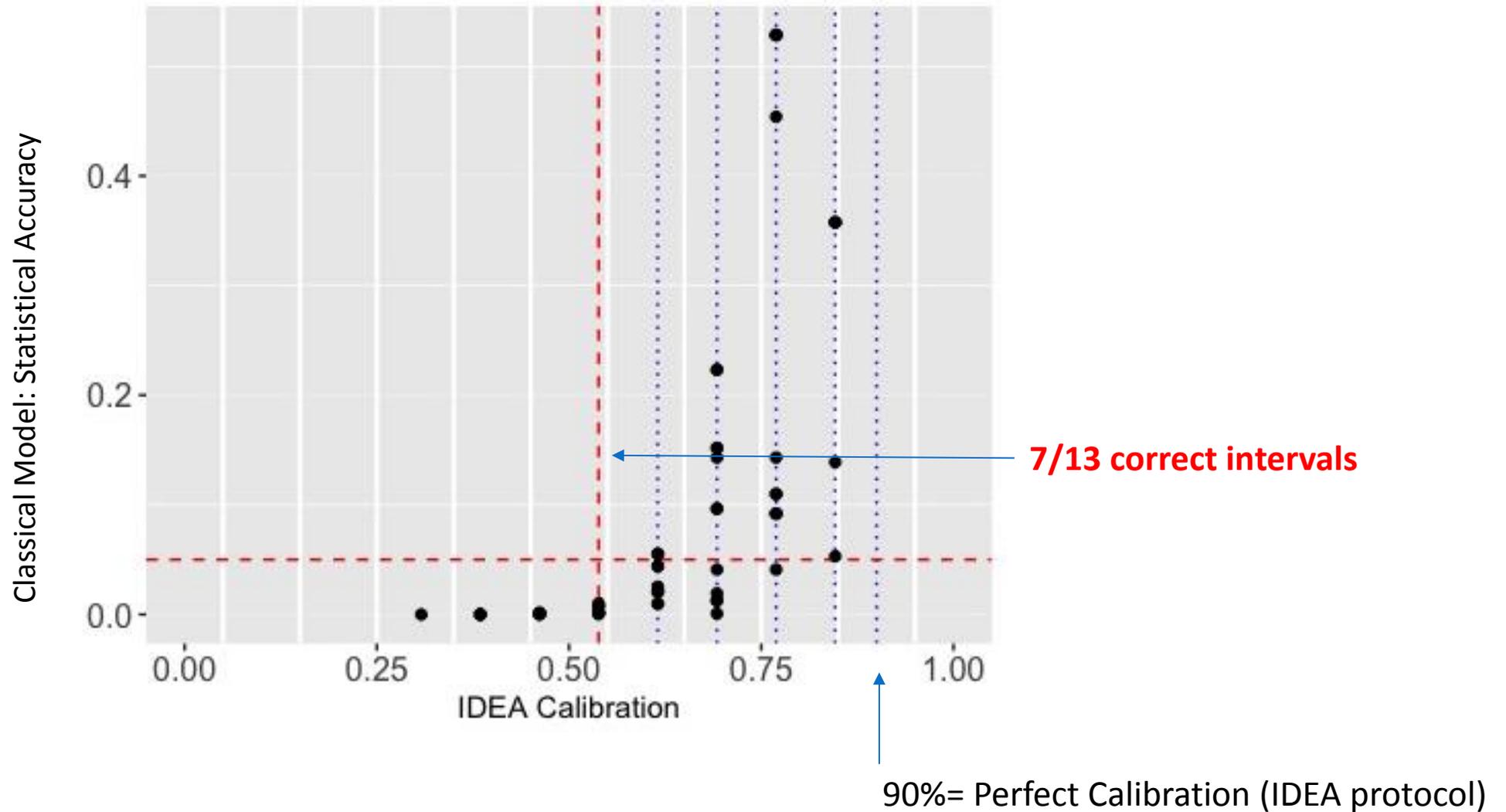
Classical Model vs IDEA Scoring



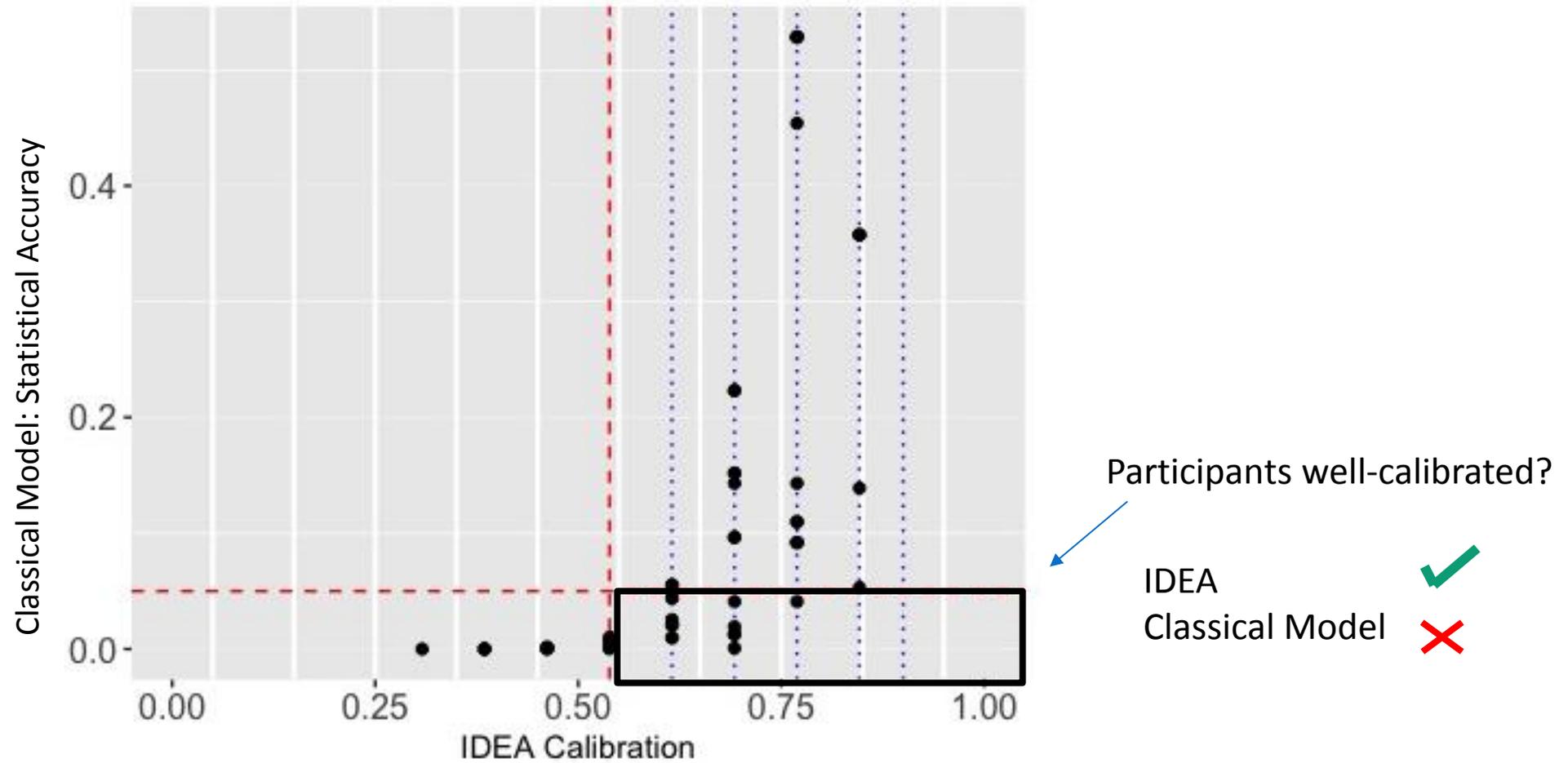
Classical Model vs IDEA Scoring



Classical Model vs IDEA Scoring



Classical Model vs IDEA Scoring



Naïve Performance-Based Weights: Classical Model

Abiotic + Geopolitical			
Weighted combination	Statistical Accuracy	Information	Rank
IT_OPT	0.6894	1.283	1
IT	0.6894	1.045	2
PW	0.6894	0.8546	3
PW_OPT	0.614	1.168	4
EW	0.614	0.8356	5

Naïve Performance-Based Weights: Classical Model

Abiotic + Geopolitical			
Weighted combination	Statistical Accuracy	Information	Rank
IT_OPT	0.6894	1.283	1
IT	0.6894	1.045	2
PW	0.6894	0.8546	3
PW_OPT	0.614	1.168	4
EW	0.614	0.8356	5

Cooke's Calibration = 0.6894			
<5 th	5 th - 50 th	50 th -.95 th	>95 th
	X X X X X X X	X X X X X X X	
X			
1	6	6	0

Cooke's Calibration = 0.614			
<5 th	5 th - 50 th	50 th -.95 th	>95 th
	X X X X X X	X X X X X X X	
X			
1	5	7	0

Naïve Performance-Based Weights: Calibration Classical Model

Abiotic + Geopolitical			
Weighted combination	Statistical Accuracy	Information	Rank
IT_OPT	0.6894	1.283	1
IT	0.6894	1.045	2
PW	0.6894	0.8546	3
PW_OPT	0.614	1.168	4
EW	0.614	0.8356	5

Biotic			
Weighted combination	Statistical Accuracy	Information	Rank
PW_OPT	0.1586	0.8688	1
EW	0.1586	0.7341	2
PW	0.1586	0.707	3
IT_OPT	0.1008	0.9319	4
IT	0.1008	1.024	5

Cooke's Calibration = 0.6894			
<5 th	5 th - 50 th	50 th -95 th	>95 th
	X X X X X X	X X X X X X	
X			
1	6	6	0

