



Universiteit Utrecht

A GENTLE INTRODUCTION TO BAYESIAN STATISTICS

RENS VAN DE SCHOOT

SARA VAN ERP

DUCO VEEN

BETH GRANDFIELD



Goal of today presentation

- Get a flavor of what you can do with expert elicitation
- What are some considerations you should think about
- Provide some case studies to give insight
- By no means, today offers an exhaustive overview of all methods



How can we use prior knowledge?

- Bayesian statistics
 - **Prior information**
- A priori 'degree of belief' – elicited from expert
 - Represented in probability distribution
 - Variance of distribution represents (un)certainty
- A priori "degree of belief" – Based on previous studies
 - Earlier research – is it comparable?



Expert elicitation - What is it?



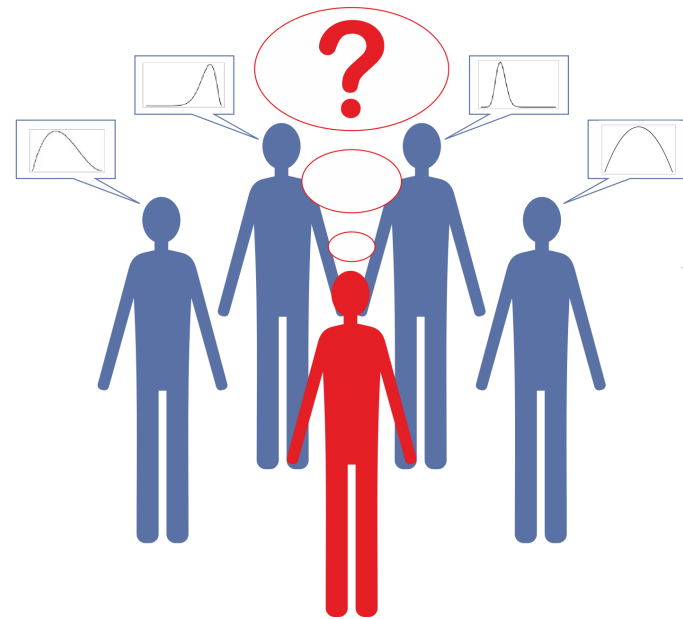
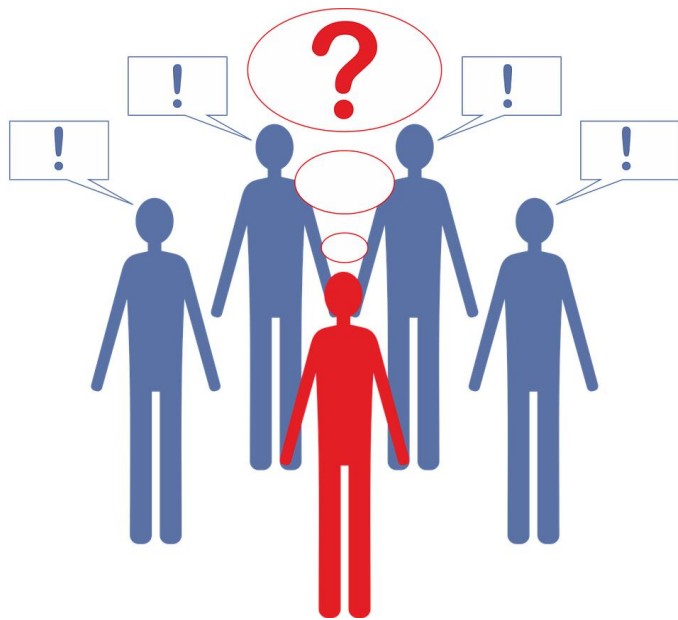
Universiteit Utrecht



“The process of creating a probabilistic representation of an experts’ beliefs is called elicitation”

O’Hagan et al., 2006







Expert elicitation – Why?

"The knowledge held by expert practitioners is too valuable to be ignored."

(Drescher et al., 2013, p. 1)





Reasons for elicitation of expert judgement

- Experts offer unique information
- It can be used to solve problems
 - As additional data to enrich the information available
 - As only data, if no data is available
- It can serve as quality control
 - Compare experts' beliefs and other data



Is expert elicitation common?

- 67,000 experts' subjective probability distributions (Cooke & Goossens, 2008)
- 57% of health economic decision models included at least one expert-knowledge elicitation parameter (Hadorn et al., 2014)
- O'Hagan et al. (chapter 10, 2006) describe examples in Medicine, Nuclear industry, Veterinary science, Agriculture, Meteorology, Business studies, Economics and Finance



Is expert elicitation common?

- ... the probability distributions (Cooke & Goossens, 2008)
- 57% of ... included at least one expert-knowledge elicitation
- O'Hagan et al. (chapter 10, 2006) described expert elicitation in the pharmaceutical industry, Veterinary science, Agriculture, Meteorology, and Economics and Finance

**Not in Psychology
(van de Schoot et al., 2016)**



Expert elicitation – What to do?

- Specific or non-specific methods
 - Suitable in general or for your problem / prior specifically?

How many parameters

- If more than one, univariate or multivariate solution?
- Direct vs. Indirect
 - Quantile elicitation
 - Predicting data



Expert elicitation – What to do?

- Group vs. Individual
 - Aggregation of priors? If so, how?
- Do they get feedback?
 - What did others say?
 - Can they adjust in multiple rounds?
- How much training is there?
 - Are your experts also statistical experts?



General reflections

- calibration questions are needed
- How much training do you experts need?
- How familiar are they with statistics?
 - Which elicitation method will suit them then?
- What is the goal of the constructed probabilistic representation?
 - Maybe suitable for some goals, not for others?



General reflections

- Do we always need a full expert prior?
 - Experts can also help to provide constraints on plausible parameter space for priors – can already be very helpful

- Do we have the same nomenclature as our experts?
 - Make sure that the systems of names and terms that are used are understood by both the statistical expert who facilitates the elicitation and the expert who have domain knowledge





Uncertain Judgements

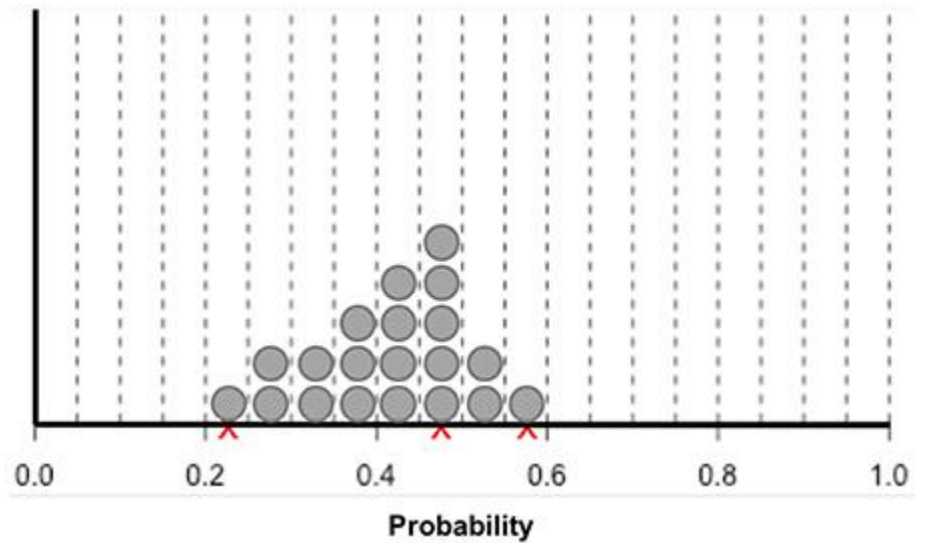
Eliciting Experts' Probabilities



ANTHONY O'HAGAN, CAITLIN E. BUCK, ALIREZA DANESHKHAH
J. RICHARD EISER, PAUL H. GARTHWAITE
DAVID J. JENKINSON, JEREMY E. OAKLEY AND TIM RAKOW

 WILEY

STATISTICS IN PRACTICE



< Articles

METHODS ARTICLE

Front. Psychol., 31 January 2017 | <https://doi.org/10.3389/fpsyg.2017.00090>



Application and Evaluation of an Expert Judgment Elicitation Procedure for Correlations

Mariëtte Zondervan-Zwijnenburg^{1*}, Wenneke van de Schoot-Hubbeek¹, Kimberley Lek¹, Herbert Hoijtink^{1,2} and Rens van de Schoot^{1,3}

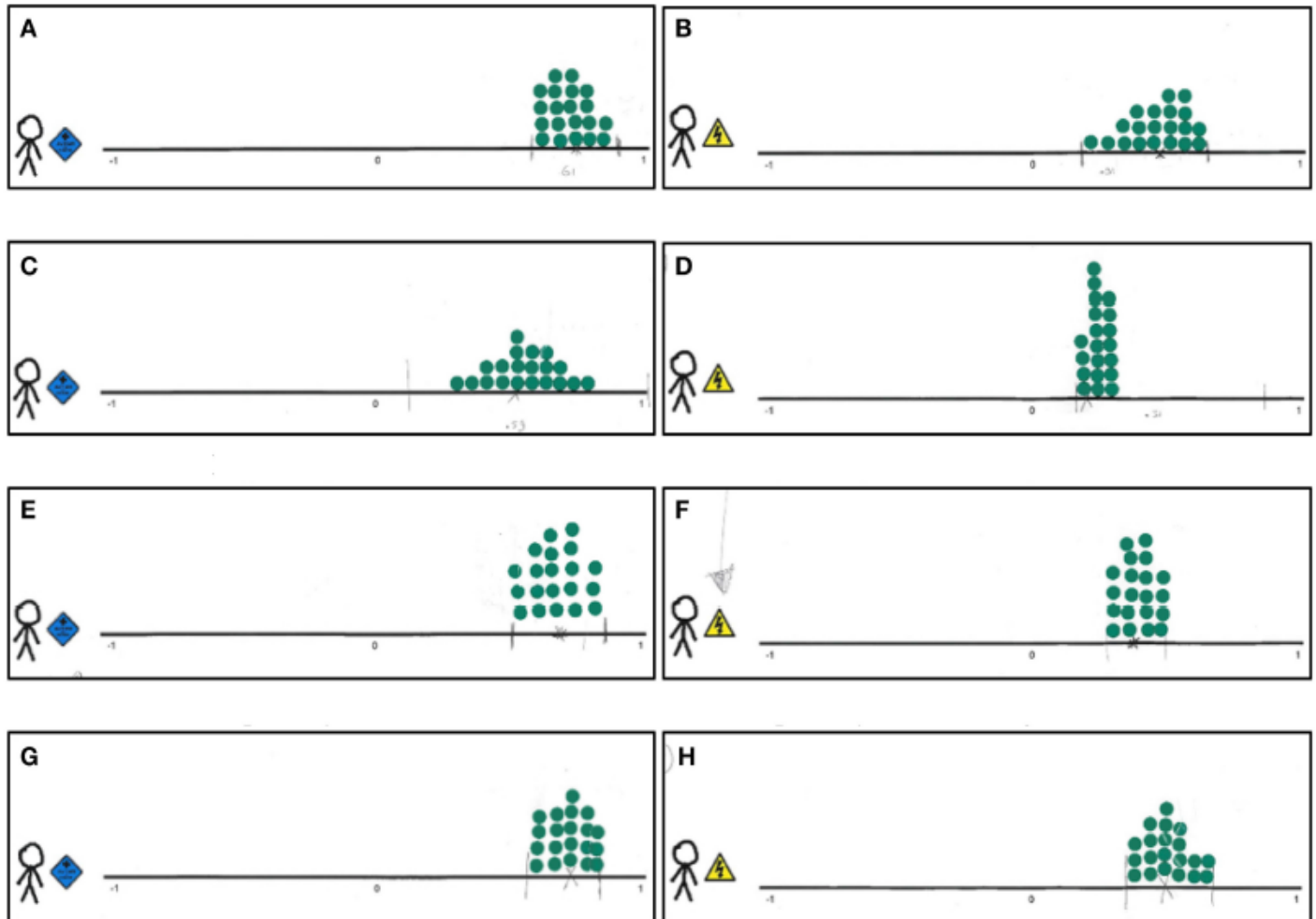
¹Department of Methods and Statistics, Utrecht University, Utrecht, Netherlands

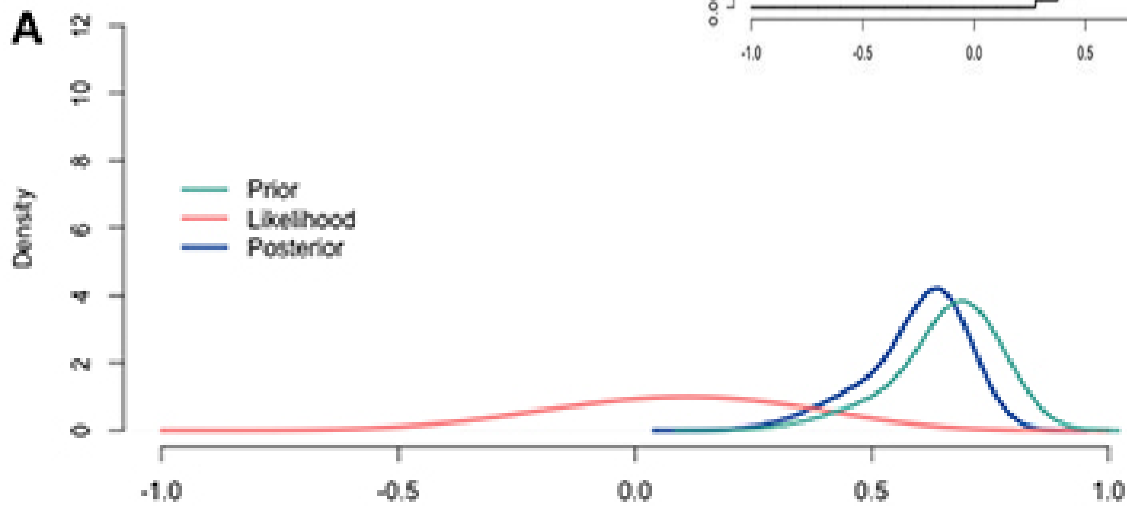
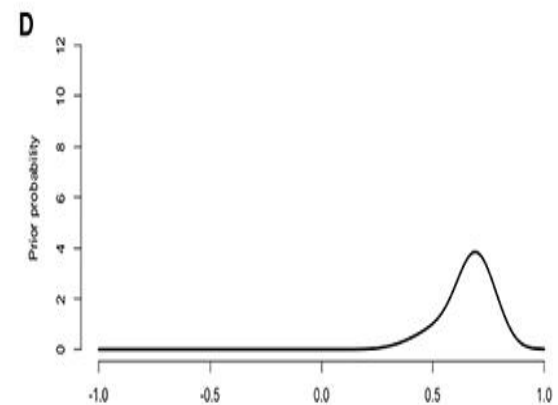
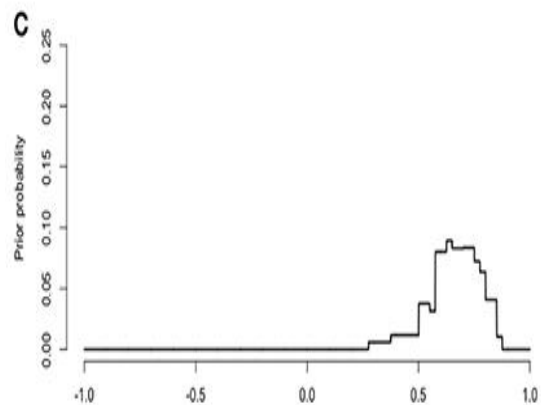
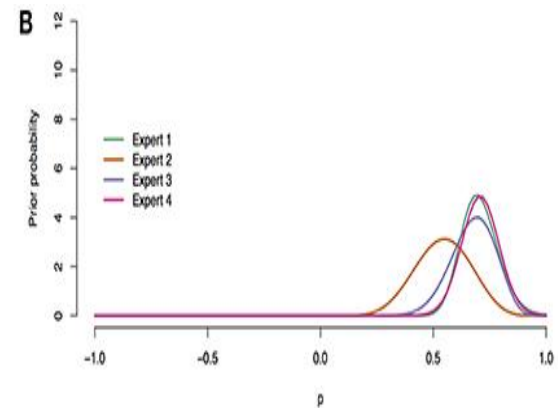
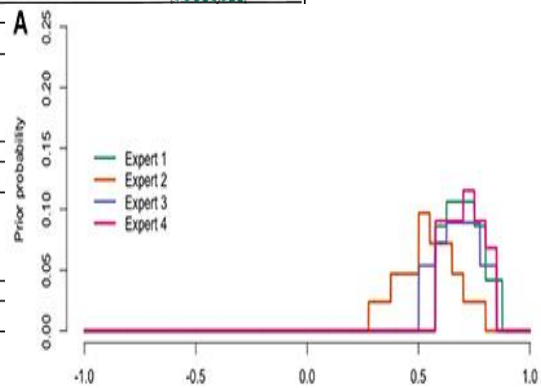
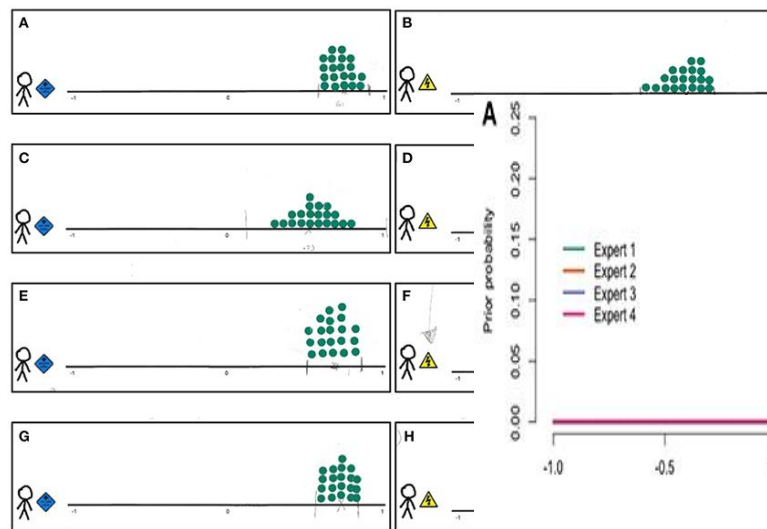
²Schreuder College Location Villerneuvestraat, Horizon Jeugdzorg en Childerwijs (Horizon Youth Care and Education), Rotterdam, Netherlands

³CITO Institute for Educational Measurement, Arnhem, Netherlands

*Openlita Research Focus Area, Nort

The purpose of the current study is to update this information with an element a trial roulette quest a concordance probability of elicitation procedure in terms means that the elicited distri





Improving elicitation quality

- Providing Feedback
 - Intuition laypeople improved through graphical elicitation techniques (Goldstein & Rothschild, 2014)
 - Interpretation expert's beliefs
 - Explicit dialogue

- Can be incorporated through software
 - Recommendation in O'Hagan et al. (2006)





