



Universiteit Utrecht

# A GENTLE INTRODUCTION TO BAYESIAN STATISTICS

SARA VAN ERP  
DUCO VEEN  
FLORIAN METWALY



## Goal of today presentation

- Discuss why priors are so important in Bayesian statistics
- Get a flavor of what you can do with priors
  - Expert information, historical data, literature
- What are some considerations you should think about
- By no means, today offers an exhaustive overview of all methods





# Goal of today presentation

- Get a feeling for priors
- What can they do?
- Hopefully some inspiration



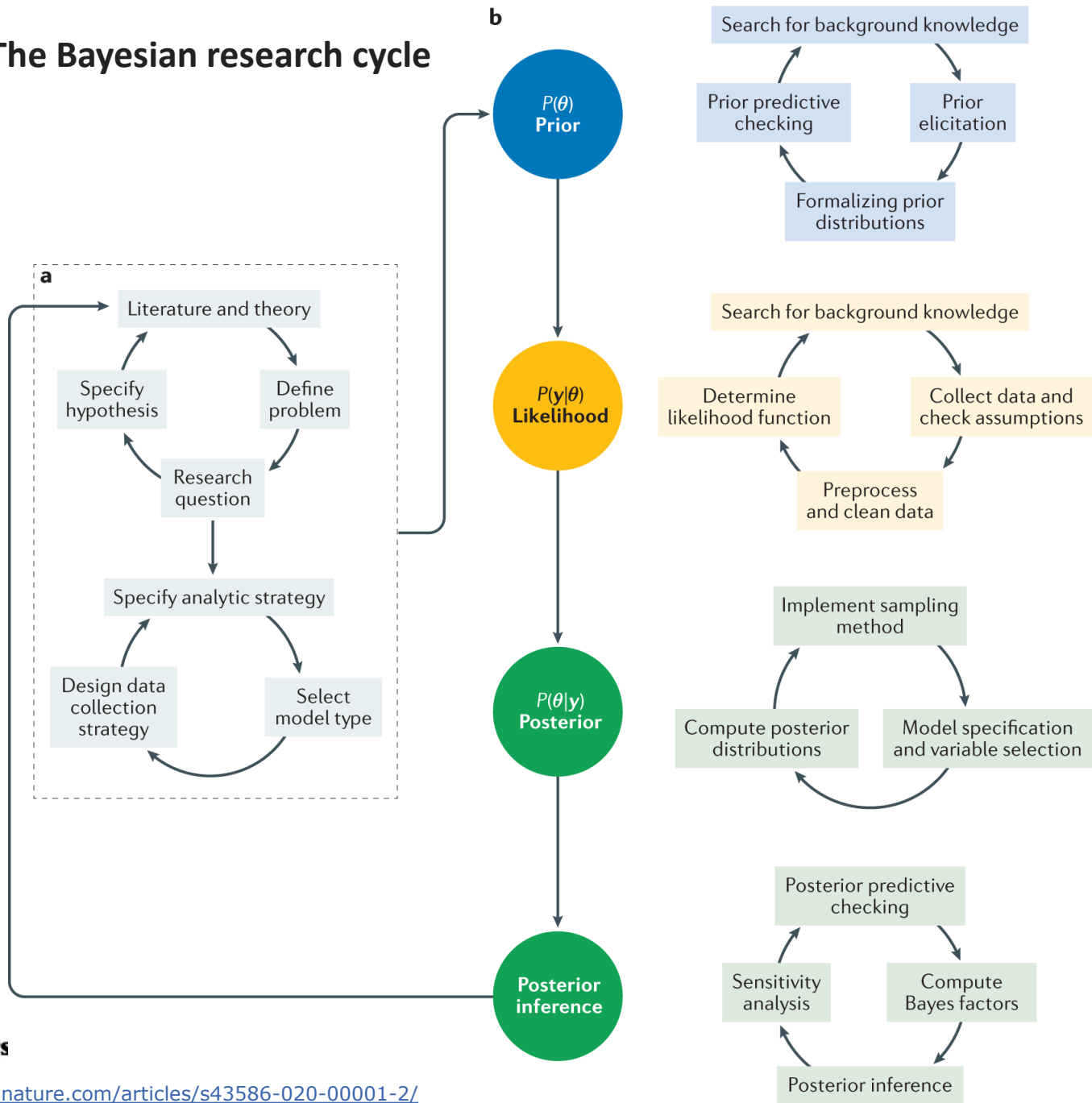
# Priors: Why all the fuss?