Cooke's Calibration = 0.614			
<5 th	5 th - 50 th	50 th -95 th	>95 th
	X X X X X	X X X X X X X	
X			
1	5	7	0

Naïve Performance-Based Weights: Calibration Classical Model

Abiotic + Geopolitical			
Weighted combination	Statistical Accuracy	Information	Rank
IT_OPT	0.6894	1.283	1
IT	0.6894	1.045	2
PW	0.6894	0.8546	3
PW_OPT	0.614	1.168	4
EW	0.614	0.8356	5

Biotic			
Weighted combination	Statistical Accuracy	Information	Rank
PW_OPT	0.1586	0.8688	1
EW	0.1586	0.7341	2
PW	0.1586	0.707	3
IT_OPT	0.1008	0.9319	4
IT	0.1008	1.024	5

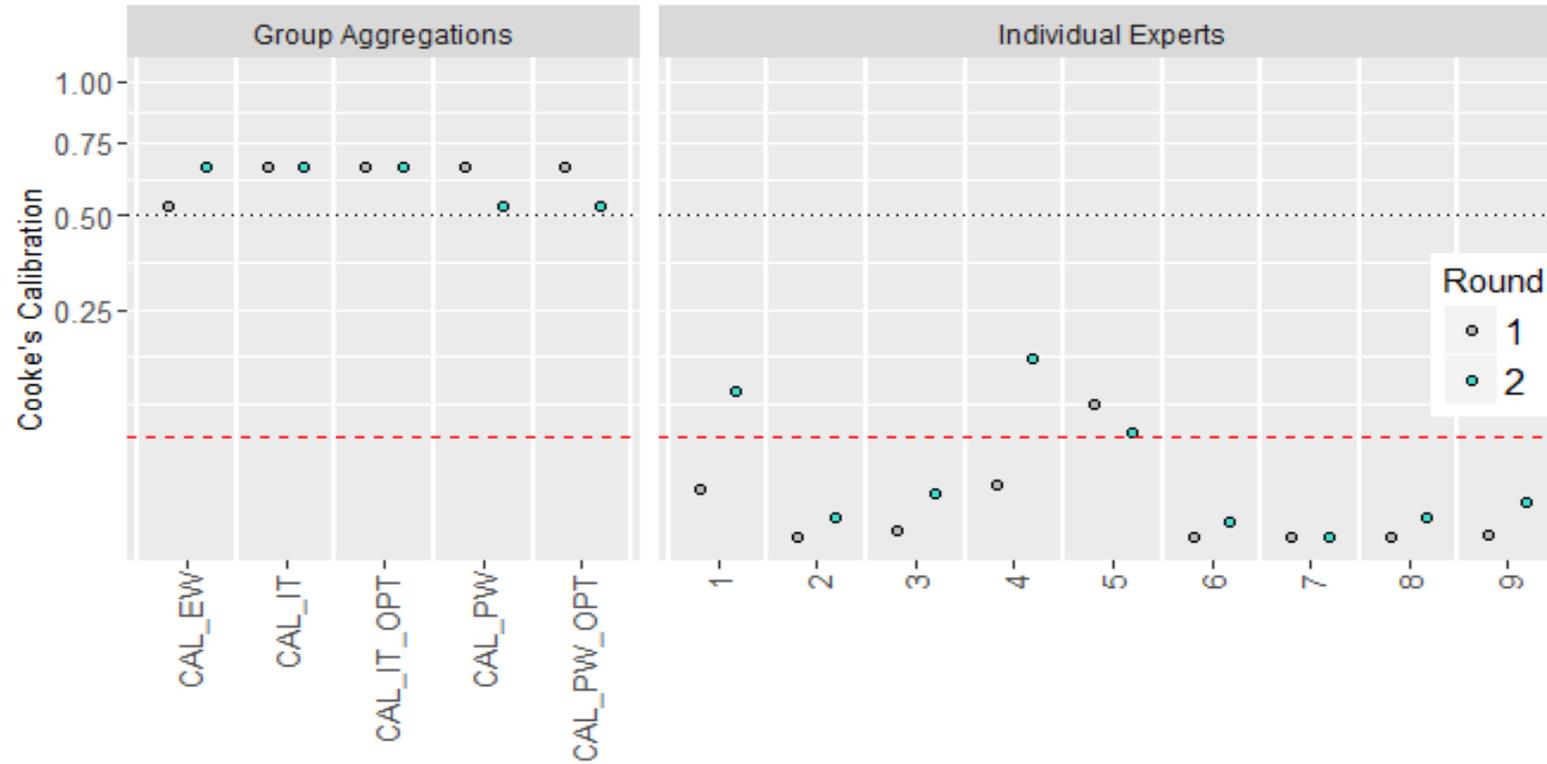
Cooke's Calibration = 0.6894			
<5 th	5 th - 50 th	50 th -95 th	>95 th
X	X X X X X	X X X X X	
1	6	6	0

Cooke's Calibration = 0.614			
<5 th	5 th - 50 th	50 th -95 th	>95 th
X	X X X X X	X X X X X X	
1	5	7	0