Improving elicitation quality

- Avoid triggering of heuristics and biases
 - For a great overview see O'Hagan et al. (2006).
- Employ face-to-face elicitation
 - Clarifications can be given
- Training experts and facilitators
 - Make sure all expert undergo the same procedure and get the same answers to potential questions they have





Expert elicitation – Five-step method

- 1) Elicit location parameter
- 2) Fit distribution and Provide feedback
- 3) Elicit scale and shape parameters
- 4) Provide feedback
- 5) Use elicited distribution



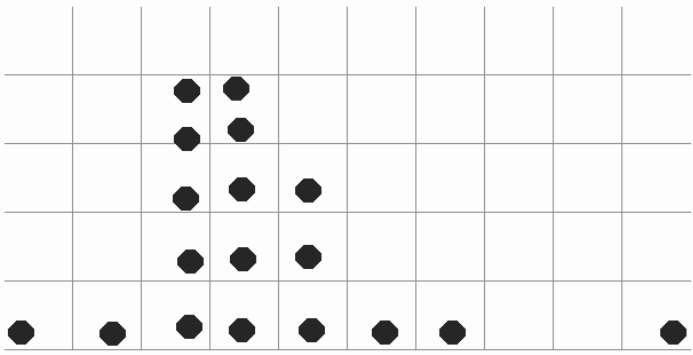
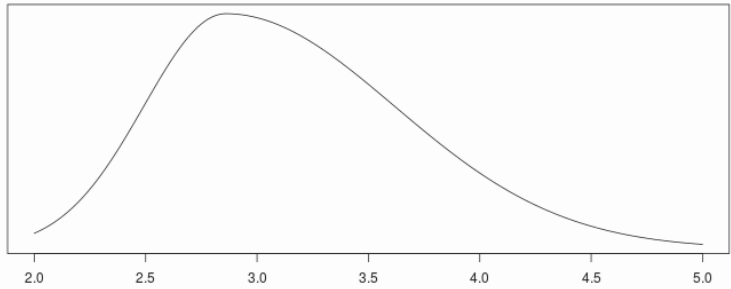
ID

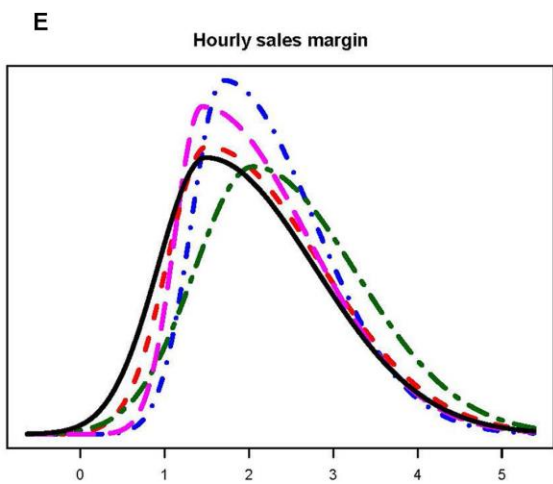
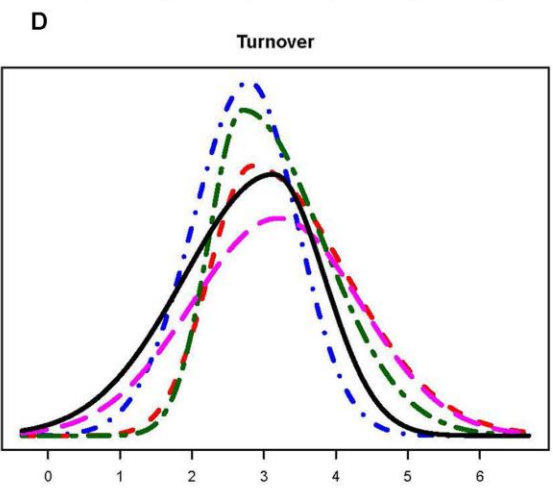
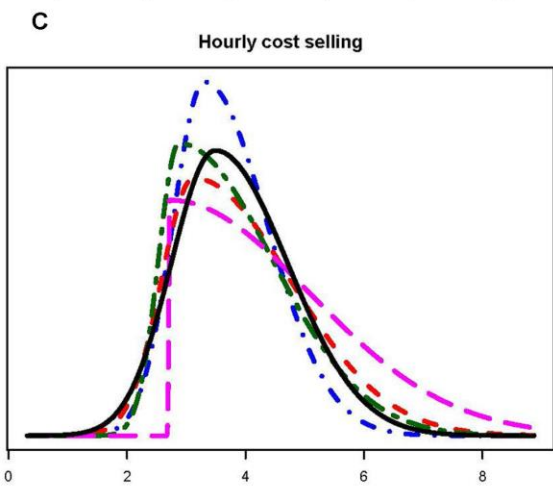
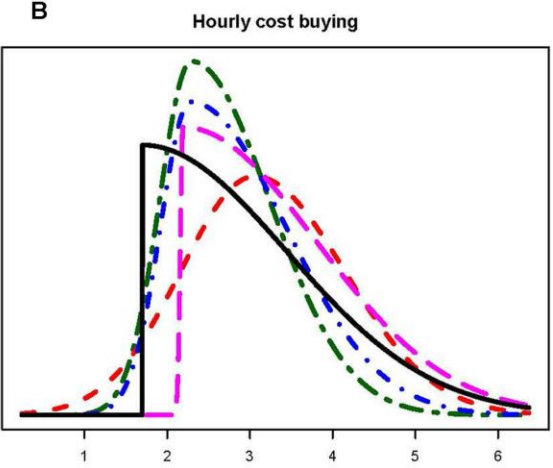
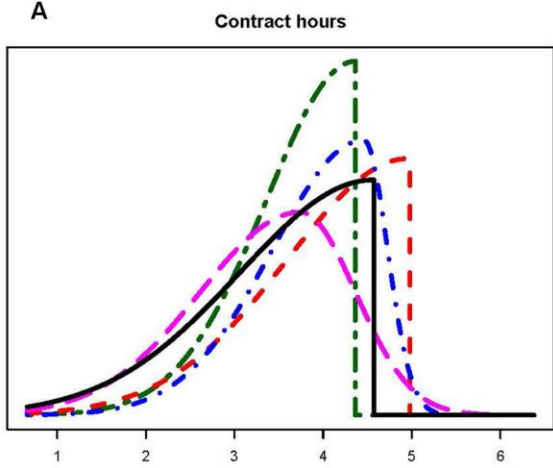
Sales results

Number of sales

Minimum sales value

Maximum sales value





-  Budget
-  Expert 1
-  Expert 2
-  Expert 3
-  Expert 4



Priors based on previous studies





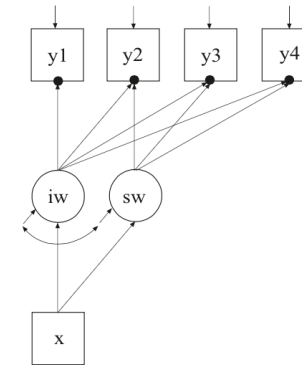
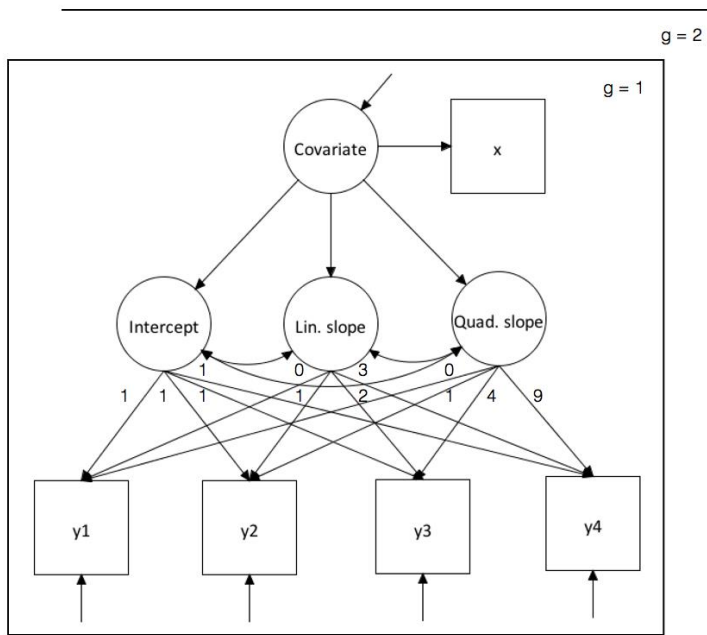
Systematically gathering information

- Search for empirical studies & Reviews
- Rate relevance of study sample for population of interest
- Example case
 - How does working memory develop in young heavy cannabis users compared to non-using peers?

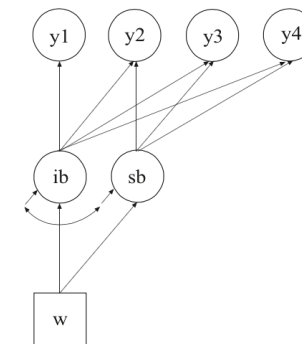




Model



Within



Between

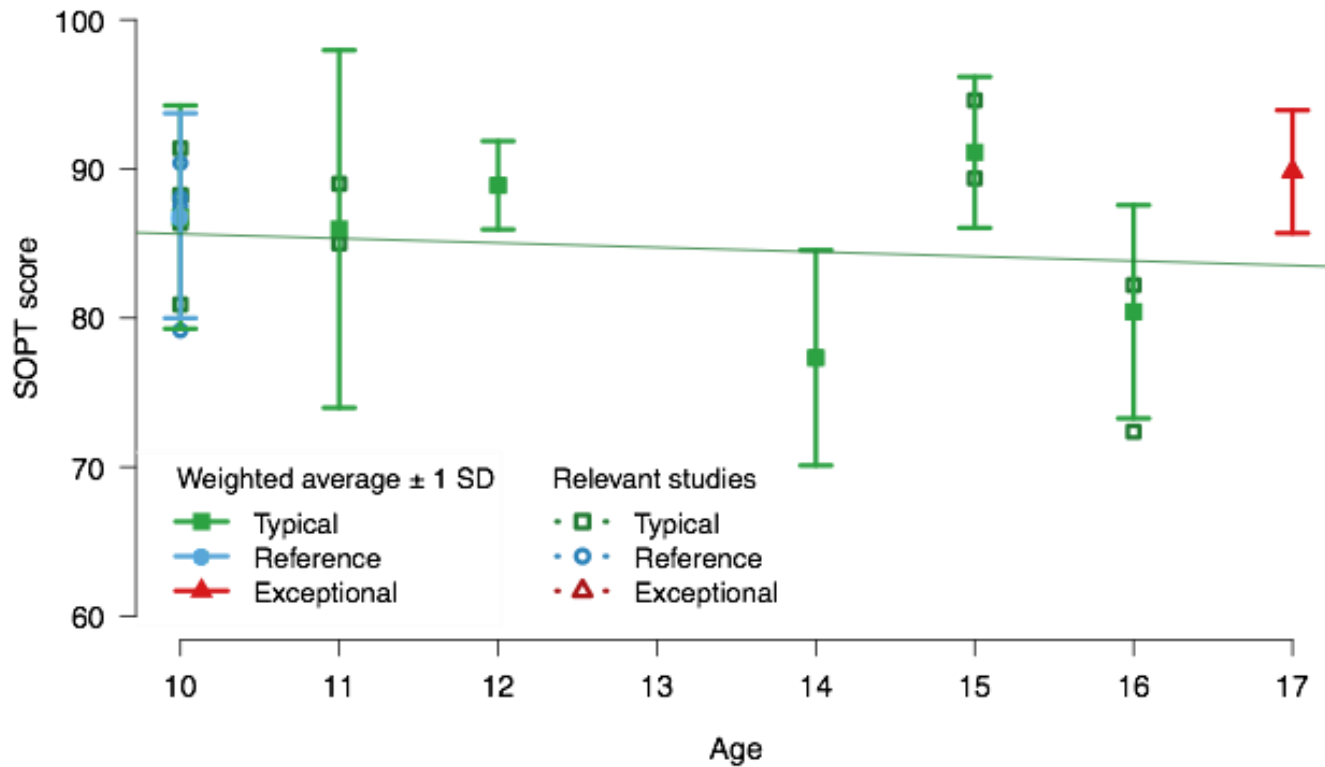


Beschrijving onderzoeksgroep	Representativiteit voor cluster 4 populatie (0-1)	Leeftijd	Verwacht % cannabis gebruikers	noot
US Children with ADHD (DSM-IV) and a mother with elevated depression level (ADHD and elevated mother depression are risk factors for CD)	.85	9.6	< 5%	
British typically developing children (no special educational needs) attending state primary schools	.01	10.1	< 1%	
Canadian adolescents that accepted a lab invitation	.01	12.5	< 1%	degenen die blowen komen vaak weg 3
Pupils from one Dutch high school, 80% boys	.1	15.6	< 12.5%	Umbo → 20% 200 → 5%
Canadian volunteers from local schools, 95% middle class families	.01	15.5	< 3%	
African-American children, 1.2% had a history of learning difficulties	.05	11.1	< 2%	
Canadian typically developing children without ADHD, mean IQ = 96.88	.01	10.3	< 1%	
Canadian children with ADHD referred to the Hyperactivity Project at the Montreal Children's Hospital for attentional and impulsivity problems. mean IQ = 96.42	.6	10.3	< 5%	
Dutch at-risk adolescents from four low-level vocational schools, 58% males	.3	16.3	50%	
Dutch children (87.8% boys) with ODD recruited from a specialised clinic for the treatment of ODD. ODD diagnosis was based on extensive psychiatric assessment and interviews with the parents. Estimated mean IQ = 99.4	.95	10.1	10%	
Dutch children (87.8% boys) with ODD in combination with ADHD recruited from a specialised clinic for the treatment of ODD. ODD diagnosis was based on extensive psychiatric assessment and interviews with the parents. Estimated mean IQ = 94.4	1	9.5	20%	
Australian adolescents from upper-working or middle-class families, recruited through the community and a Lutheran secondary school. Participants were competent English language speakers, and readers mean IQ = 112	.01	14.6	< 2%	

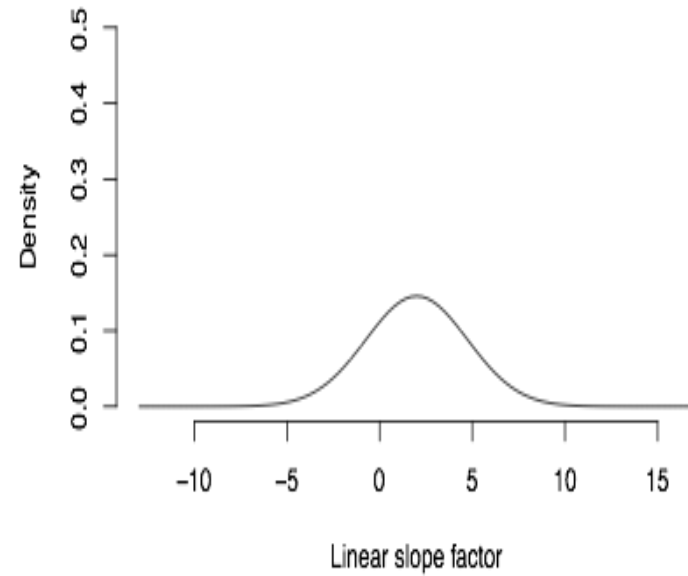
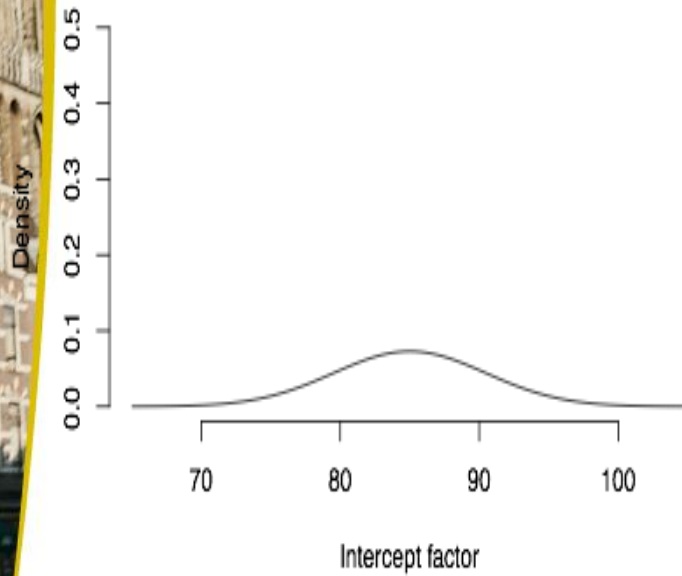
Result:

- 4 study samples relevant for non-users
- 1 relevant for heavy users
- 8 remaining (typically developing)








Weighted by relevance *
sample size for each group



Prior information



Systematically Defined Informative Priors in Bayesian Estimation: An Empirical Application on the Transmission of Internalizing Symptoms Through Mother-Adolescent Interaction Behavior

 Susanne Schulz^{1*},  Mariëlle Zondervan-Zwijenburg²,  Stefanie A. Nelemans¹,  Duco Veen^{3,4},  Albertine J. Oldehinkel⁵,  Susan Branje¹ and  Wim Meeus¹

¹ Youth and Family, Utrecht University, Utrecht, Netherlands

² Methodology and Statistics, Utrecht University, Utrecht, Netherlands

³ Julius Global Health, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, Netherlands

⁴ Optentia Research Program, North-West University, Potchefstroom, South Africa

⁵ Interdisciplinary Center Psychopathology and Emotion Regulation, University of Groningen, University Medical Center Groningen, Groningen, Netherlands

Background: Bayesian estimation with informative priors permits updating previous findings with new data, thus generating cumulative knowledge. To reduce subjectivity in the process, the present study emphasizes how to systematically weigh and specify informative priors and highlights the use of different aggregation methods using an empirical example that examined whether observed mother-adolescent positive and negative interaction behavior mediate the associations between maternal and adolescent internalizing symptoms across early to mid-adolescence in a 3-year longitudinal multi-method design.