Univers

# The Bayesian research cycle





# Advantages & Disadvantages of Bayesian analyses

## Advantages:

- Natural approach to express uncertainty
- Ability to incorporate prior knowledge
- Increased model flexibility
- Full posterior distribution of the parameters
- Natural propagation of uncertainty

## Disadvantage:

- Slow speed of model estimation
- Some reviewers don't understand you ("give me the p-value")

Taken from: <https://bstat-edubron24.netlify.app/#part-1>





# Advantages & Disadvantages of Bayesian analyses

## Advantages:

- Natural approach to express uncertainty
- Ability to incorporate prior knowledge
- Increased model flexibility
- Full posterior distribution of the parameters
- Natural propagation of uncertainty

## Disadvantage:

- Slow speed of model estimation
- Some reviewers don't understand you ("give me the p-value")
- **Defending any prior information**
- **Or lack of prior information**
- **Or both**

Taken from: <https://bstat-edubron24.netlify.app/#part-1>





# Advantages & Disadvantages of Bayesian analyses

## Advantages:

- Natural approach to express uncertainty
- Ability to incorporate prior knowledge
- Increased model flexibility
- Full posterior distribution of the parameters
- Natural propagation of uncertainty

## Disadvantage:

- Slow speed of model estimation
- Some reviewers don't understand you ("give me the p-value")
- **Defending any prior information**
- **Or lack of prior information**
- **Or both**
- **You need to convince readers that what you did is reasonable**

Taken from: <https://bstat-edubron24.netlify.app/#part-1>





# Priors

- I want to start with some examples and challenge some ideas about priors
- See if we have the same intuition





## Example IRT model

- English exam results of the 2017–2018 academic year from No. 11 Middle School of Wuhan.
- IRT model – with item discrimination parameter and item difficulty parameter
- Estimates for the discrimination parameter
  - How well can the item differentiate examinees?

Liu, Y., Hu, G., Cao, L., Wang, X., & Chen, M. H. (2019). A comparison of Monte Carlo methods for computing marginal likelihoods of item response theory models. *Journal of the Korean Statistical Society*, 48(4), 503-512.

# Example IRT model

English exam results of the 2017–2018 academic year from No. 11 Middle School of Wuhan.