Cooke's Calibration = 0.159			
<5 th	5 th - 50 th	50 th -95 th	>95 th
X X	X X	X X	
2	2	2	0

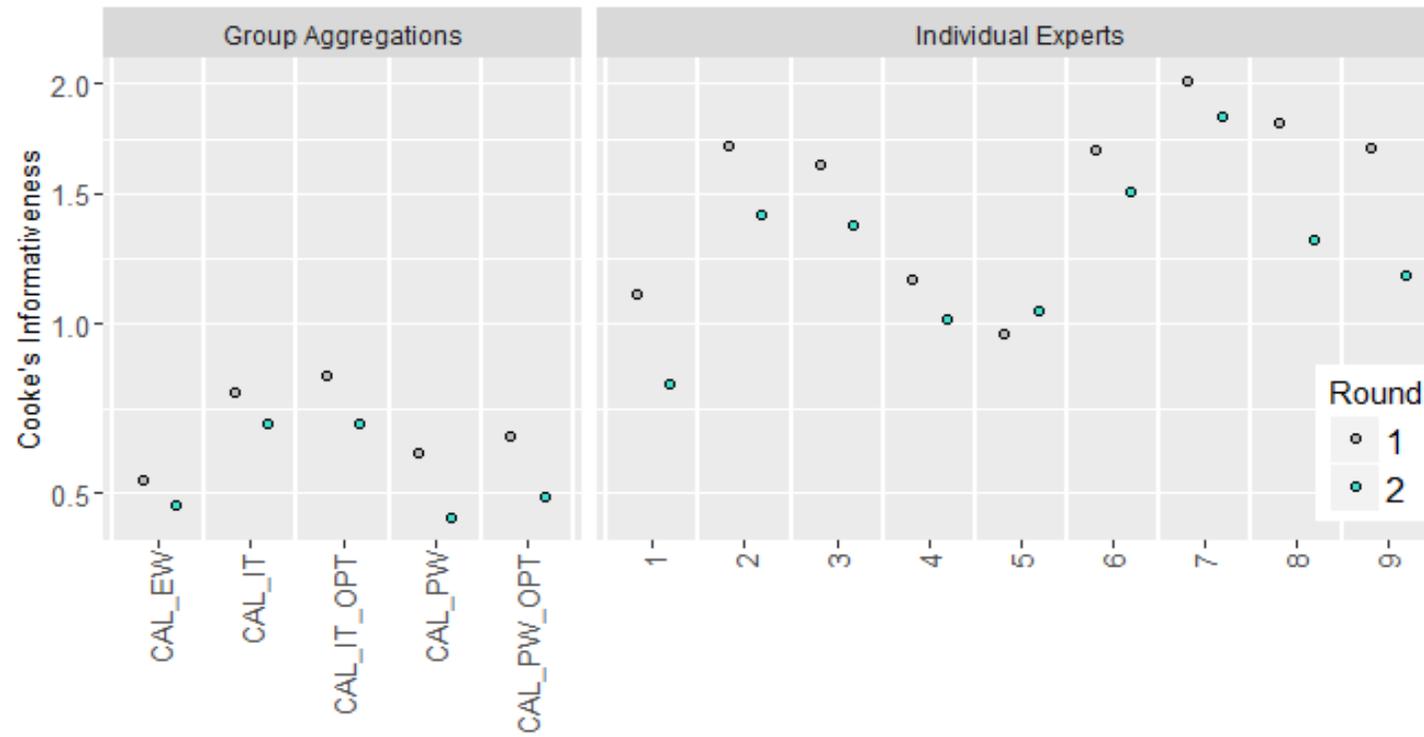
Cooke's Calibration = 0.101			
<5 th	5 th - 50 th	50 th -95 th	>95 th
X X	X X X	X	
2	3	1	0

A Second Example: Fault Engineering



14 Questions

A Second Example: Fault Engineering



Should Test Questions Be Developed?

YES

- To confirm the group/ Individuals have knowledge, and can accurately communicate their knowledge.
- Avoid analyst bias.
- Avoid pre-judgement of expertise.
- Important for validating expertise (legal challenges?).

Should performance-based weight be used?

Yes

- If they are better calibrated then **Yes**.
- Can't tell unless you develop test questions
- Study indicates that the weights may not “over-optimize”.

Are there challenges still to be overcome?

Yes

- More studies required in conservation domains to show how test questions could be developed.
- More cross-validation studies required to explore if there are conditions under which development of test questions could lead to over-optimisation.
- Guidance for developing test questions ([I have started to compile this.- would love your input](#))
- Showing the difference that it makes to a decision. Does it save lives, money etc.
- Excalibur recoded in an open access program like R.

Conclusions

- Difficult to define good test questions.
- Weighting may be sensitive to the questions asked but did not over-optimize.
- Difference in how the aggregations would be ranked between IDEA and Classical Model.
- Further exploration on the best way to adapt 4-Step Elicitation to suit Classical Model / performance-based weighting
- Can only investigate the performance if we include test questions.