Methods: The sample consisted of 102 mother-adolescent dyads (39.2% girls, M_{age} T1 = 13.0). Mothers and

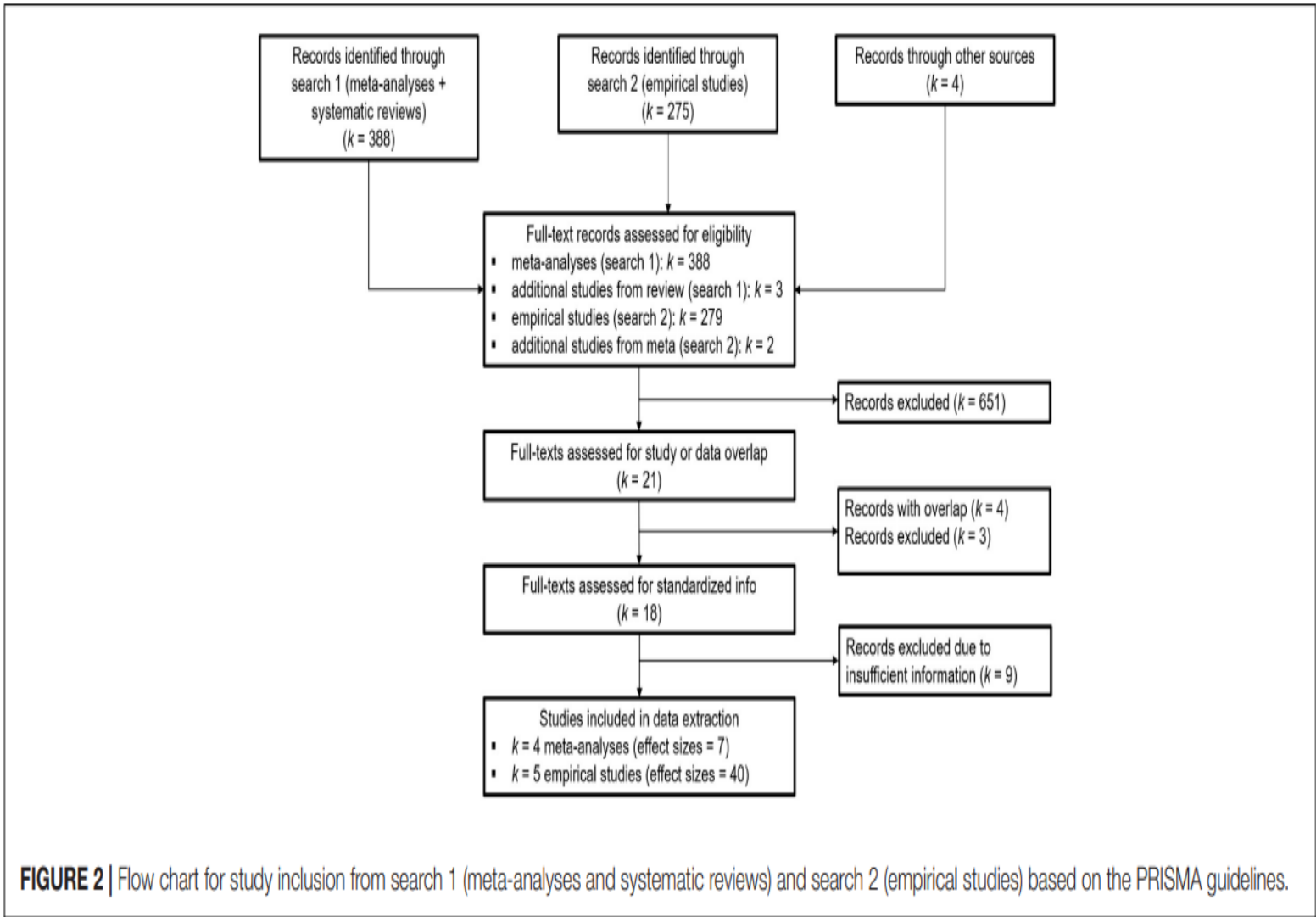


TABLE 1A | Weighting scheme for informative priors.

Category	Points	Details
T1-T2 (longitudinal)	10	The estimates of longitudinal studies are usually smaller than those of cross-sectional studies. As our parameter are longitudinal estimates as well, longitudinal designs should receive most weight in relation other categories.
- controlling for symptoms at T1	20	Longitudinal studies that do not control for symptoms at T1 might have quite large estimates and cannot indicate change. As this is the most crucial aspect of longitudinal research, studies that also control for T1 symptoms should receive more weight. <i>Not applicable for T1 → T2 associations (deleted from final score)!</i>
- Same time lag - (1 year)	5	Studies that use the same time lag as we do are closer to our study design and thus deserve more weight.
Observation	15	The study list only includes empirical studies with observational assessments of the parent-adolescent interaction as these (multi-method) estimates are usually smaller than self-reports. However, meta-analyses often include a combination of observations and self-reports, which is difficult to disentangle. Therefore, estimates from "pure" observations should receive more weight than mixed studies (and most weight in relation to other categories as this is another main aspect of our study).
Early adolescence (12–16)	10	Some studies, and particularly the meta-analyses, used a broader age range than our study or even just adolescence (but all studies include adolescence). As our study focuses on early-mid adolescence, studies that included a similar age group should receive some more weight.
Internalizing symptoms include both anxiety and depression, or anxiety only	10	Most studies do not focus on a combination of depression and anxiety symptoms, but only include one of those symptoms (mostly depression). As we will use a combination of both, studies that include measures on internalizing symptoms or both depression and anxiety symptoms should receive more weight. <i>Most studies focus on mother or adolescent depression (rather than anxiety). To counterbalance that, we will also award 5 points if the study only focused on anxiety (i.e., either combined or anxiety only).</i>
Including covariates - parental symptoms	5	If studies include other relevant covariates that might better reflect our study associations, such as parental symptoms (for T2-T3 parameters), they might receive additional weight.
- other interaction behaviors	5	
Community sample (does not include clinical/diagnostic groups)	10	Many (older) studies include two subsamples, of which one is usually clinical. Therefore, the final sample includes participants who may have higher levels of internalizing symptoms than our participants. For these participants, the associations may be stronger. Thus, studies with a community sample which is closer to our sample should receive more weight.
Meta-analysis	10	Meta-analyses combine information from several studies and thus provide the most comprehensive evidence. Therefore they should receive somewhat more weight than individual studies.
10 categories (standard 5)	100 (80)	Each study can score between 0 and 100 points (or between 0 and 80 points for T1 → T2 associations).



TABLE 1B | Final scoring of all included studies.

Study	T1-T2	lag	cT1	obs	Age	M _{dep+anx (or anx)}	A _{dep+anx (or anx)}	cov _s	cov _i	comm	MA	Score
Points	10	5	20	15	10		10	5	5	10	10	100
Lovejoy et al. (2000)				x							x	25
Simons et al. (1993)*	x				x							20
McCabe (2014)				x		x					x	35
Pinquart (2017)	x		x				x			x	x	60
Weymouth et al. (2016)							x			x	x	30
Allen et al. (2006)	x	x	x	x	x					x		70
Asbrand et al. (2017)	x			x			x					35
Dadds et al. (1992)				x								15
Dietz et al. (2008)				x		x						25
Griffith et al. (2019), (neg)	x		x	x				x		x		60
Griffith et al. (2019), (pos)	x		x	x						x		55
Hofer et al. (2013)	x		x	x	x		x		x	x		80
Jackson et al. (2011)				x								15
Milan and Carbone (2018), (only cs)				x				x	x	x		30
Milan and Carbone (2018)	x		x	x				x	x	x		60
Nelson et al. (2017)	x			x					x			30
Olino et al. (2016)	x		x	x				x				50
Schwartz et al. (2012)	x		x	x	x		x	x				70
Szwedo et al. (2017)	x		x	x						x		55
van Doorn et al. (2016)				x				x	x			25

Note. T1-T2 = longitudinal assessment, lag = same time lag used (for longitudinal studies), cT1, controlling for T1 symptoms (for longitudinal studies); obs, observational assessment of parent-adolescent interaction; age, age range early adolescence; N, sample size; M, maternal; A, adolescent; year, publication year; cov_s, controlling for parental symptoms; cov_i, controlling for other interaction behaviors; comm, community sample; MA, meta-analysis; x, indicates that the category is met, gray studies were excluded from the final analyses due to insufficient standardized information.

*Study included in aforementioned meta-analysis.





TABLE 2 | Informative priors for the regression parameters in Model A and Model B.

Parameter description and names	Linear pool	Logarithmic pool	Fitted normal	Image
Maternal internalizing symptoms T1 → Maternal positive interaction T2 <i>MPonMint</i> <i>b_meanMP2[1]</i>	$N(-0.18, 0.0179)^{0.4375} +$ $N(-0.21, 0.1040)^{0.3125} +$ $N(-0.29, 0.0015)^{0.3750}$	$N(-0.29, 0.01)$	$N(-0.23, 0.20)$	
Adolescent internalizing symptoms T1 → Maternal positive interaction T2 <i>MPonAint</i> <i>b_meanMP2[2]</i>	$N(-0.06, 0.0077)^{0.5000} +$ $N(-0.09, 0.0950)^{0.3125} +$ $N(-0.12, 0.1755)^{0.1875} +$ $N(-0.16, 0.6407)^{0.3750}$	$N(-0.06, 0.03)$	$N(-0.10, 0.98)$	
Maternal internalizing symptoms T1 → Adolescent positive interaction T2 <i>APonMint</i>	$N(-0.06, 0.0704)^{0.3750}$	$N(-0.06, 0.19)$	$N(-0.06, 0.19)$	



Contrasting experts' beliefs and data



Universiteit Utrecht



Expert elicitation – Why?

“The knowledge held by expert practitioners is too valuable to be ignored.”

(Drescher et al., 2013, p. 1)





Expert elicitation – Why?

“The knowledge held by expert practitioners is too valuable to be ignored. But only when thorough methods are applied, can the application of expert knowledge be as valid as the use of empirical data. The responsibility for the effective and rigorous use of expert knowledge lies with the researchers”

(Drescher et al., 2013, p. 1)





Expert elicitation – Quality control

- Classical method
 - Calibration questions

- But what if you don't have many questions to calibrate on?
 - Maybe one of the reasons why expert elicitation is not common in psychology?



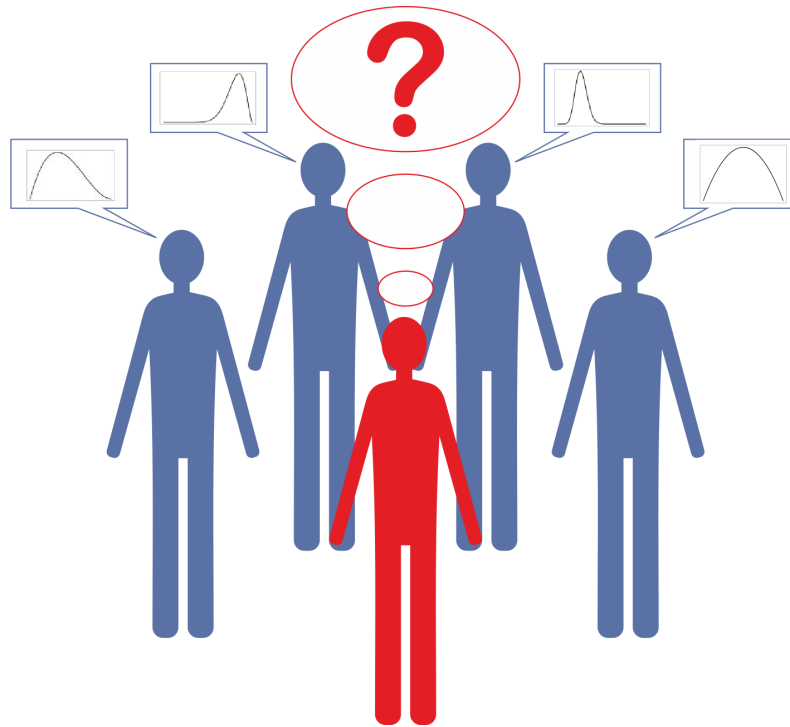


Expert elicitation – Quality control

- Direct comparison expert priors and data
 - Prior predictive distributions – save bet to be uncertain
 - Prior-data conflict measure