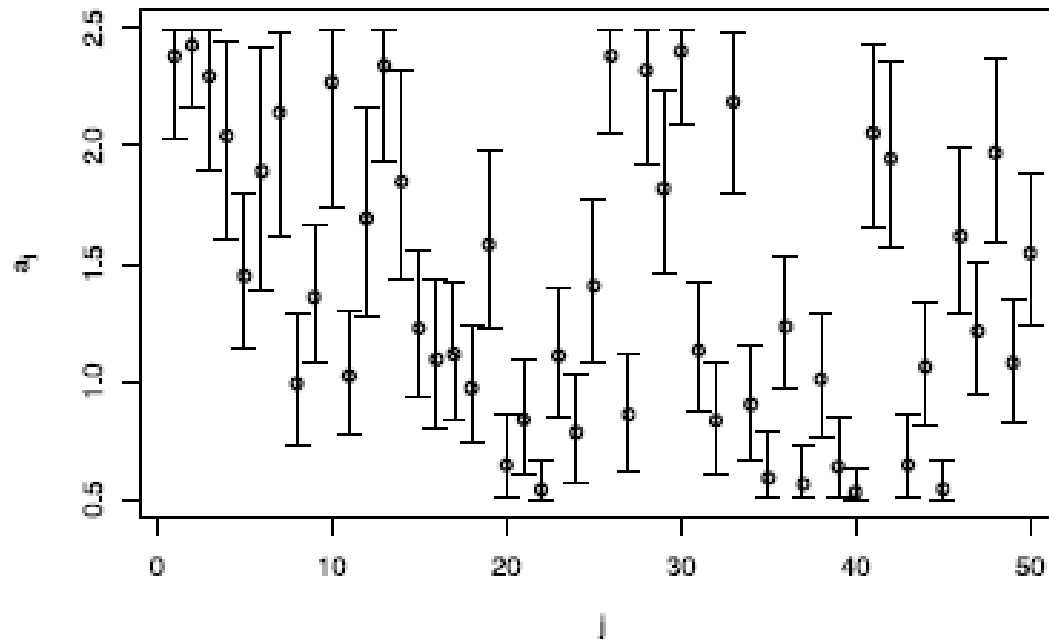


Fig. 1. Posterior medians and 95% credible intervals of  $a_j$ 's under the 2PL model.

Liu, Y., Hu, G., Cao, L., Wang, X., & Chen, M. H. (2019). A comparison of Monte Carlo methods for computing marginal likelihoods of item response theory models. *Journal of the Korean Statistical Society*, 48(4), 503-512.



# Example IRT model

What prior?  $U(0.5, 2.5)$

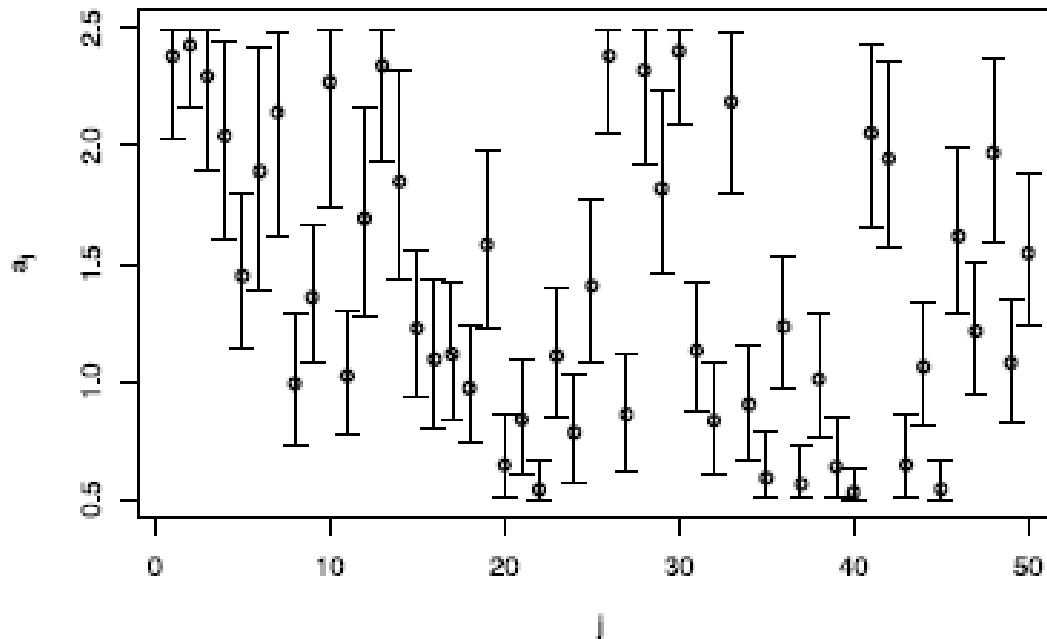


Fig. 1. Posterior medians and 95% credible intervals of  $a_j$ 's under the 2PL model.