Acknowledgements

Supervisors: Prof. Mark Burgman, Dr. Anca Hanea, Dr. Terry Walshe, Dr Fiona Fidler

All of the experts who enabled me to test them 😊

Victoria Hemming
PhD Candidate

Centre of Excellence for Biosecurity Risk
Analysis

The University of Melbourne
hemmingv@student.unimelb.edu.au



@v_hemming



THE UNIVERSITY OF
MELBOURNE

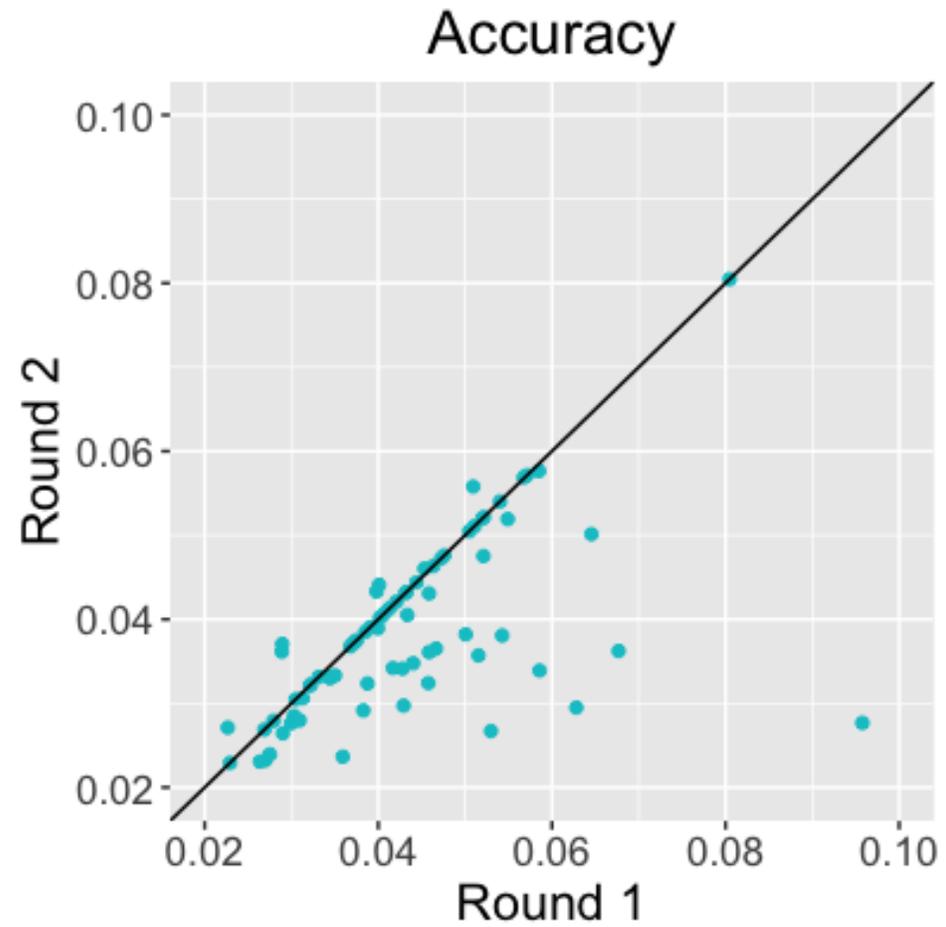
Caveats:

- Results of 1 study.

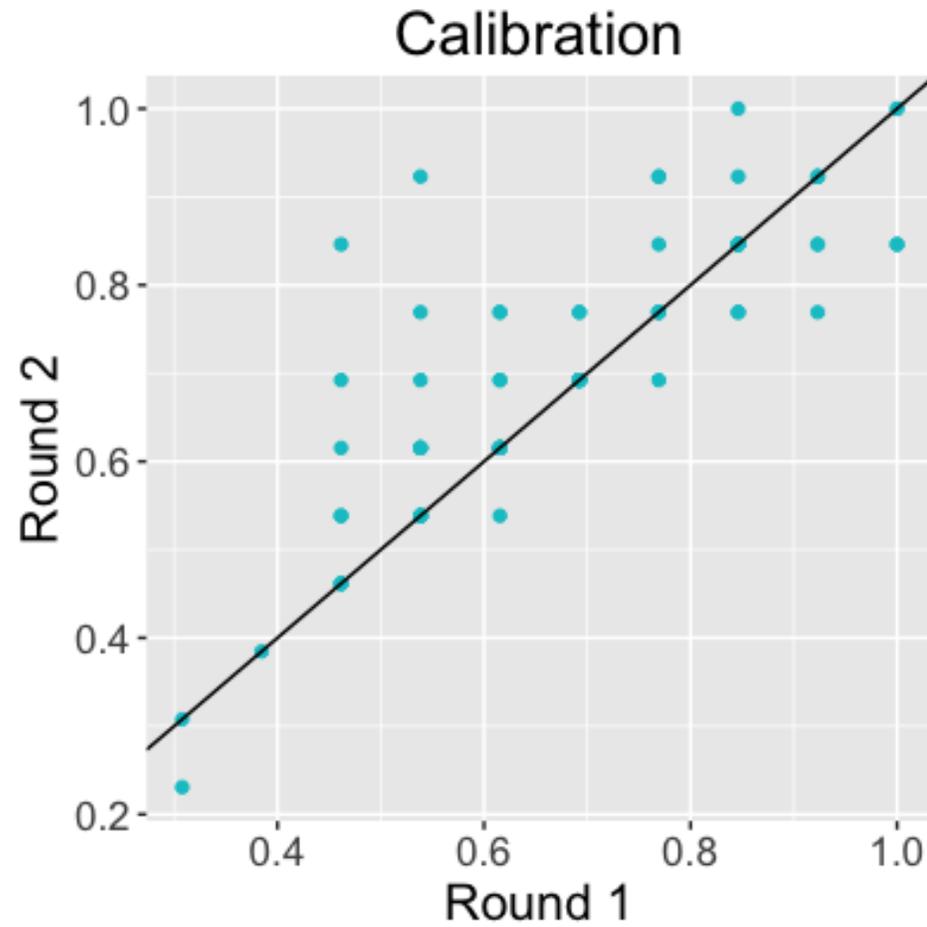
Future directions:

- Does it improve Decision Quality?
- To explore groups comprised of experts vs novices.

The value of a second round?



The Value of a Second Round?





1.
Conservation reliant
on
Expert judgement



2.

Unstructured and
opaque methods
abundant

Vague questions...



3.

Results ambiguous



3.

IDEA protocol and
structured elicitation
proposed as
alternatives



3. The next step?

Performance-based weights

Great Barrier Reef Elicitation

*Images not my own

Biotic: Bleaching, Crown of Thorns, Invasive Species, Disease, Threatened Species, Predators, Culling



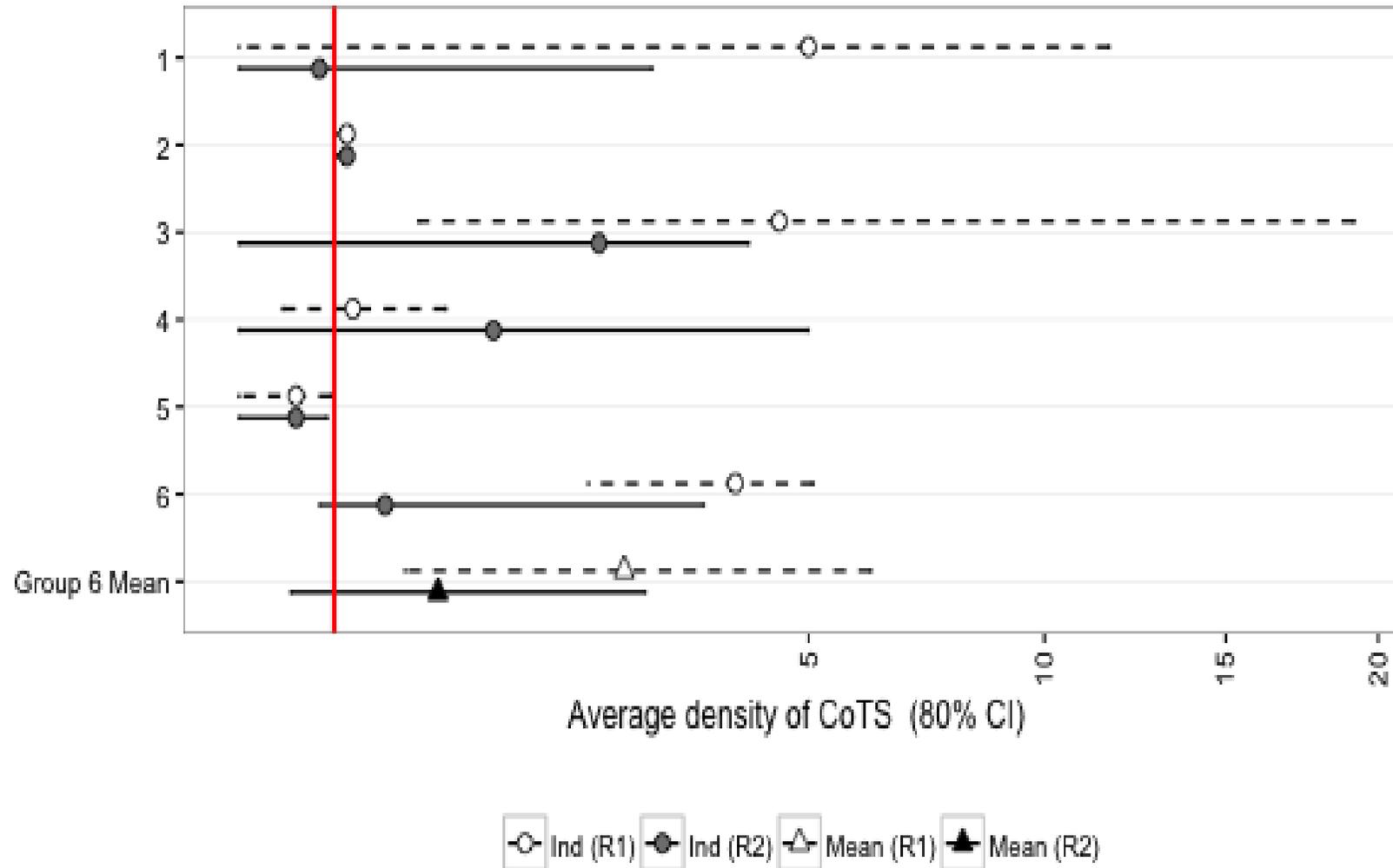
Abiotic: River discharge, El-Nino, Wind speed, Turbidity, Water Temperature, Chlorophyll, Air Temperature