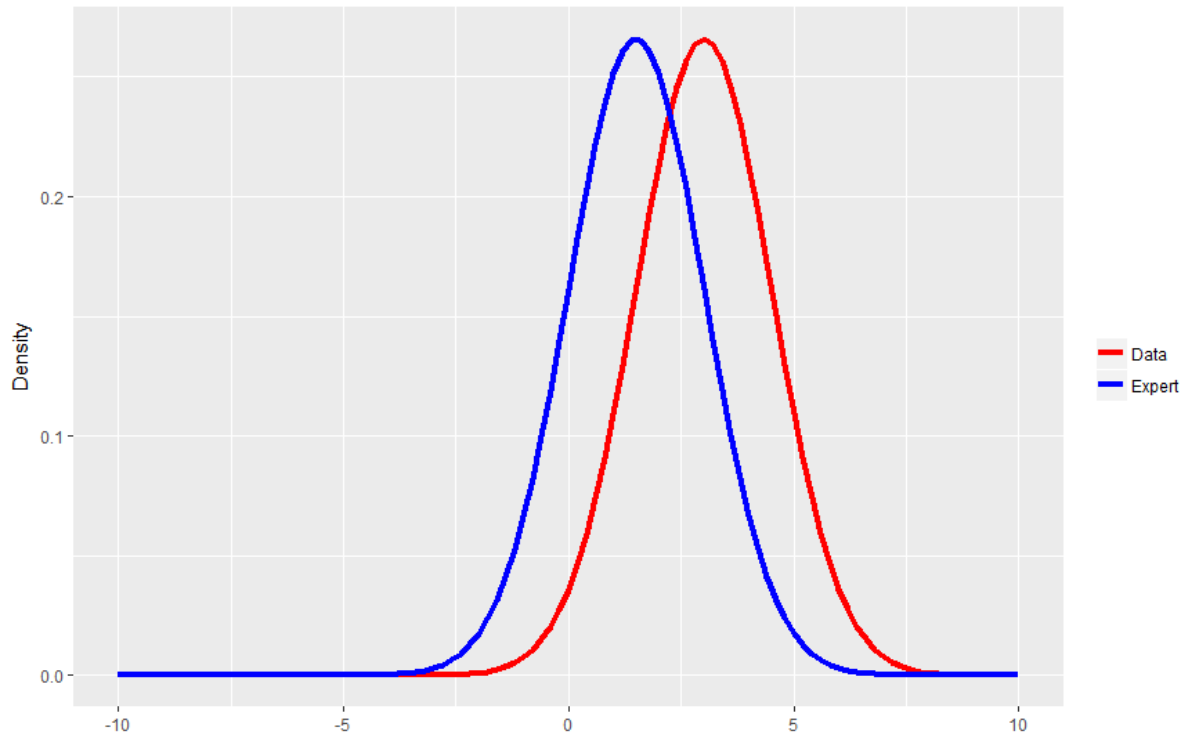


Quality Control

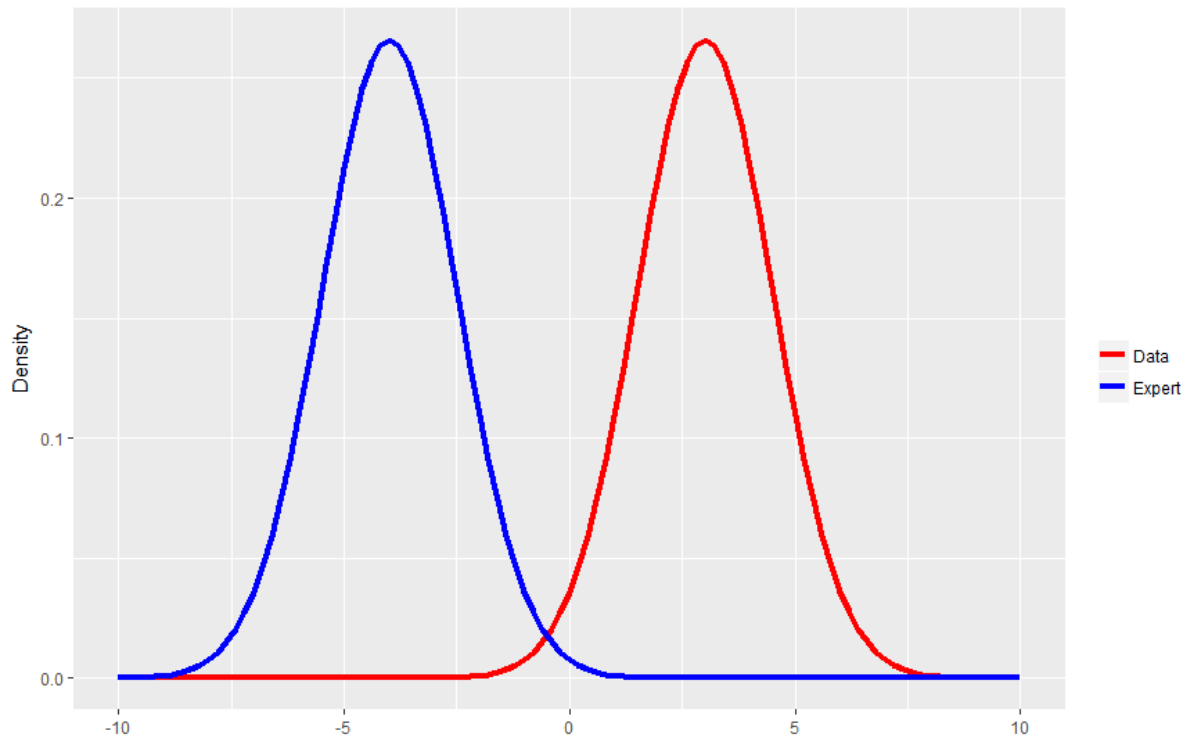




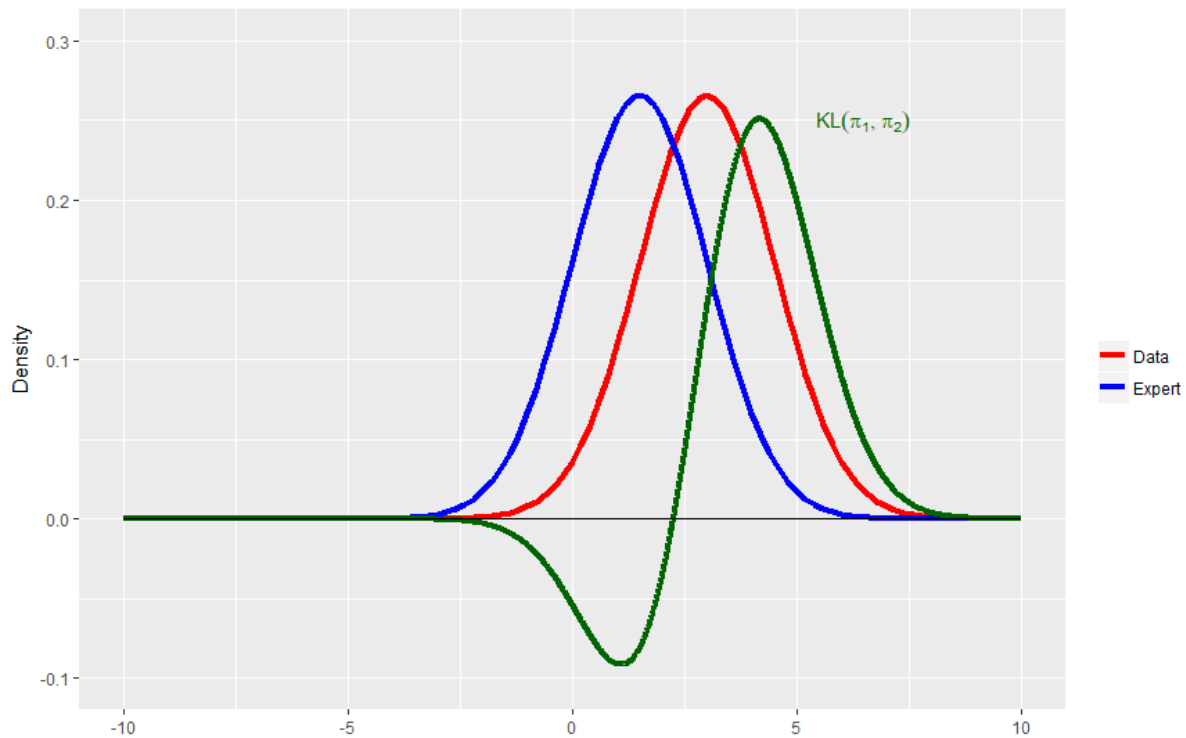
Prior-data Agreement



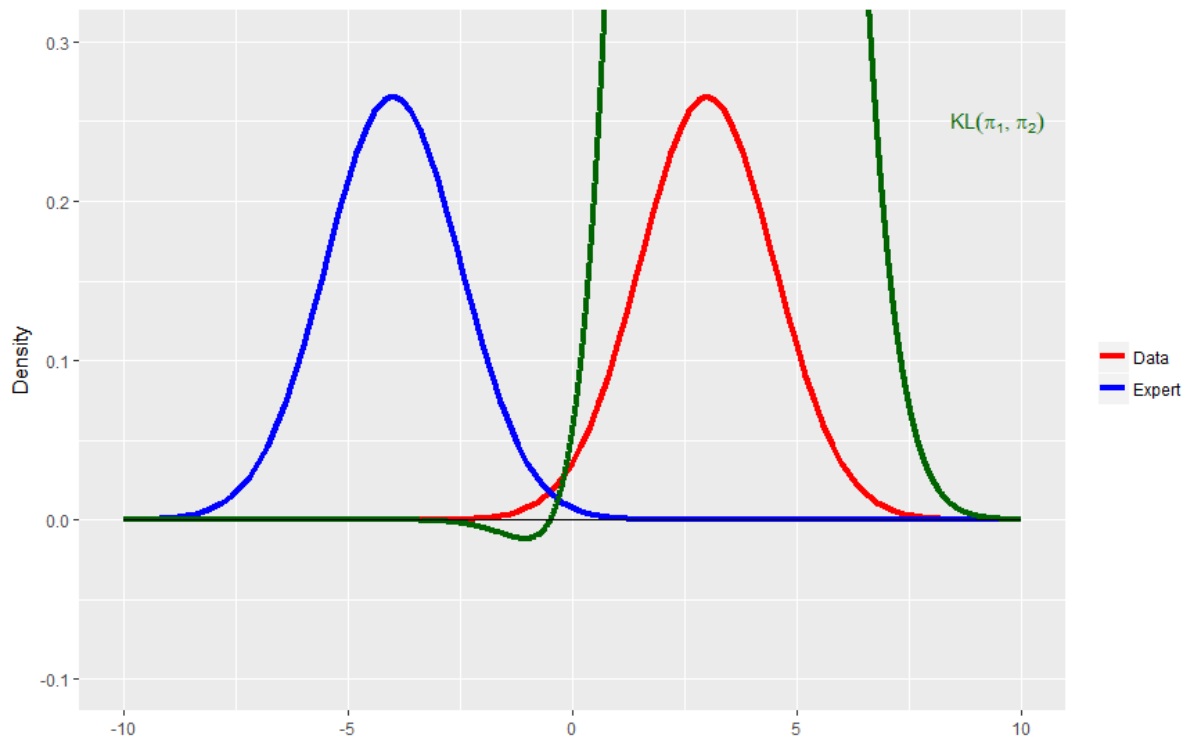
Prior-data Disagreement



Kullback-Leibler Divergence



Kullback-Leibler Divergence





Data Agreement Criterion

- Bousquet (2008)
 - Take a benchmark prior
 - Compute a posterior based on the data and the benchmark prior
 - Get KL-divergence between computed posterior and the benchmark prior
 - Get KL-divergence between computed posterior and the candidate (expert) prior
 - Compute the ratio of candidate KL / benchmark KL





Data Agreement Criterion

- Ratio smaller than 1
 - No prior-data conflict
 - The candidate prior resembles the data more closely than the benchmark prior









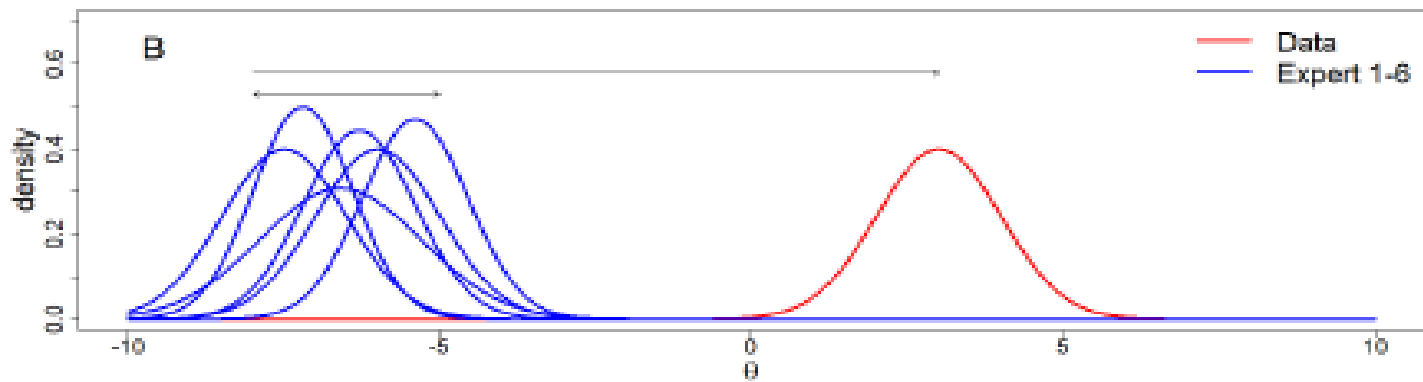
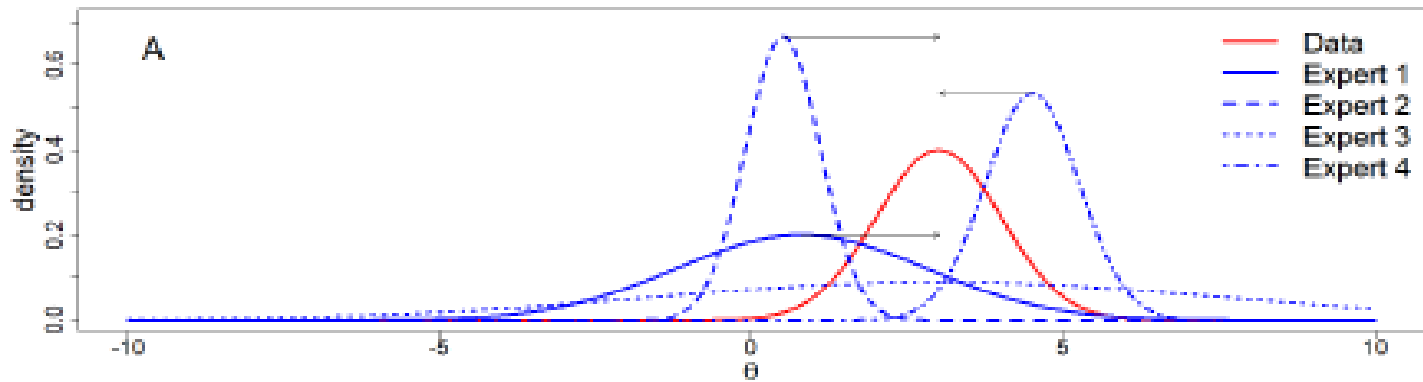
Data Agreement Criterion

- This leaves the choice for the benchmark
 - Needs to be of low information compared to the data
- When we have multiple experts
 - We can compare their KL divergences directly or all to the benchmark
 - Always look at data visually too





Data Agreement Criterion





Case studies



Universiteit Utrecht



Experts in a financial institution

- How good are the prior beliefs of experts?
- Regional directors provided their beliefs regarding average turnover per professional in the upcoming quarter
 - They are experts concerning market opportunities, market dynamics and estimating the capabilities of the professionals to seize opportunities
 - They were used to providing a single digit estimate
 - We got them to specify their beliefs in terms of priors

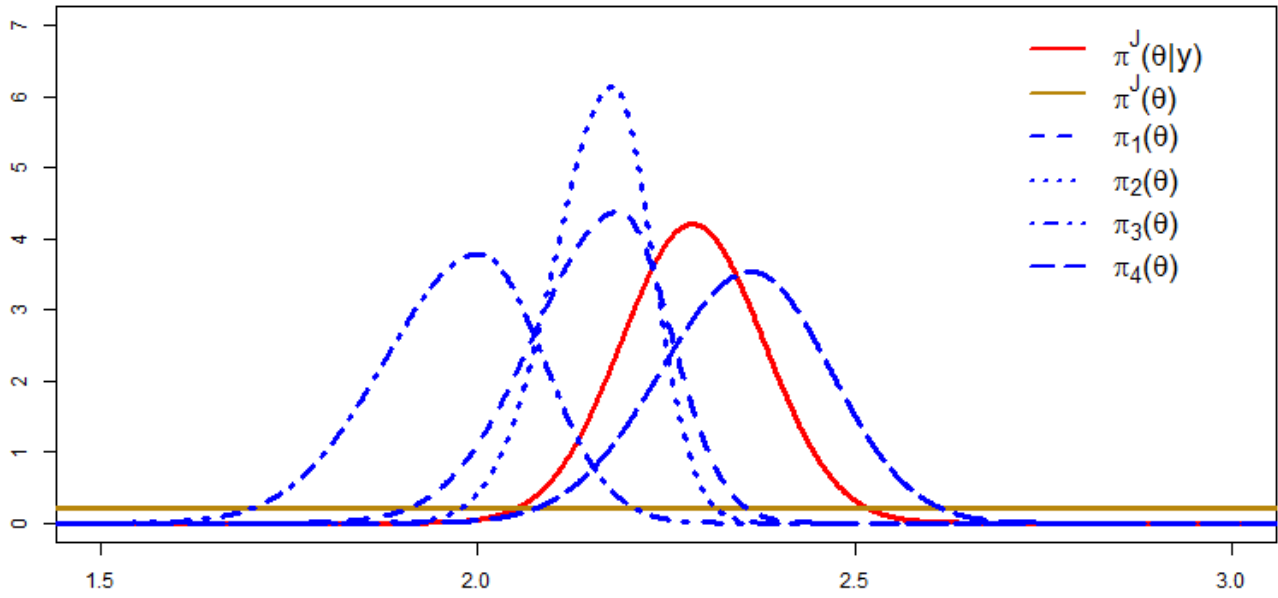




Experts in a financial institution

- We compared their prior beliefs to the actual realization of that quarter
 - Benchmark used was uniform prior ranging from 0 (no turnover) up to a large value that could not reasonably be attained.



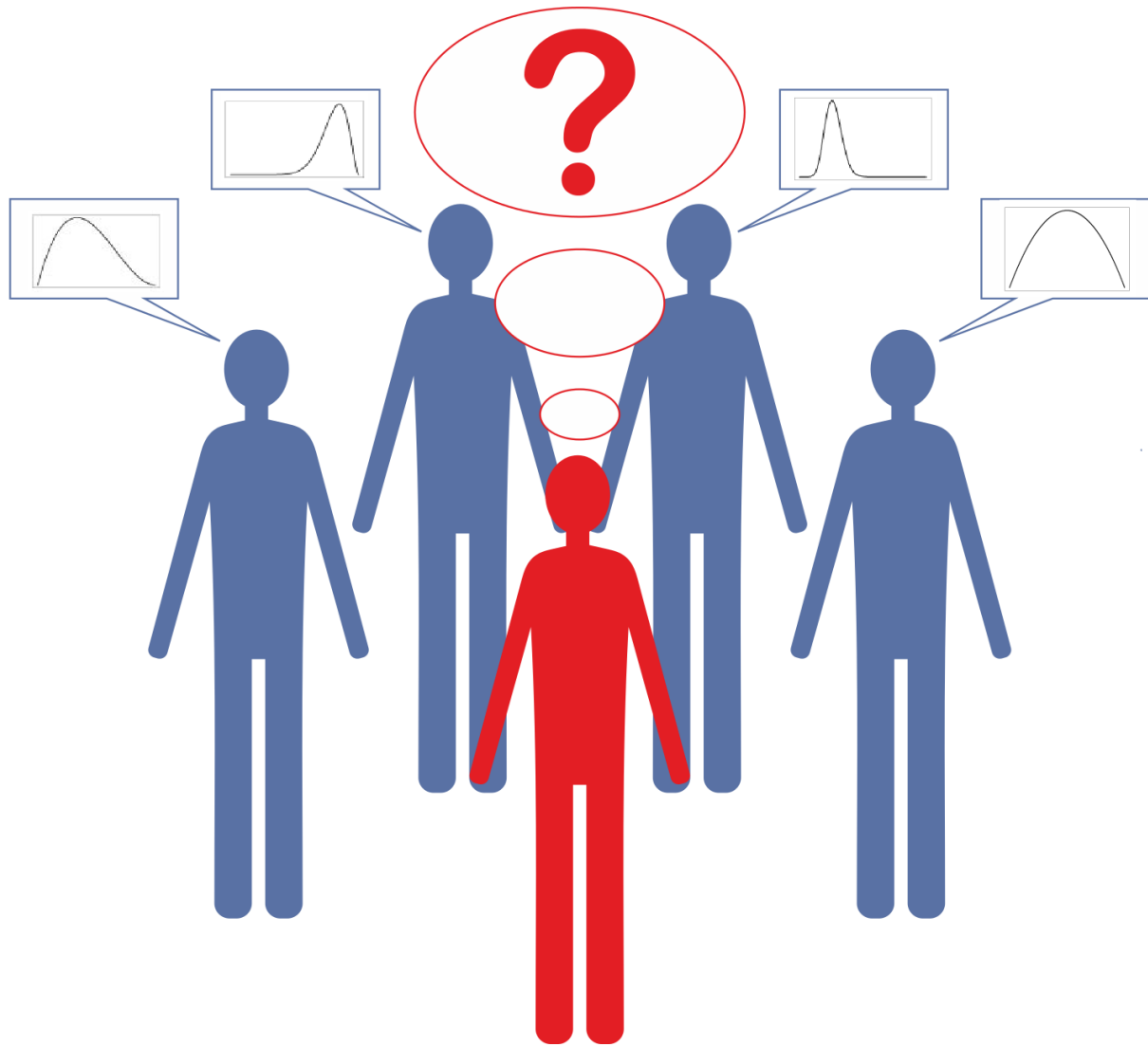


	KL divergence	DAC _d	Ranking
Expert 1	1.43	0.56	2
Expert 2	2.86	1.12	3
Expert 3	5.76	2.26	4
Expert 4	0.19	0.07	1
Benchmark	2.55	-	-