Liu, Y., Hu, G., Cao, L., Wang, X., & Chen, M. H. (2019). A comparison of Monte Carlo methods for computing marginal likelihoods of item response theory models. *Journal of the Korean Statistical Society*, 48(4), 503-512.



# Example IRT model

What prior?  $U(0.5, 2.5)$  - > hard stop, can't be more or less!

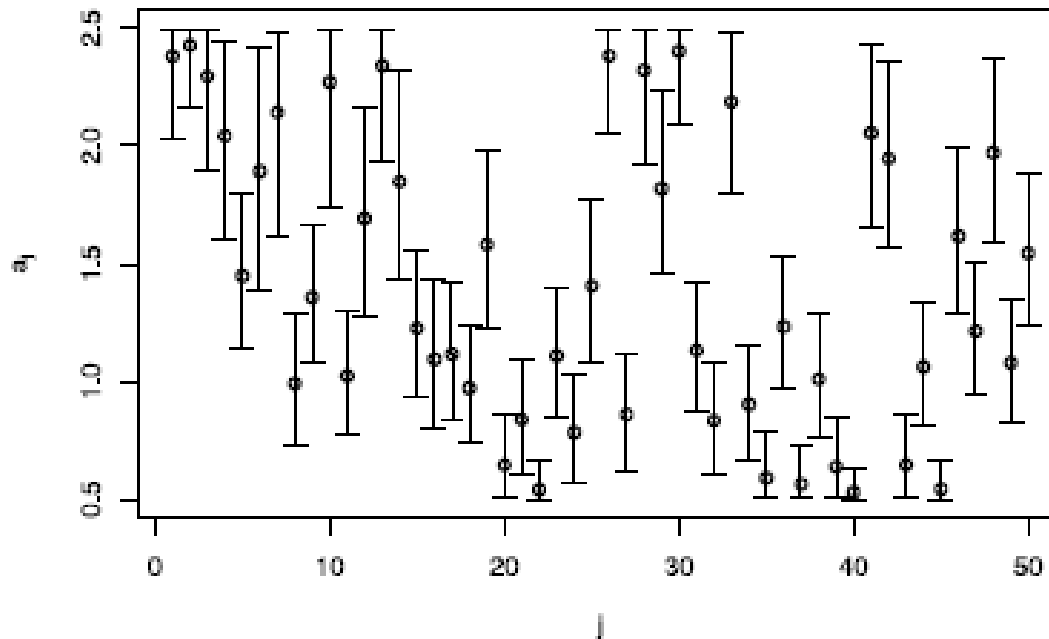


Fig. 1. Posterior medians and 95% credible intervals of  $a_j$ 's under the 2PL model.

Liu, Y., Hu, G., Cao, L., Wang, X., & Chen, M. H. (2019). A comparison of Monte Carlo methods for computing marginal likelihoods of item response theory models. *Journal of the Korean Statistical Society*, 48(4), 503-512.