Geopolitical: Zika Virus, Twitter Price, Gold, Space Launches, Refugees, Coal, Brexit.



IDEA protocol: Crown of Thorns



Performance-based questions: A case study in conservation

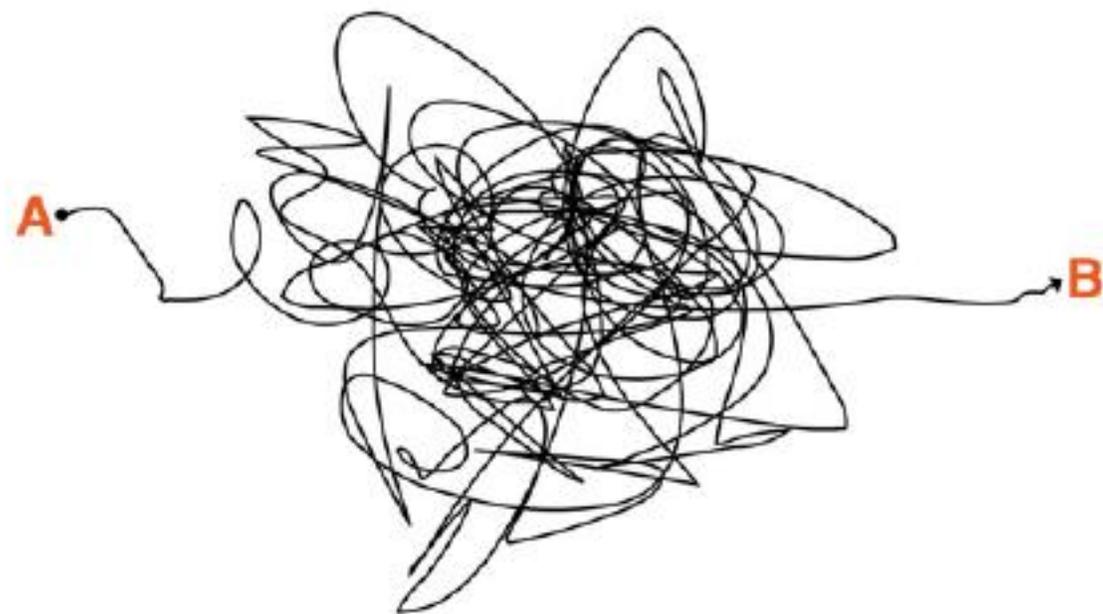


Outline

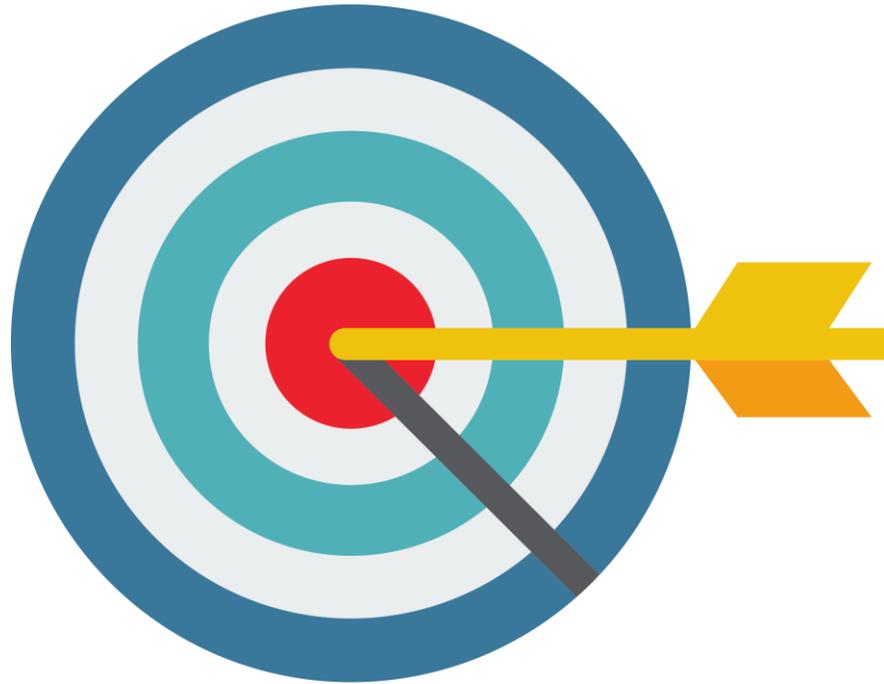
1. Should we aggregate distributions or quantiles?
2. Can **performance-weights** improve the group aggregation?
3. What makes a **good / bad question**?