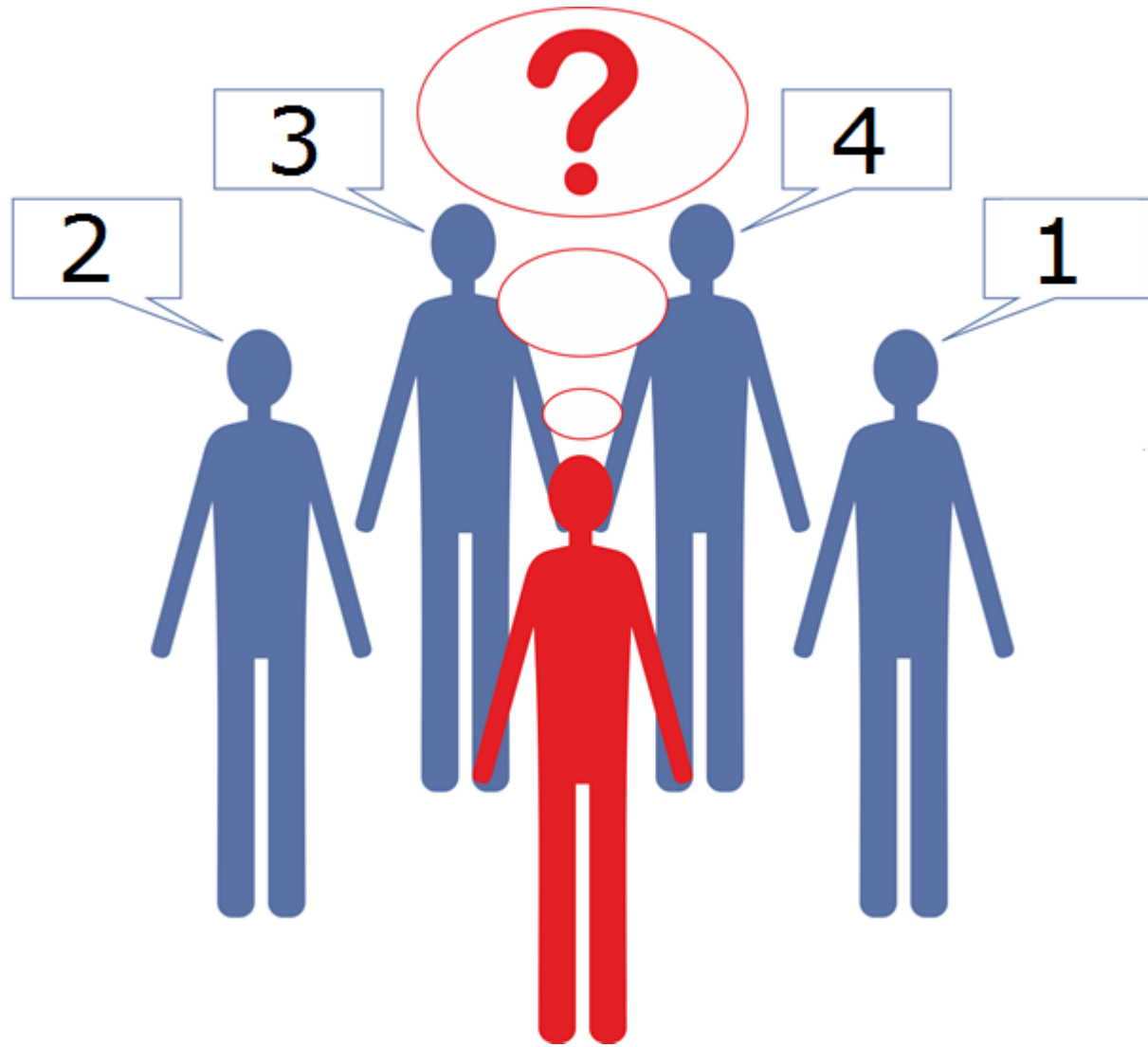


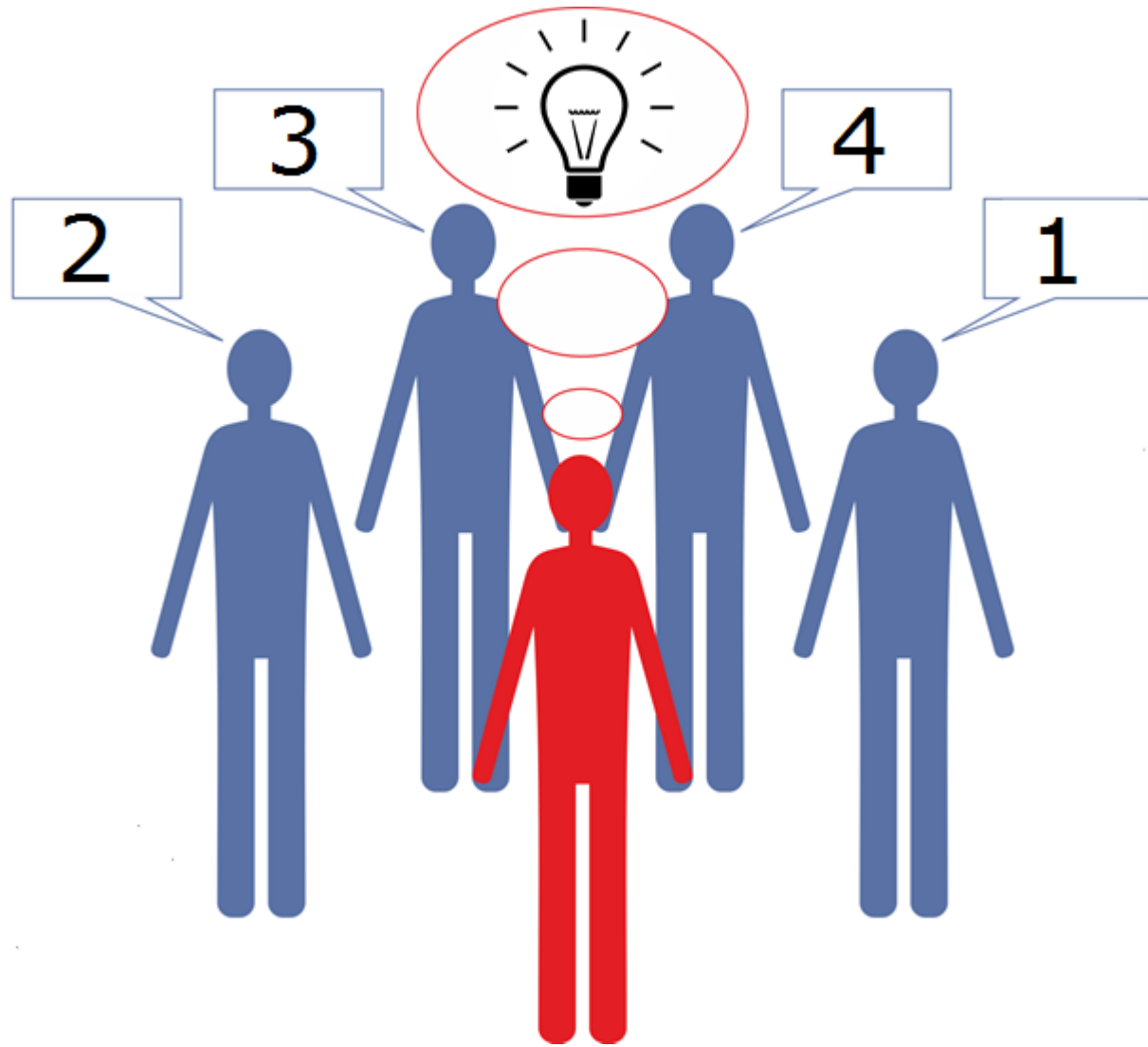


Universiteit Utrecht



Universiteit Utrecht





Impact of pediatric burn injuries



Impact of pediatric burn injuries





Impact of pediatric burn injuries

- How do Posttraumatic Stress Symptoms (PTSS) develop in children with burn injuries?
- 8–18-year old from Netherlands and Belgium
- Minimal 24-hour stay
- Minimal percentage of body burned of 1%
- Self-reported posttraumatic stress symptoms





Experts in burn-injuries and PTSS

- 7 nurses specialized at working with burn-injuries
- 7 psychologists working with the children
- From all 3 Dutch burn-institutes
- Audio recordings of elicitations for qualitative information



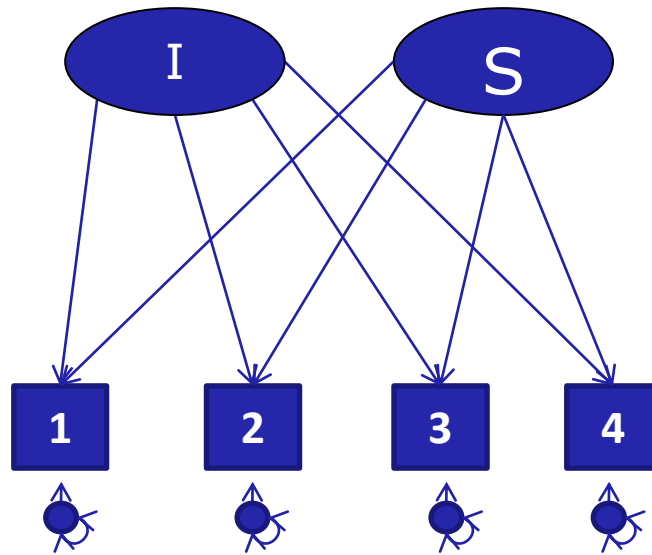


Expert elicitation

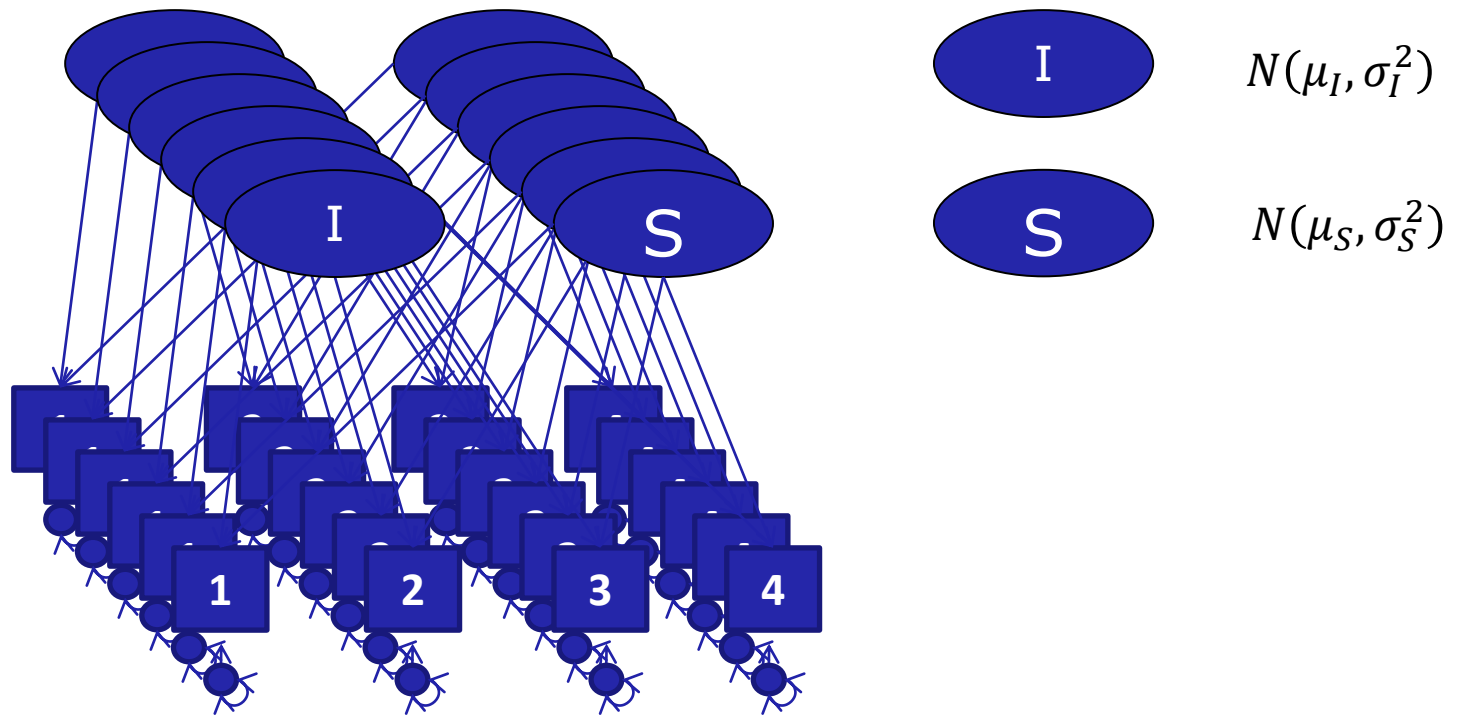
- Extending the Five-step method from before
- Adjusted the method for the elicitation of hierarchical model



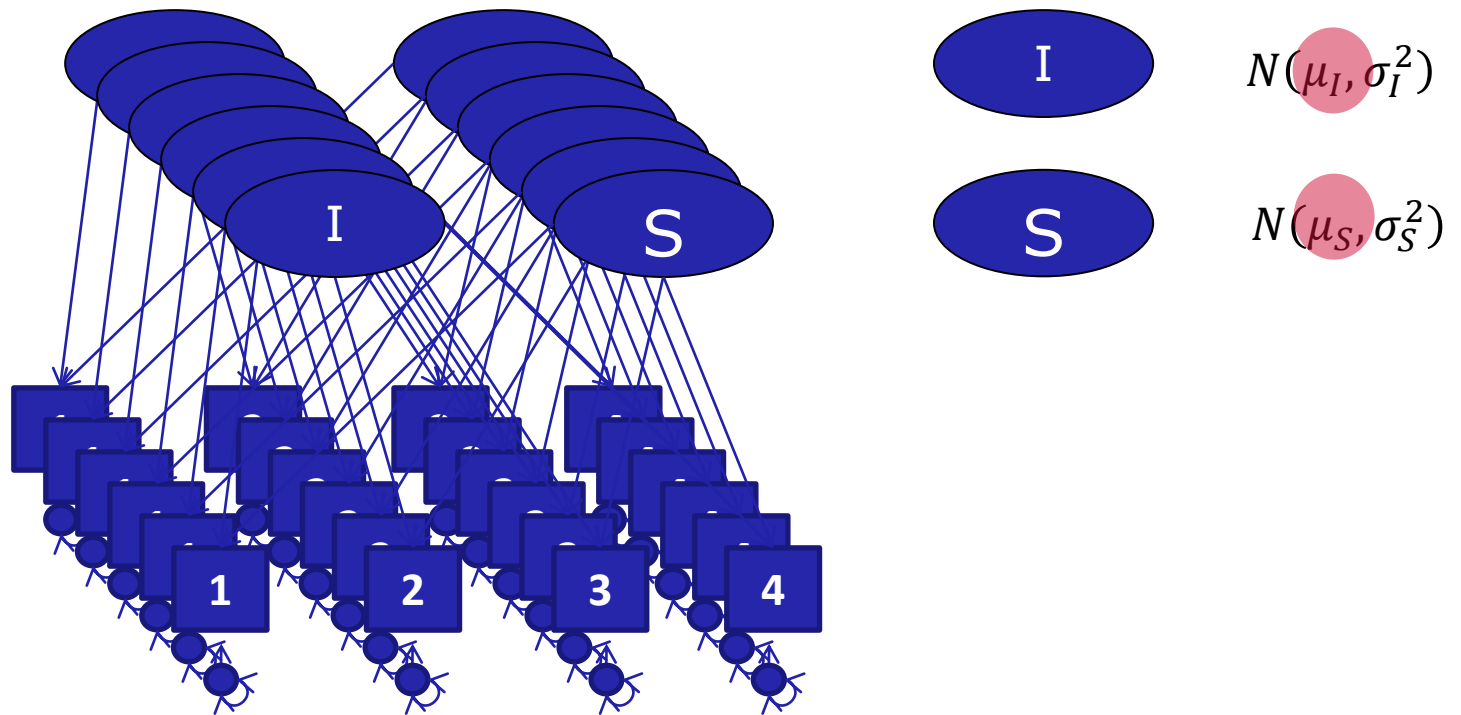
Model per child



Hierarchical model

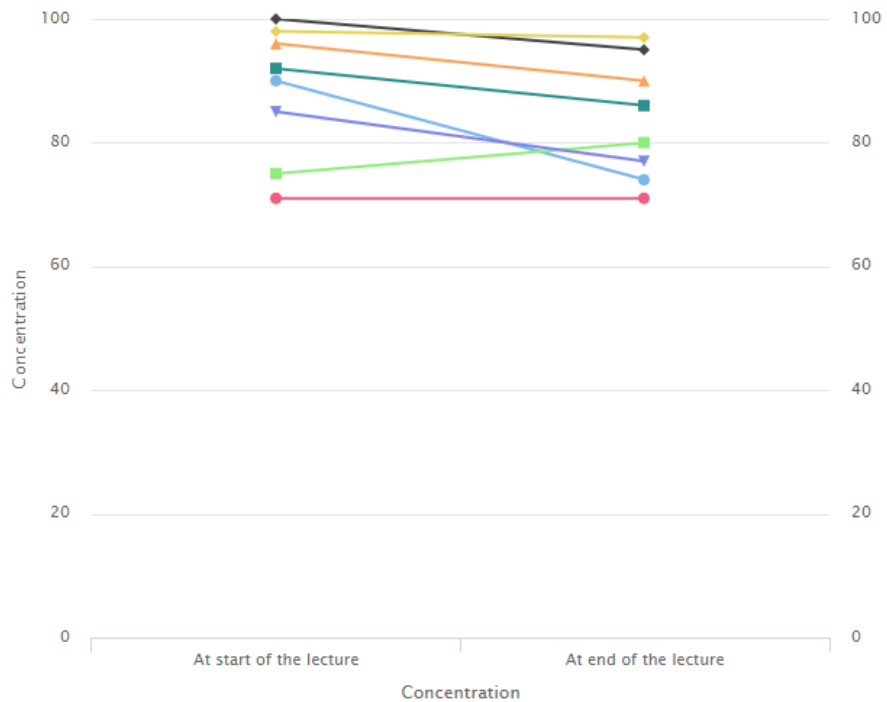


Hierarchical model





Add Additional Line
 Show average trajectory
Submit



	data
Average concentration at start of the lecture	88.38
Average change in concentration from start to end of the lecture	-4.62

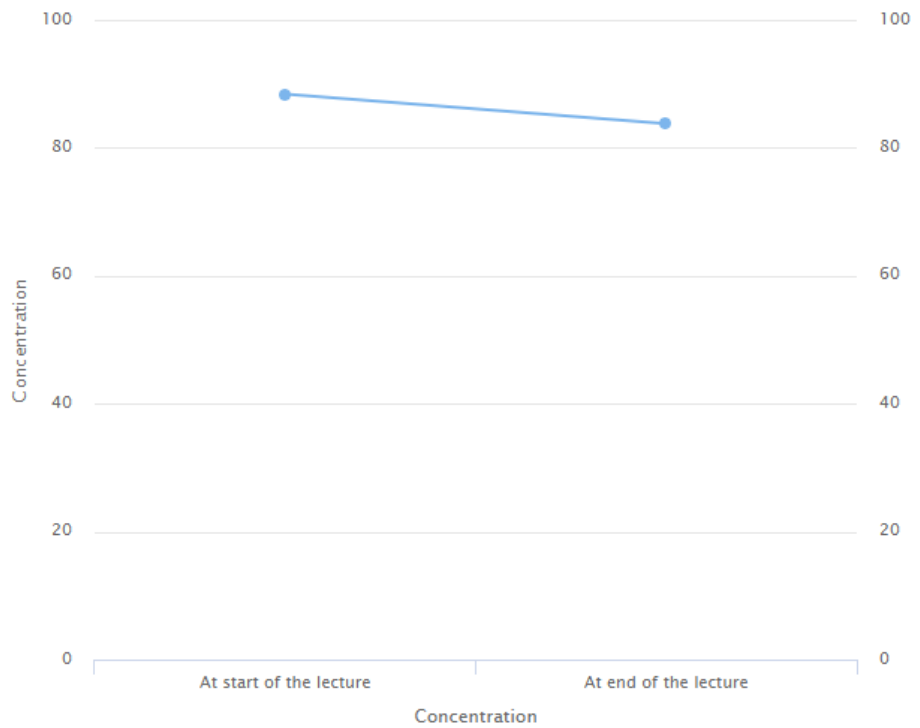




Add Additional Line

Show average trajectory

Submit



	data
Average concentration at start of the lecture	88.38
Average change in concentration from start to end of the lecture	-4.62





Reasonable lowerbound average concentration at start of lecture

95

Average average concentration at start of lecture

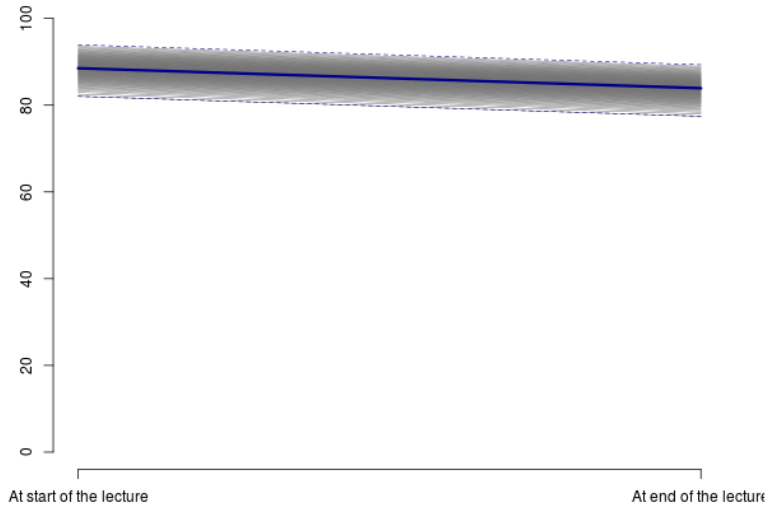
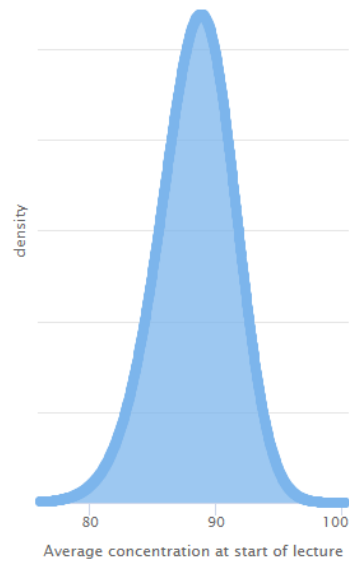
88,4

Reasonable upperbound average concentration at start of lecture

80

Fit distribution

Show implications



Concentration				
2.5%	25%	50%	75%	97.5%
82	86.5	88.6	90.5	93.9
95% CI		50% CI		
[82, 93.9]		[86.5, 90.5]		