# Example IRT model

What if we change the prior? **LN(0.5, 1)**

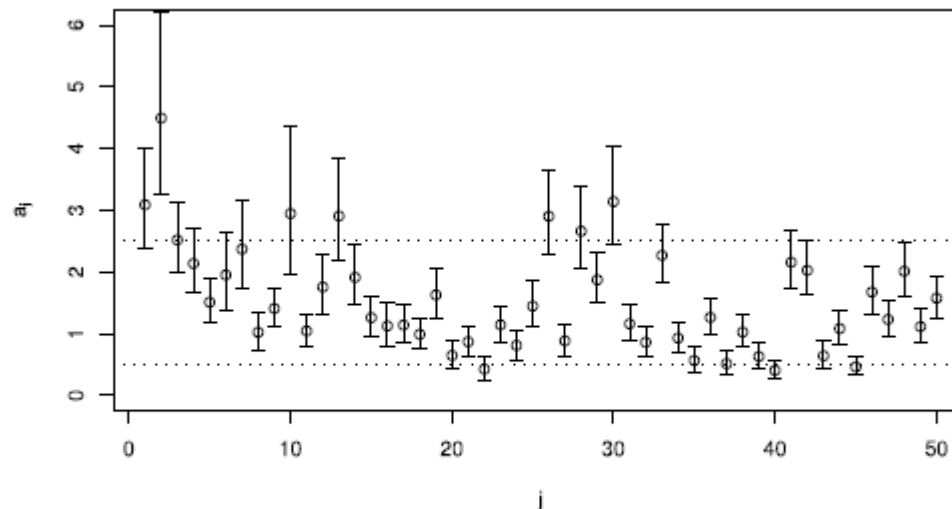
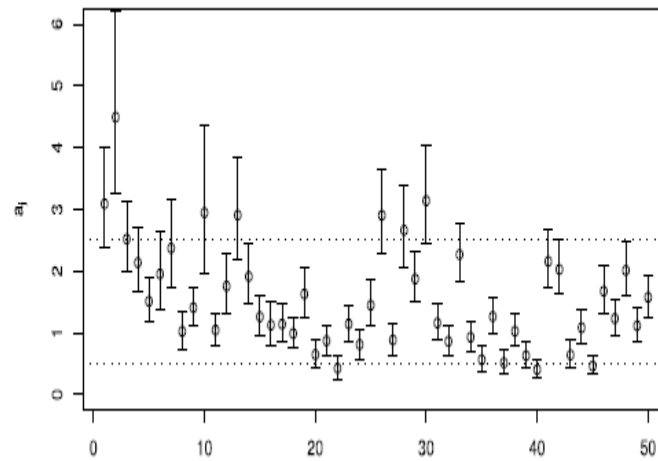
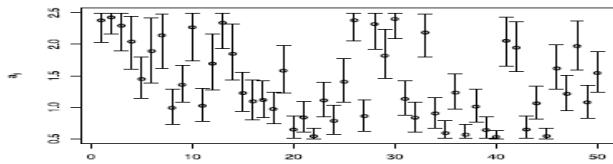


Fig. 1. Posterior medians and 95% credible intervals of  $a_j$ 's under the 2PL model using  $LN(0.5, 1)$  prior on  $a_j$ 's. Dashed lines indicate bounds of  $U(0.5, 2.5)$  prior.

Veen, D., & Klugkist, I. (2019). Standard errors, priors, and bridge sampling: A Discussion of Liu et al. *Journal of the Korean Statistical Society*, 48(4), 515-517.