Some background

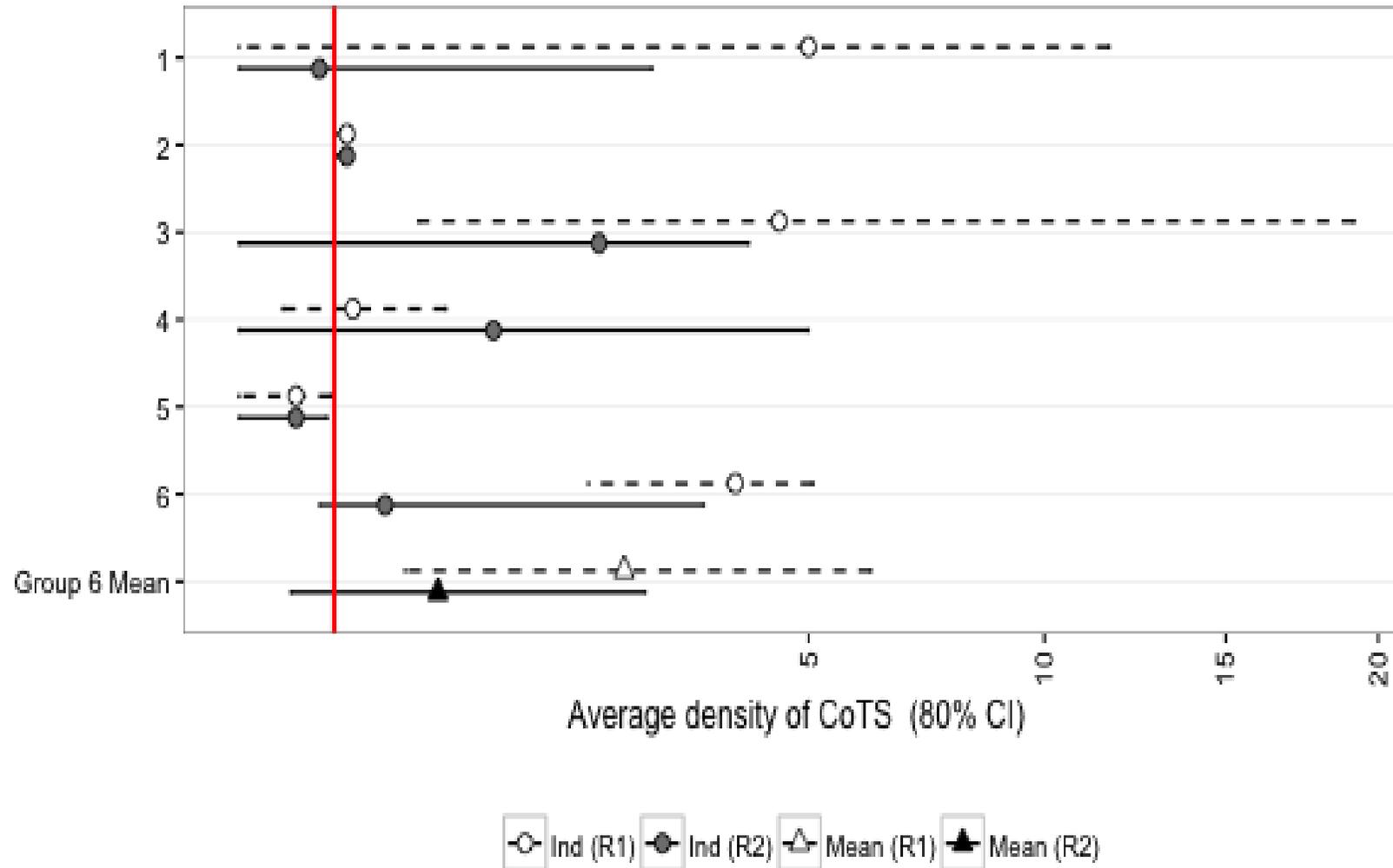
The problem?



One truth. No uncertainty.



IDEA protocol: Crown of Thorns



Experimental Design

- 21 questions
- 76 participants (8 random groups)
- Numerical estimates
- Future events validated with data

Is There a Another Option?



© CartoonStock

How Bad Can It Be?

- 7 Biotic Questions
- 7 Abiotic Questions
- 7 Geopolitical Questions

IDEA protocol: Crown of Thorns