Reasonable lowerbound average change in concentration

0

Average change in concentration

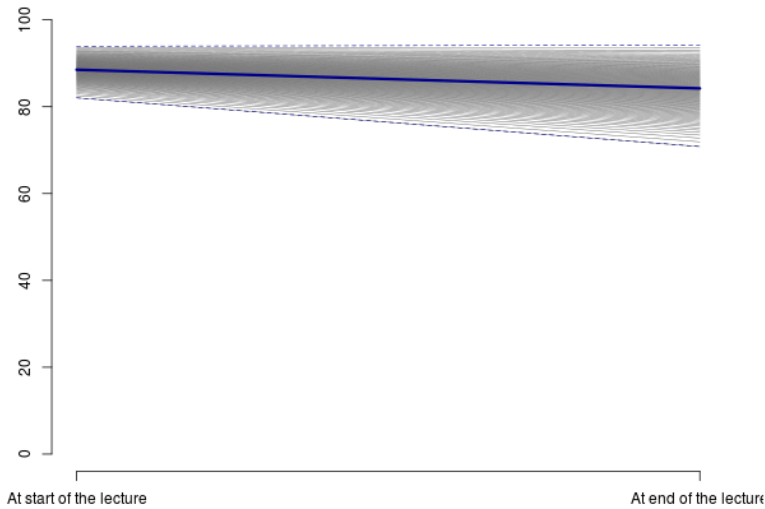
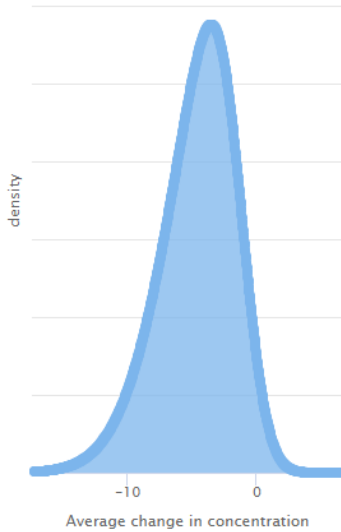
-4,6

Reasonable upperbound average change in concentration

-15

Fit distribution

Show implications

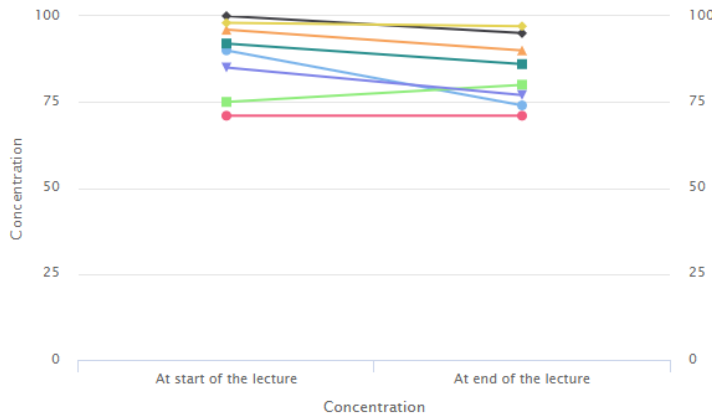


concentration				
2.5%	25%	50%	75%	97.5%
-11.2	-6.4	-4.3	-2.5	0.3
95% CI		50% CI		
[-11.2, 0.3]		[-6.4, -2.5]		

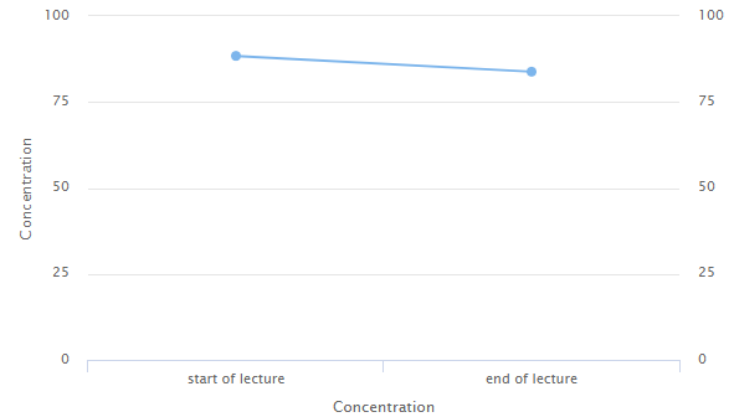




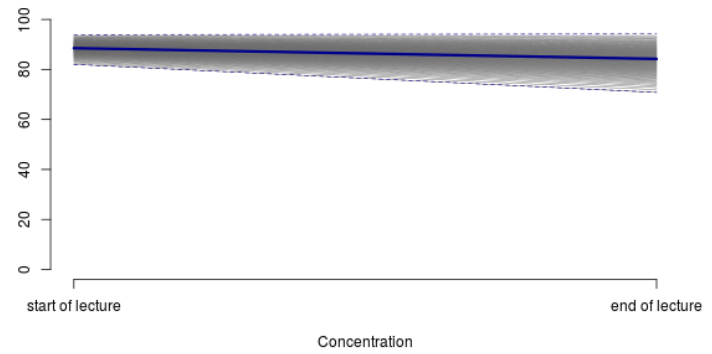
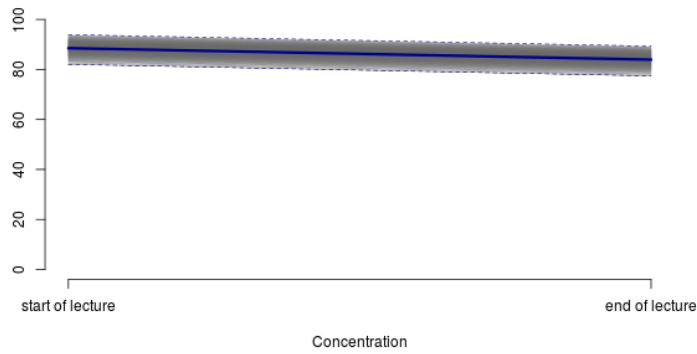
In this tab we provide a final summary of how we interpret your elicited beliefs and you can either agree to this or we go back to the relevant section of the procedure to adapt your input and our interpretation of your beliefs.



These are the concentration levels for your imagined individual children

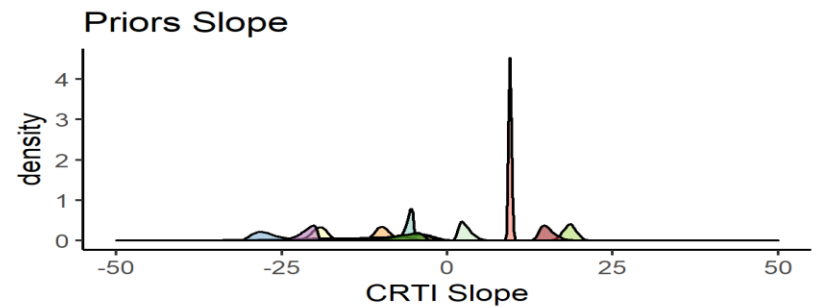
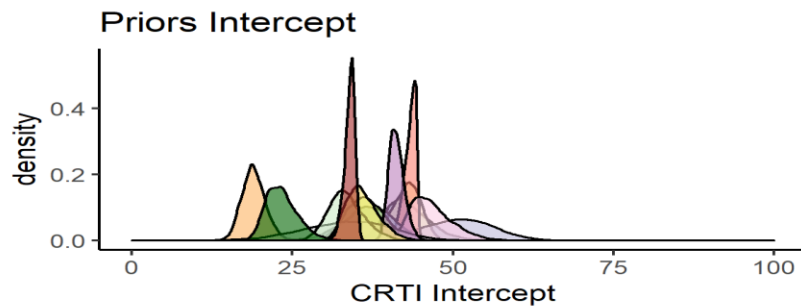
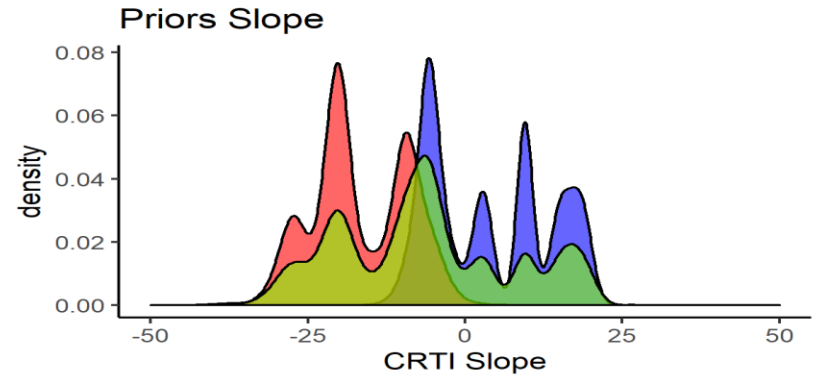
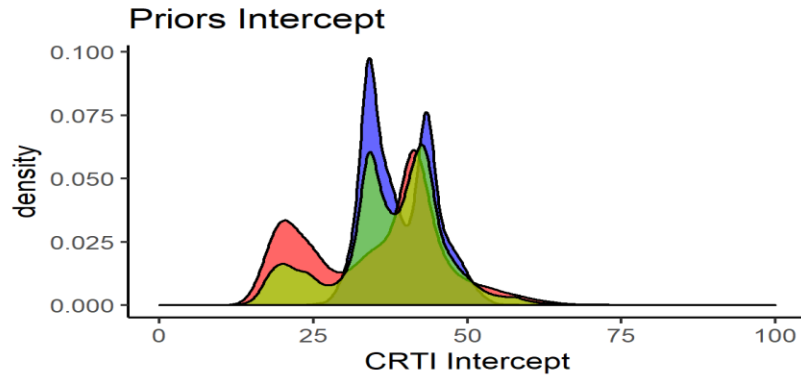


This is our interpretation of your beliefs regarding the average concentration levels at the start and the end of the lecture.





■ All experts ■ Nurses ■ Psychologists



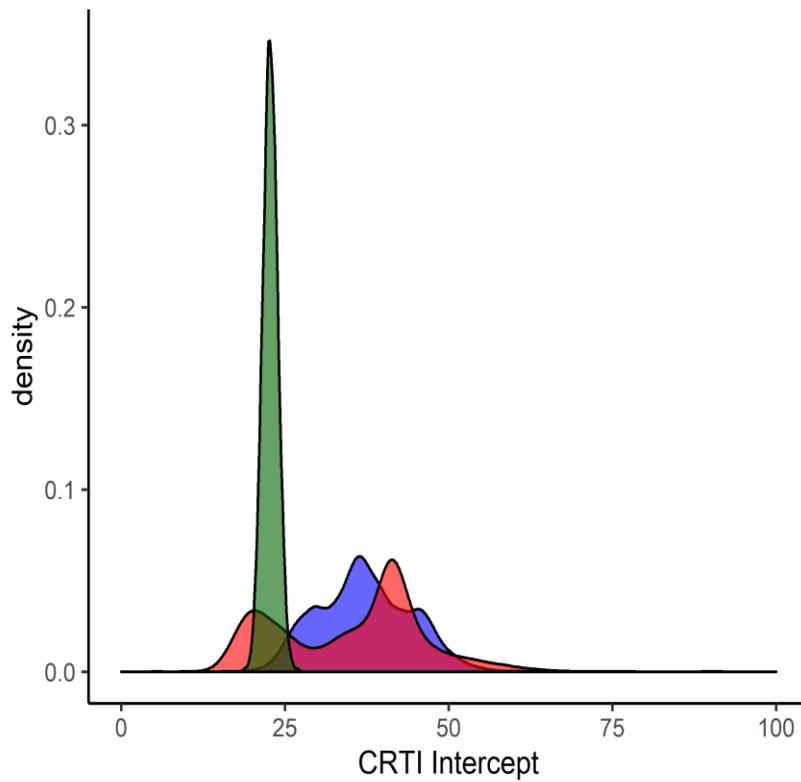
■ expert 1	■ expert 12	■ expert 2	■ expert 5	■ expert 8
■ expert 10	■ expert 13	■ expert 3	■ expert 6	■ expert 9
■ expert 11	■ expert 14	■ expert 4	■ expert 7	



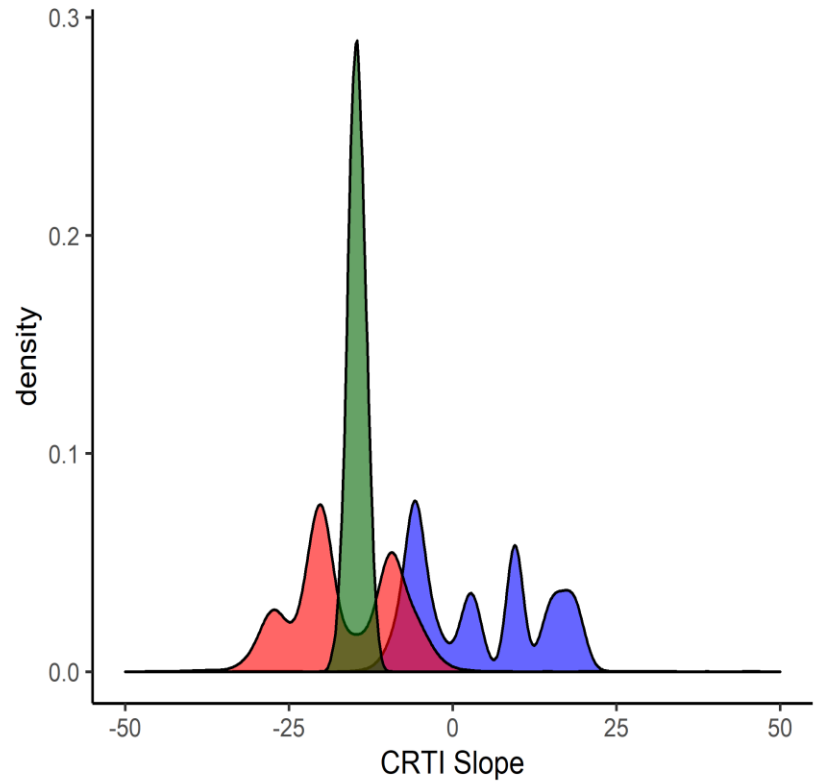


■ Nurses ■ Psychologists ■ Reference Posteriors

Mean of latent Intercept



Mean of latent Slope



Results – KL divergences

	Intercept	Slope
Benchmark 1	3.04	3.56
Benchmark 2	8.56	8.39
Nurses	8.19	5.88
Psychologists	1.99	2.18
All	2.72	2.63
Expert 1	42.87	59.18
Expert 2	45.16	25.87
Expert 3	6.71	1.23
Expert 4	72.86	55.38
Expert 5	5.66	98.32
Expert 6	2.1	22.17
Expert 7	79.2	59.61
Expert 8	46.97	4.37
Expert 9	2.48	1.28
Expert 10	43.74	67.55
Expert 11	12.78	64.56
Expert 12	99.94	4.88
Expert 13	0.35	3.62
Expert 14	75	74.11





Results – Audio recordings

- Referring specifically to (concepts of) PTSS
 - All psychologists
 - Only two nurses, though lost of mention of stress
- Expressing sentiment of more severe cases come to mind
 - 5 nurses – 1 psychologist
- Three psychologists reflected on linearity assumption of model





Results – Audio recordings

- Three experts actively reflected based on visual feedback and adjusted their input
 - One psychologist and two nurses
- One expert stated that although they were sure about the direction of the trajectory, they felt unsure about the associated numerical representation
- Finally, one expert repeatedly mentioned that they found the task hard to do





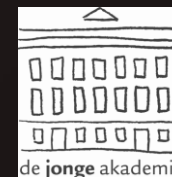
Universiteit Utrecht

Rens van de /SCHOOT

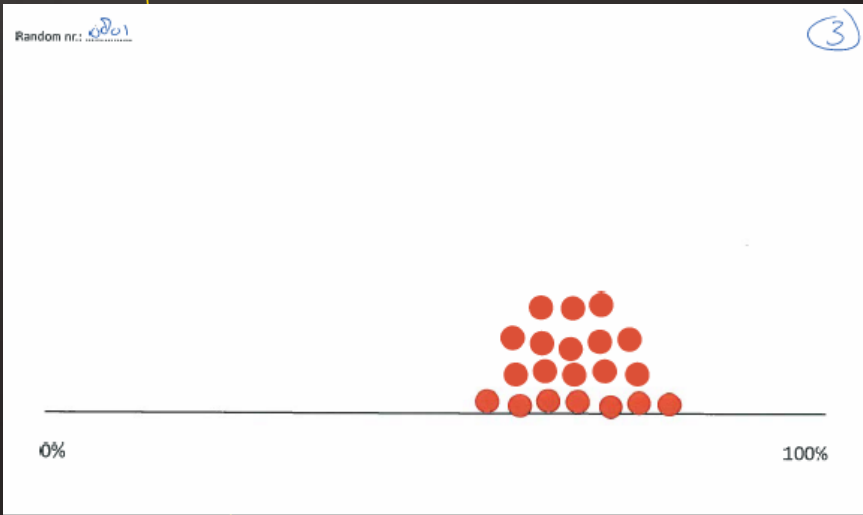
Dealing with Expert–Data (Dis)Agreement

A case study on Using Questionable Research
Practices to Survive in Academia

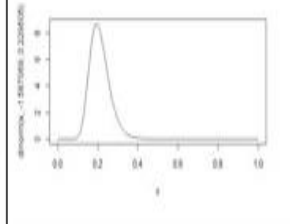
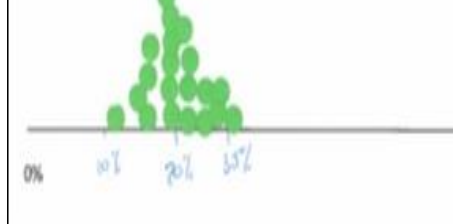
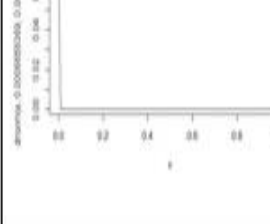
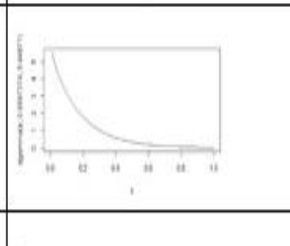
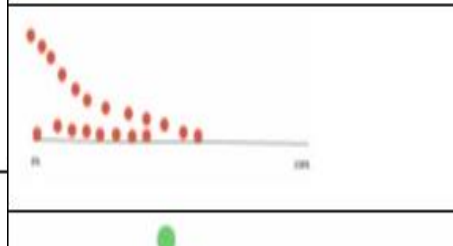
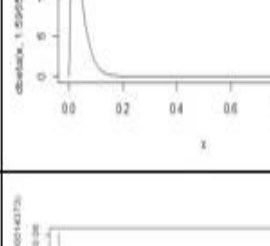
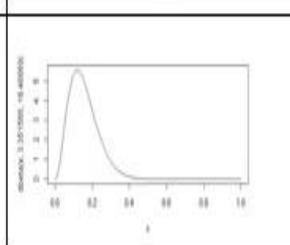
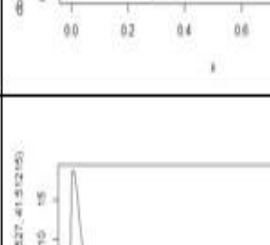
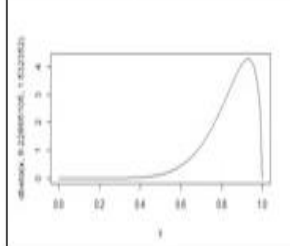
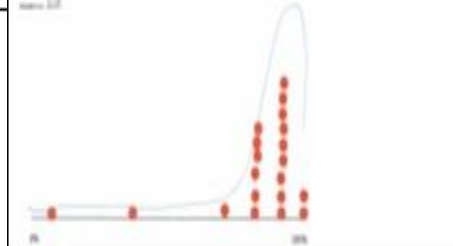
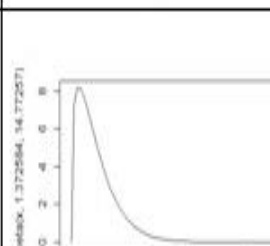
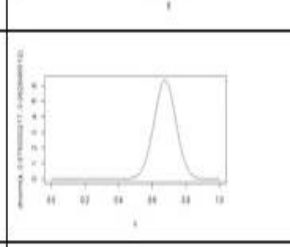
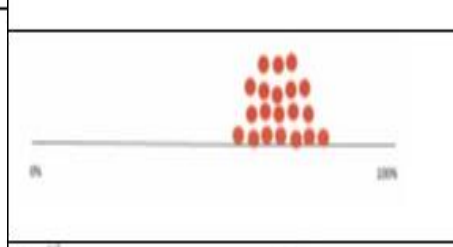
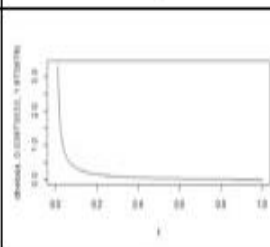
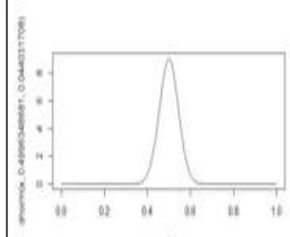
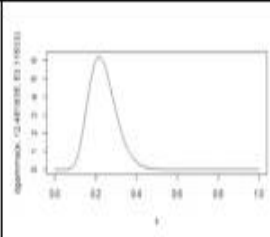
www.rensvandeschoot.com
[@RensvdSchoot](https://twitter.com/RensvdSchoot)
www.linkedin.com/in/rensvandeschoot