# Example when we restrict to much



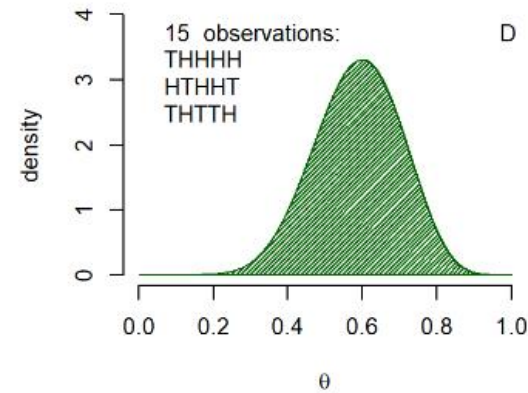
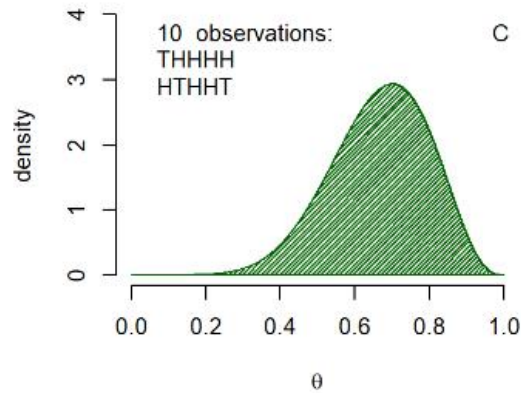
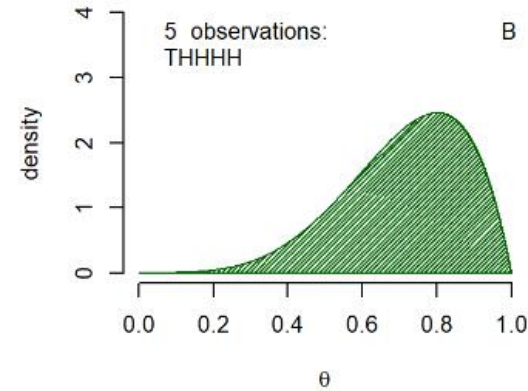
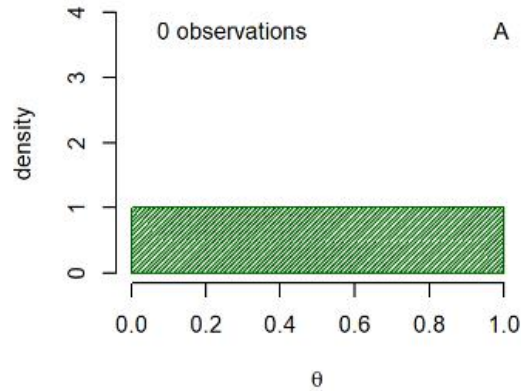
# Types of priors

- Non-informative (uninformative, diffuse, flat)
  - Weakly informative
  - Informative
  - Improper priors
- 
- This is sometimes misleading:
    - Is  $[-\infty, \infty]$  uninformative?
    - Is 132 as likely as 2.4, really?
    - What about a transformed parameter space?

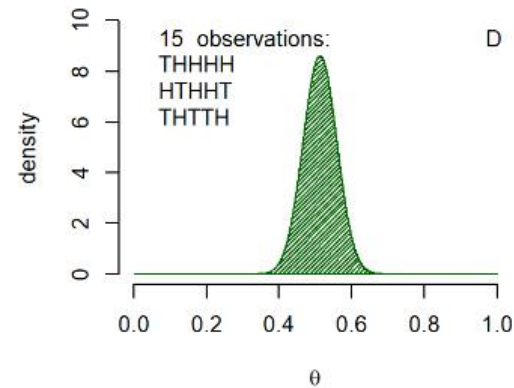
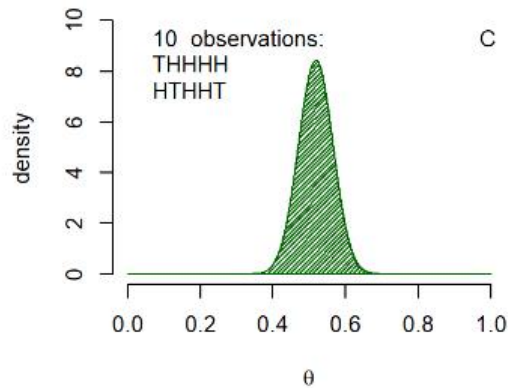
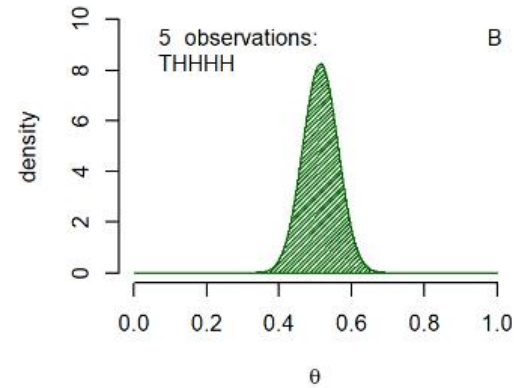
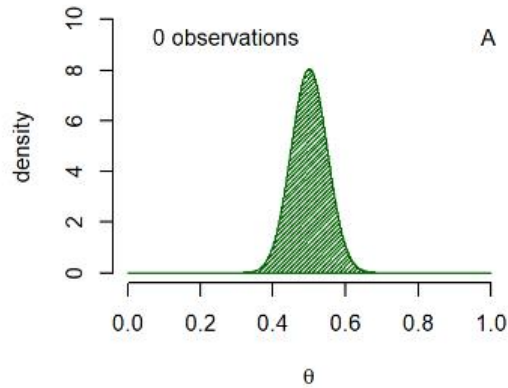