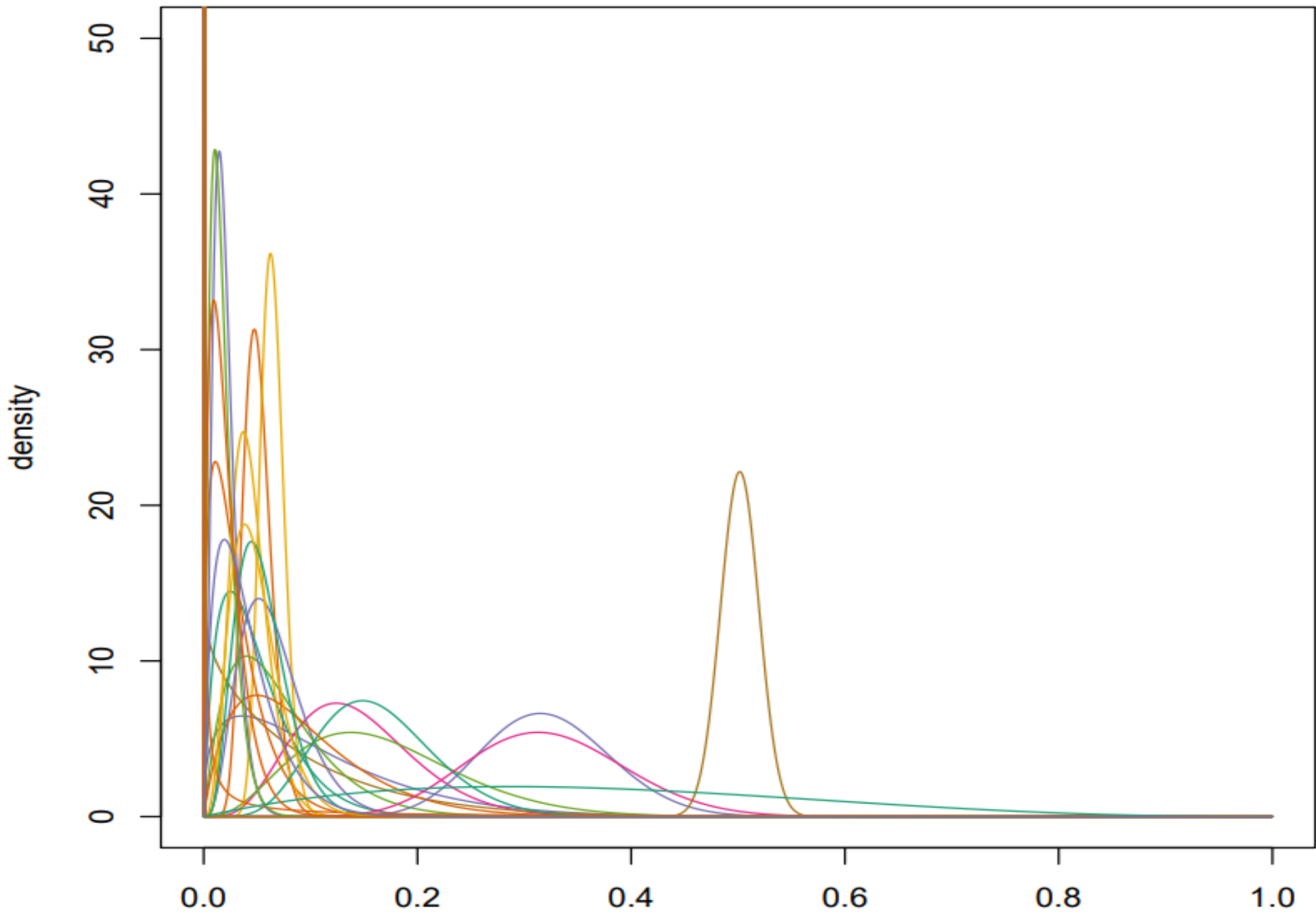


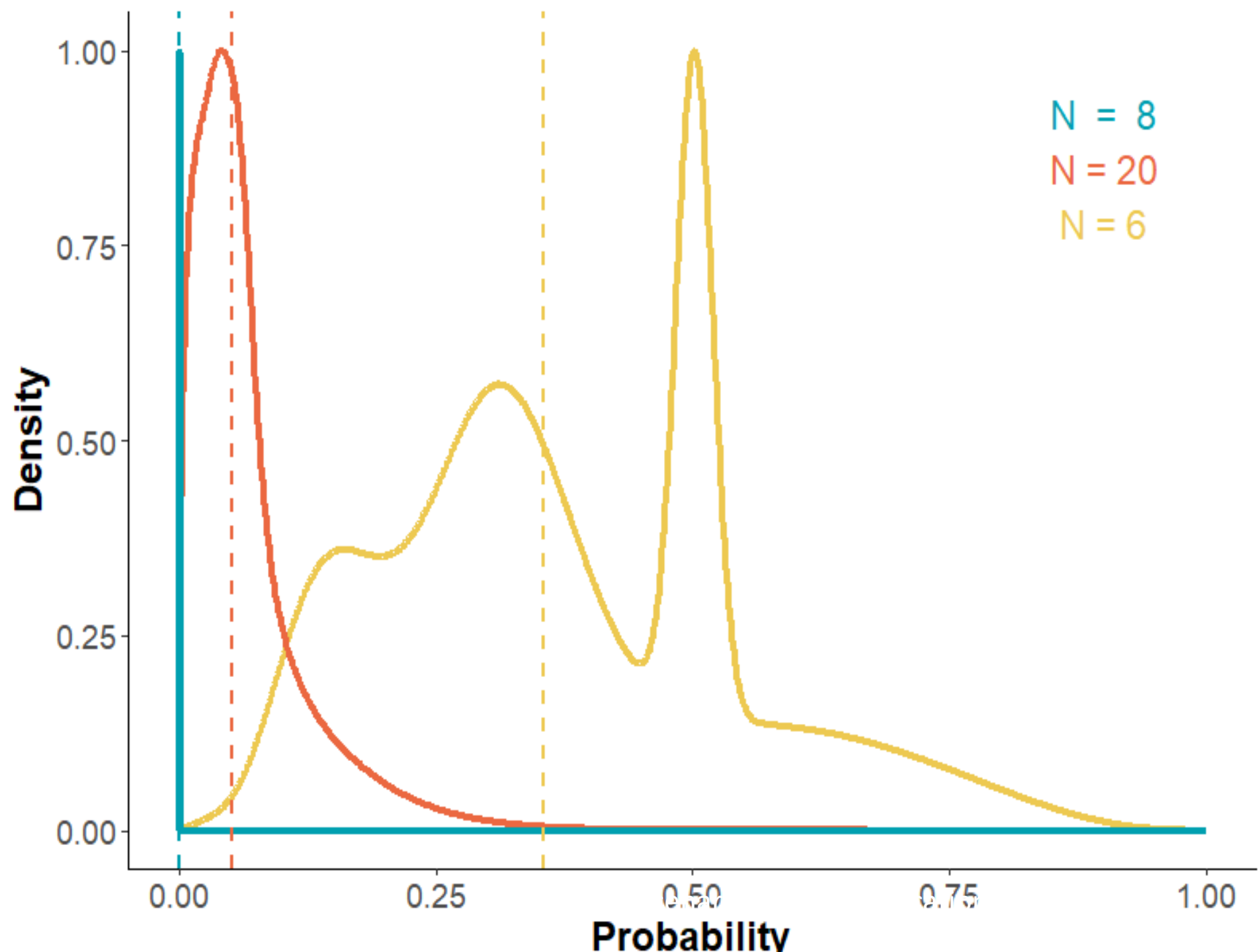


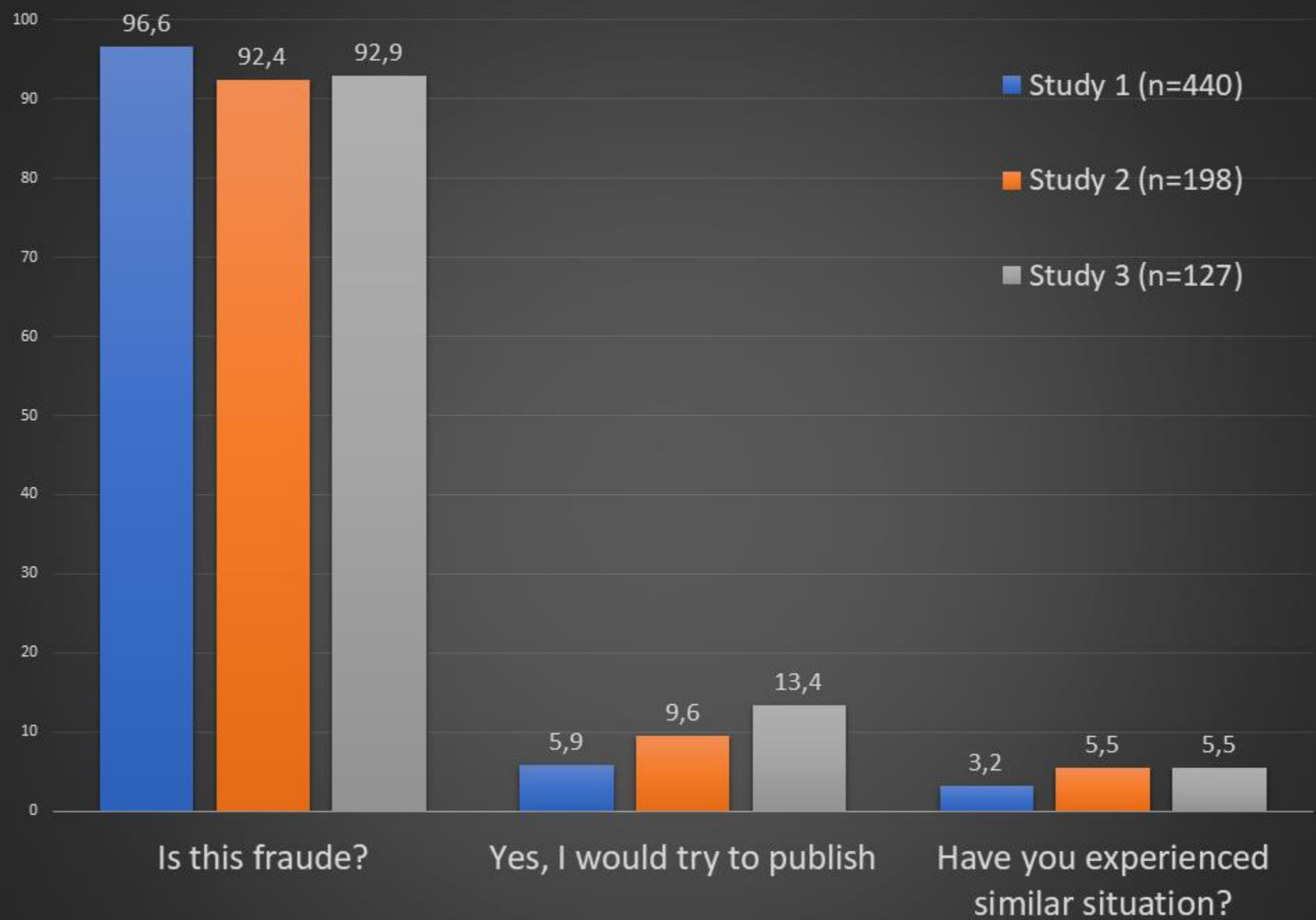


$N =$









Using the Data Agreement Criterion to Rank Experts' Beliefs

Duco Veen ^{1,*}, Diederick Stoel ², Naomi Schalken ¹, Kees Mulder ¹ and Rens van de Schoot ^{1,3}

¹ Department of Methods and Statistics, Utrecht University, 3584 CH 14 Utrecht, The Netherlands

² ProfitWise International, 1054 HV 237 Amsterdam, The Netherlands

³ Optentia Research Focus Area, North-West University, Vanderbijlpark 1900, South Africa

* Author to whom correspondence should be addressed.

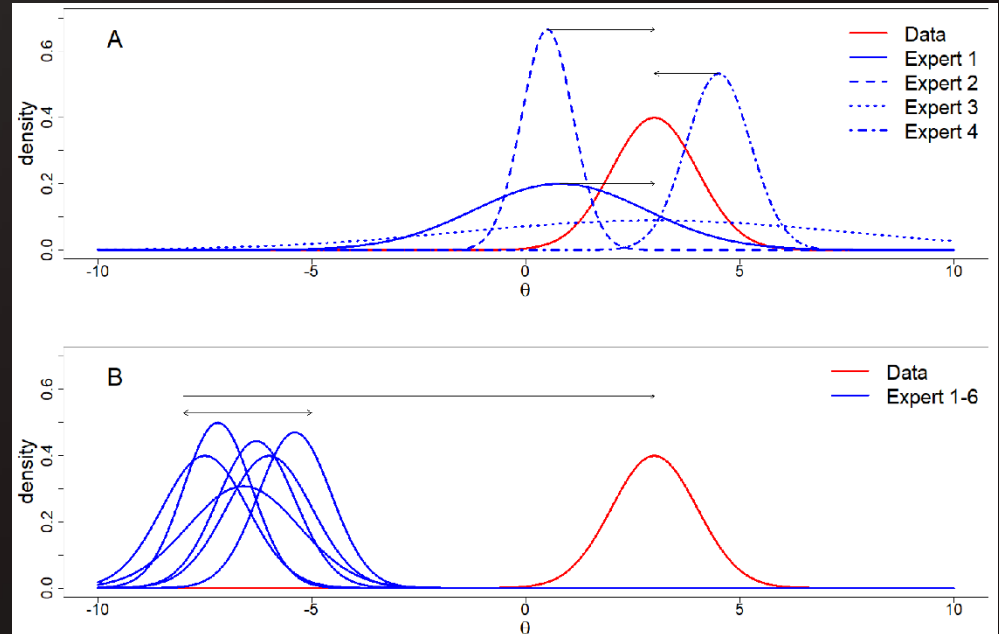
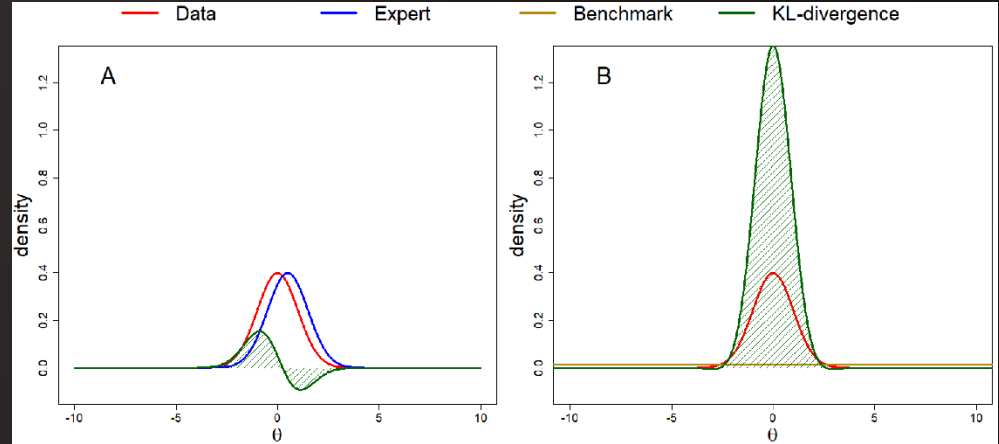
Received: 30 May 2018 / Revised: 7 August 2018 / Accepted: 7 August 2018 / Published: 9 August 2018

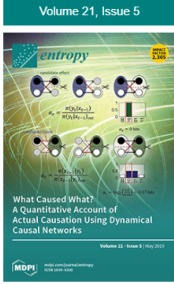
(This article belongs to the Special Issue Bayesian Inference and Information Theory)

Full-Text | PDF [493 KB, uploaded 9 August 2018] | Figures

Abstract

Experts' beliefs embody a present state of knowledge. It is desirable to take this knowledge into account when making decisions. However, ranking experts based on the merit of their beliefs is a difficult task. In this paper, we show how experts can be ranked based on their knowledge and their level of (un)certainly. By letting experts specify their knowledge in the form of a probability distribution, we can assess how accurately they can predict new data, and how appropriate their level of (un)certainly is. The expert's specified probability distribution can be seen as a prior in a Bayesian statistical setting. We evaluate these priors by extending an existing prior-data (dis)agreement measure, the Data Agreement Criterion, and compare this approach to





- Article Versions
- Abstract
 - Full-Text PDF (5764 KB)
 - Full-Text HTML
 - Full-Text XML
 - Full-Text Epub
 - Article Versions Notes

- Related Info
- Google Scholar
 - Order Reprints

- More by Authors
- on DOAJ
 - on Google Scholar

Entropy 2019, 21(5), 446; <https://doi.org/10.3390/e21050446>

Article
How the Choice of Distance Measure Influences the Detection of Prior-Data Conflict

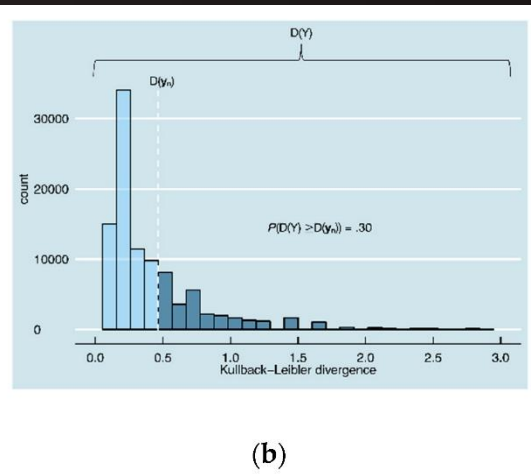
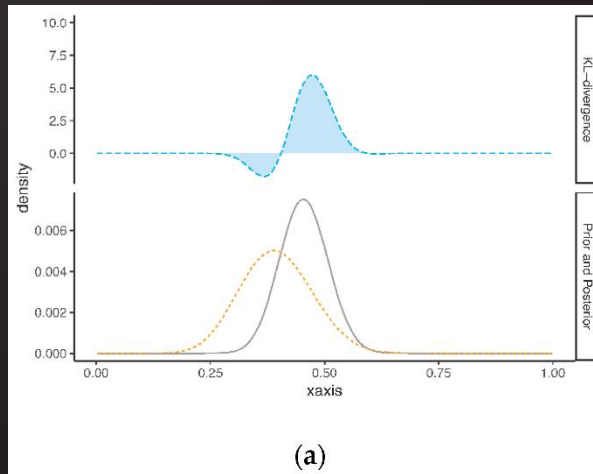
Kimberley Lek 1,* and Rens Van De Schoot 1,2

- 1 Department of Methods and Statistics, Utrecht University, 3584 CH 14 Utrecht, The Netherlands
- 2 Optentia Research Program, Faculty of Humanities, North-West University, Vanderbijlpark 1900, South Africa
- * Author to whom correspondence should be addressed.

Received: 28 March 2019 / Accepted: 23 April 2019 / Published: 29 April 2019

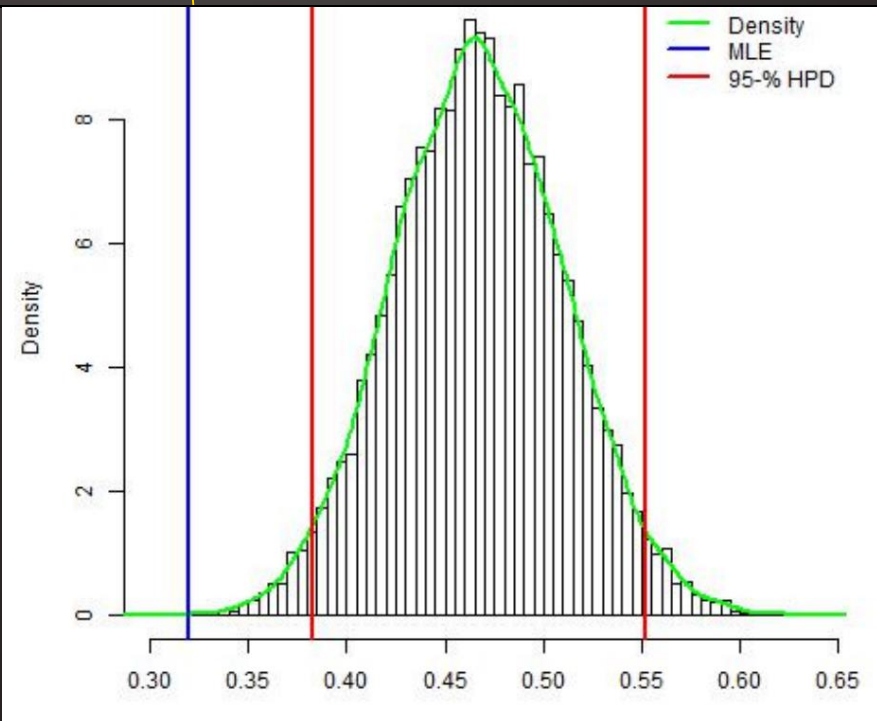
Abstract: The present paper contrasts two related criteria for the evaluation of prior-data conflict: the Data Agreement Criterion (DAC; Bousquet, 2008) and the criterion of Nott et al. (2016). One aspect that these criteria have in common is that they depend on a distance measure, of which dozens are available, but so far, only the Kullback-Leibler has been used. We describe and compare both criteria to determine whether a different choice of distance measure might impact the results. By means of a simulation study, we investigate how the choice of a specific distance measure influences the detection of prior-data conflict. The DAC seems more susceptible to the choice of distance measure, while the criterion of Nott et al. seems to lead to reasonably comparable conclusions of prior-data conflict, regardless of the distance measure choice. We conclude with some practical suggestions for the user of the DAC and the criterion of Nott et al.

Keywords: prior-data conflict; distance measure; Kullback-Leibler; data agreement criterion

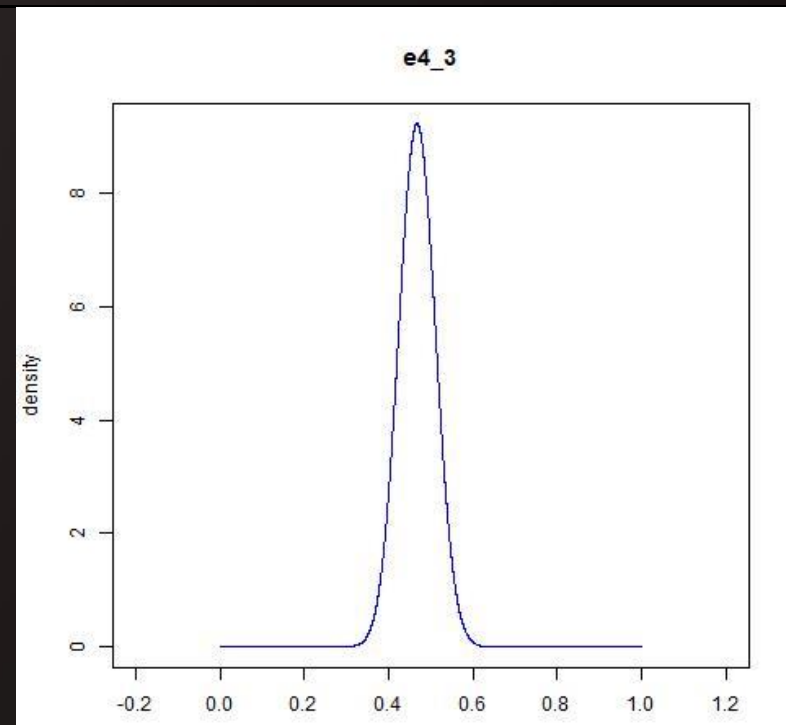




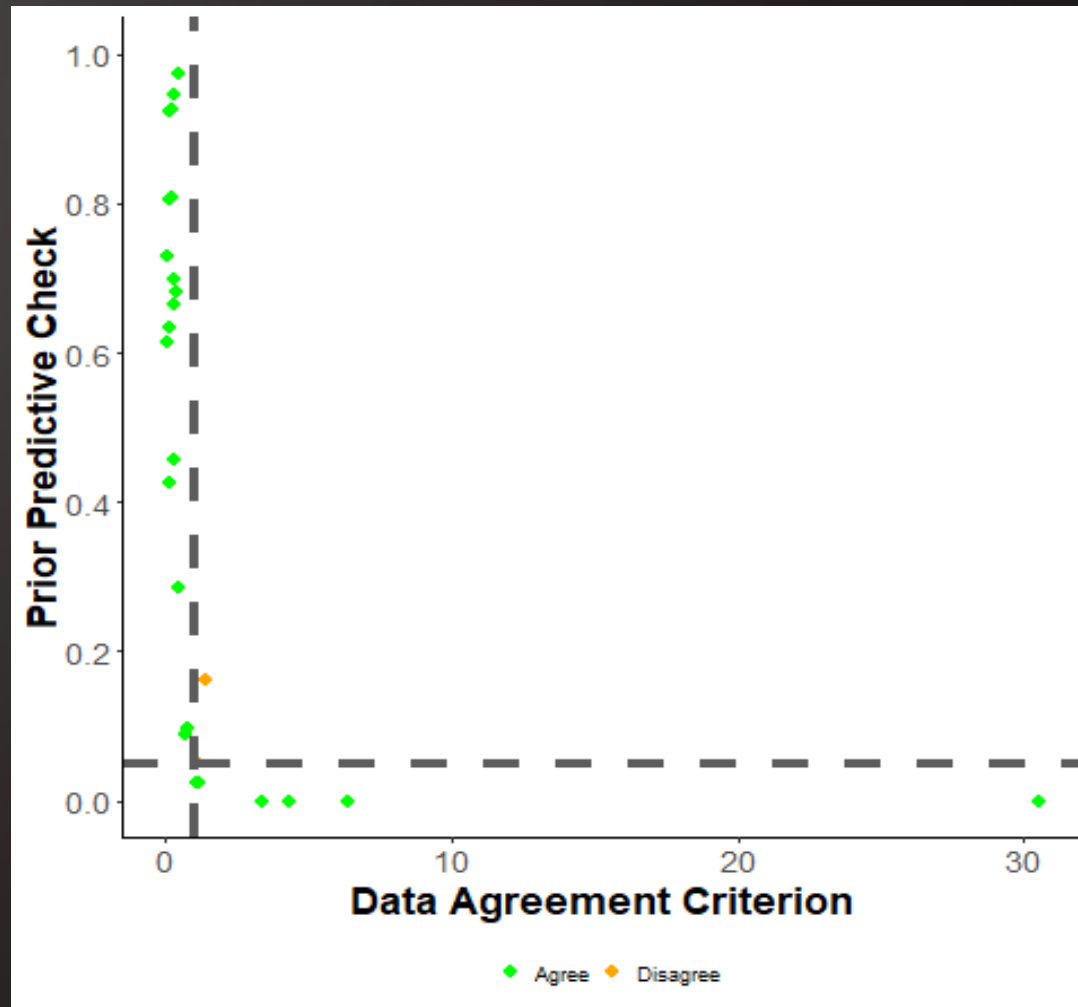
Prior Predictive Checking



(A)



(B)



20 experts (58.8%) showed no significant conflict with the data. Nine experts (26.5%) significantly underestimated the percentage of PhD-candidates who would be willing to publish with fabricated data, while the remaining 5 experts (14.7%) significantly overestimated this percentage.



	Percentage "Yes, I would try to publish"		
	Study 2	Study 3	Study 4
Scenario 1: data fabrication	5.9 (n=440)		
Scenario 1 (revised): data fabrication		9.6 (n=198)	13.4 (n=127)
Scenario 2: deleting outliers to get significant results	12.3 (n=407)		
Scenario 3: Salami slicing	32.0 (n=397)		
Scenario 3 (revised): Salami slicing		38.9 (n=185)	32.8 (n=119)
Scenario 4: gift authorship		59.2 (n=184)	58.8 (n=119)
Scenario 5: excluding information		12.1 (n=182)	16.1 (n=118)

SHARE REPORT



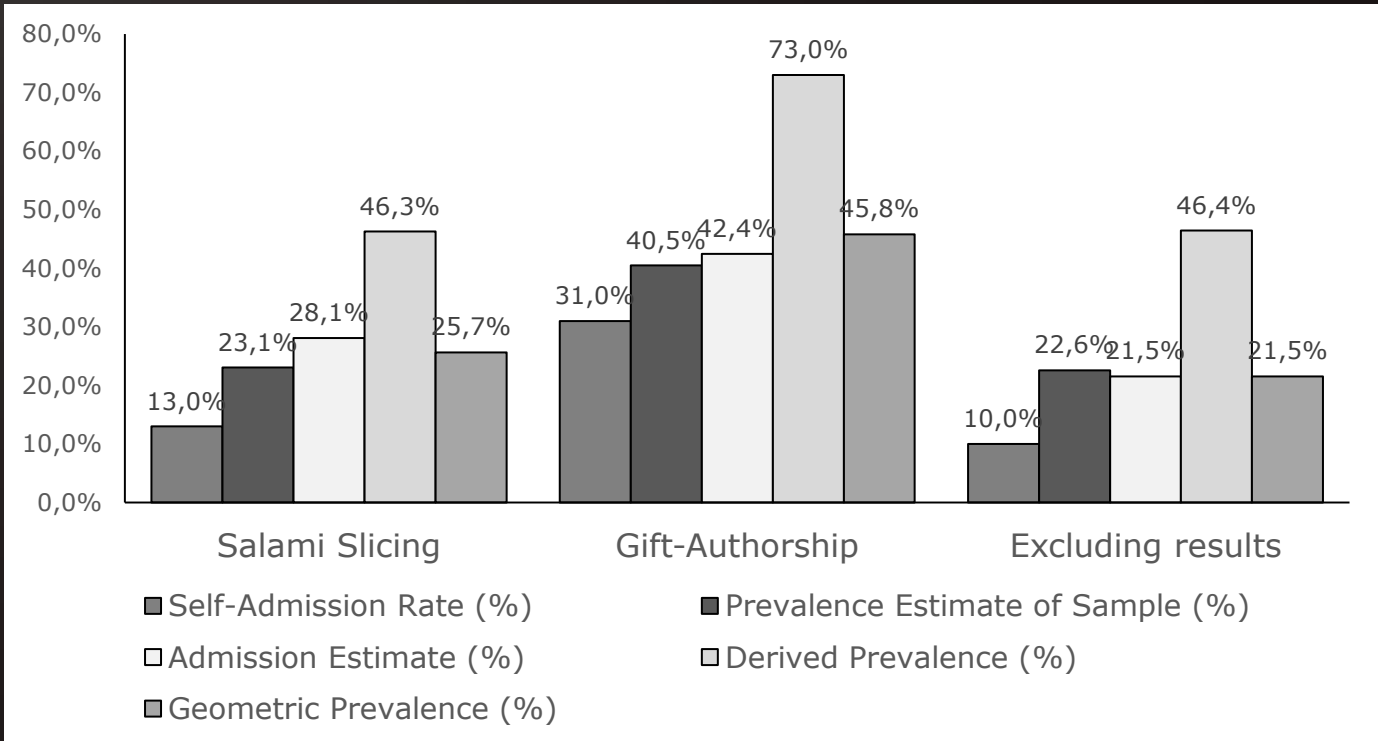
A Bayesian Truth Serum for Subjective Data

Dražen Prelec
+ See all authors and affiliations
Science 15 Oct 2004;
Vol. 306, Issue 5695, pp. 462-466
DOI: 10.1126/science.1102081

Article Figures & Data Info & Metrics eLetters PDF

Abstract

Subjective judgments, an essential information source for science and policy, are problematic because there are no public criteria for assessing judgmental truthfulness. I present a scoring method for eliciting truthful subjective data in situations where objective truth is unknowable. The method assigns high scores not to the most common answers but to the answers that are more common than collectively predicted, with predictions drawn from the same population. This simple adjustment in the scoring criterion removes all bias in favor of consensus: Truthful





Results Bayes Factor tests applied to Self Question.

Note. Bayes Factors > 1 favor the hypothesis that effect is present

	Proportion	Rep B_{r0}	JZS B_{10}	Equality B_{01}
Scenario 3				
Original	0.13		4.27E+05	
Replication	0.14	3.50E+04	2.64E+03	14.94
Scenario 4				
Original	0.31		3.12E+16	
Replication	0.29	1.20E+10	8.44E+08	13.70
Scenario 5				
Original	0.10		5.34E+03	
Replication	0.13	1.05E+04	9.27E+02	12.55



6-13% would publish ... why?



Just to it to survive in academia

"since it will get me closer to obtaining my PhD"

"It's not a solid yes, but a tentative one. I can image, just to be realistic, in terms of publishing pressures and not wanting to be out of contract, that this would be the best bet after all."



Pressure of supervisor

"No, unless the project leader also insists. In that case I would have a hard time refusing"

"If the supervisors tell me it's okay, I would try to publish the data."

"since it will get me closer to obtaining my PhD"



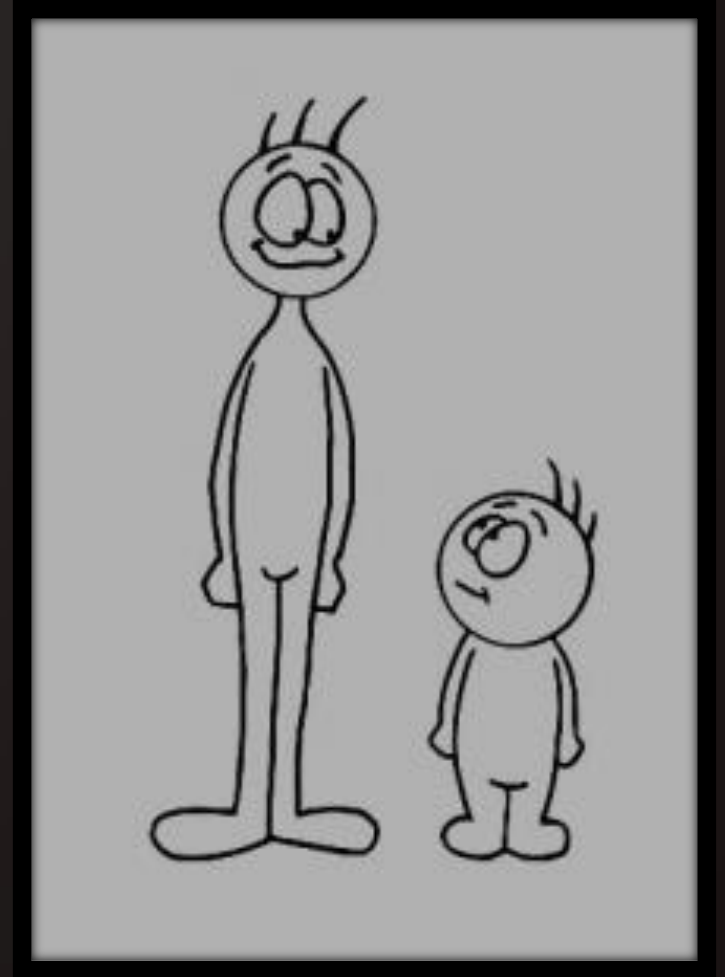
Why not publish...???

- Belief in a better world (8.8%)
- Afraid it will come back in the future (10.6%)
- To be safe -> conservative (15.9%)
- First ask a senior (22.4%)
- Because of moral conflict (34.2%)



"Never, this goes against all I stand for and this is not what research is about,

I feel very annoyed that this question is even being asked".



Expert elicitation only plan B?





It might be worth the effort!

NO!

- Experts provide unique information
 - Can be used to solve problems!
 - As additional data (enrich data)
 - As quality control



You might not want an alternative....





So... what's next??



Universiteit Utrecht