# Classical examples – 0, 1 data

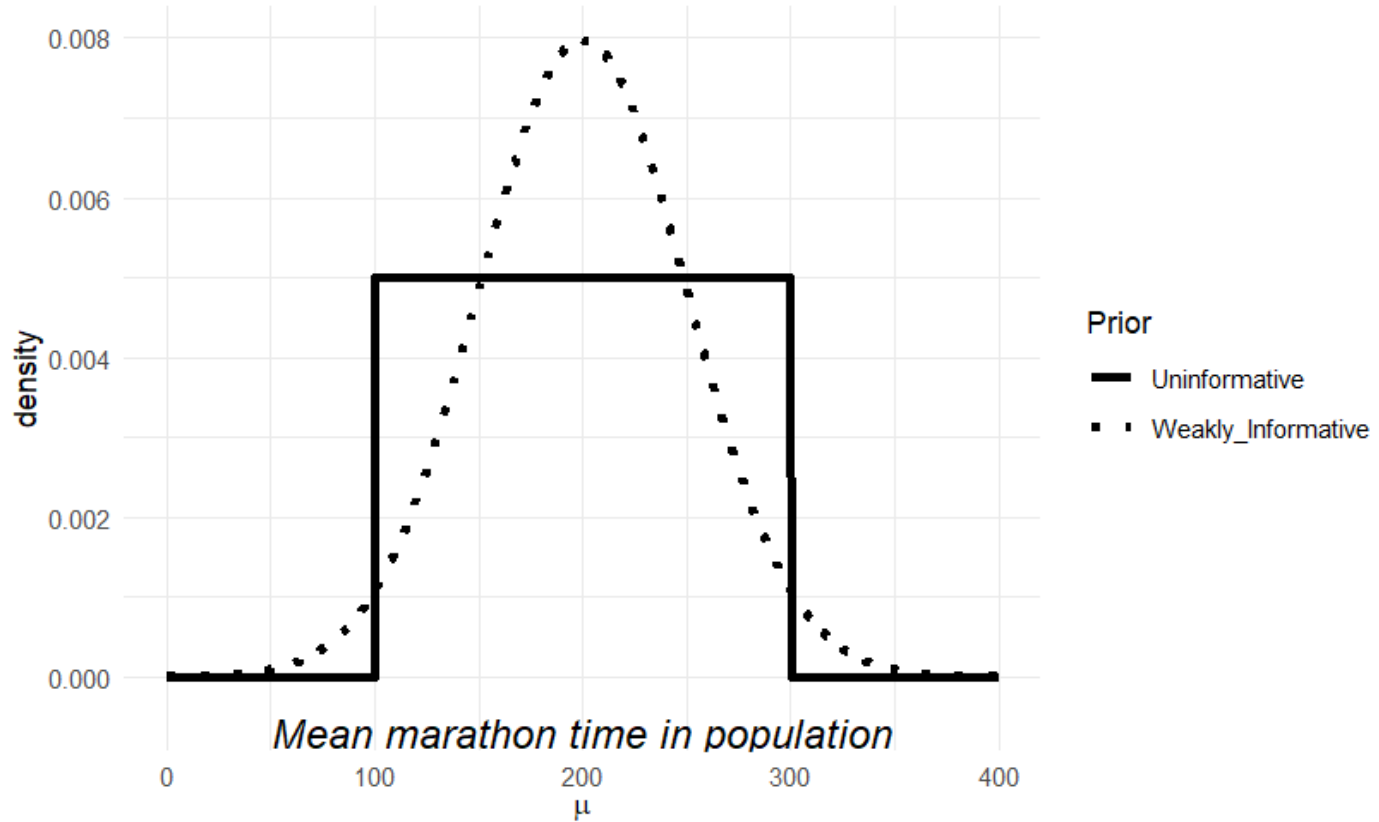


# Classical examples – 0, 1 data

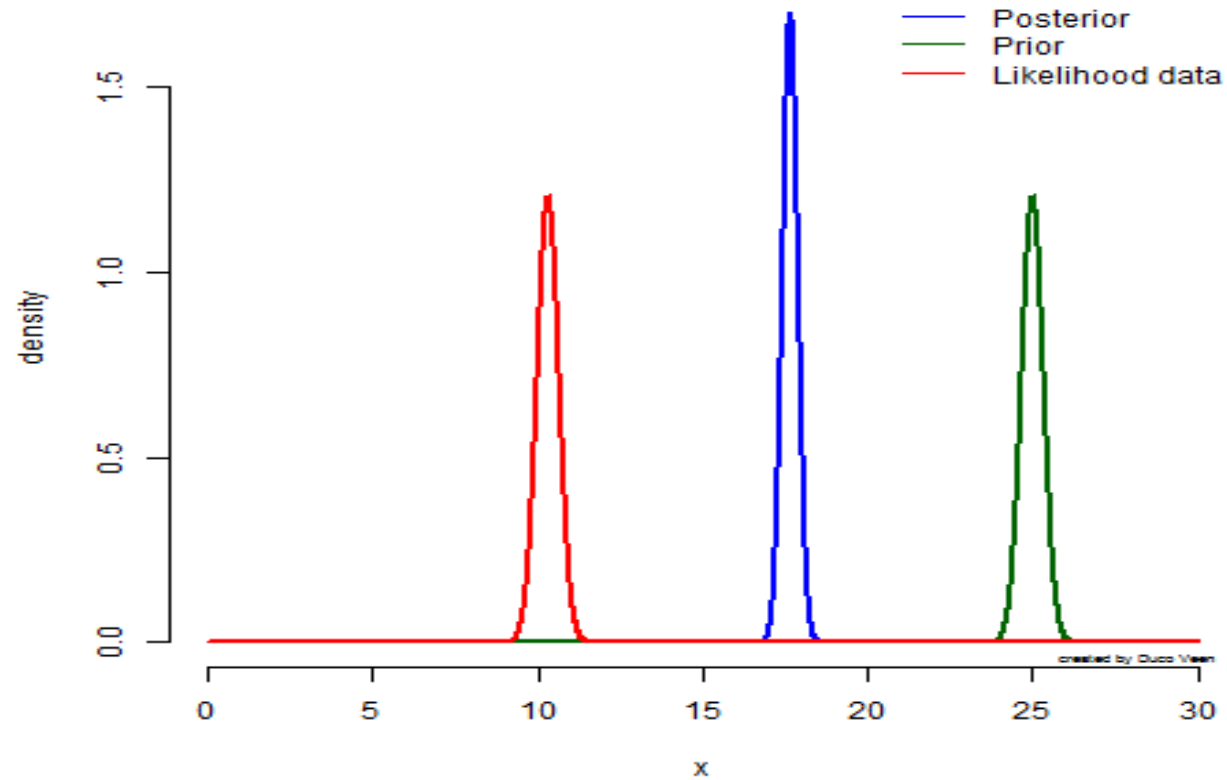


# Is this distinction that easy?

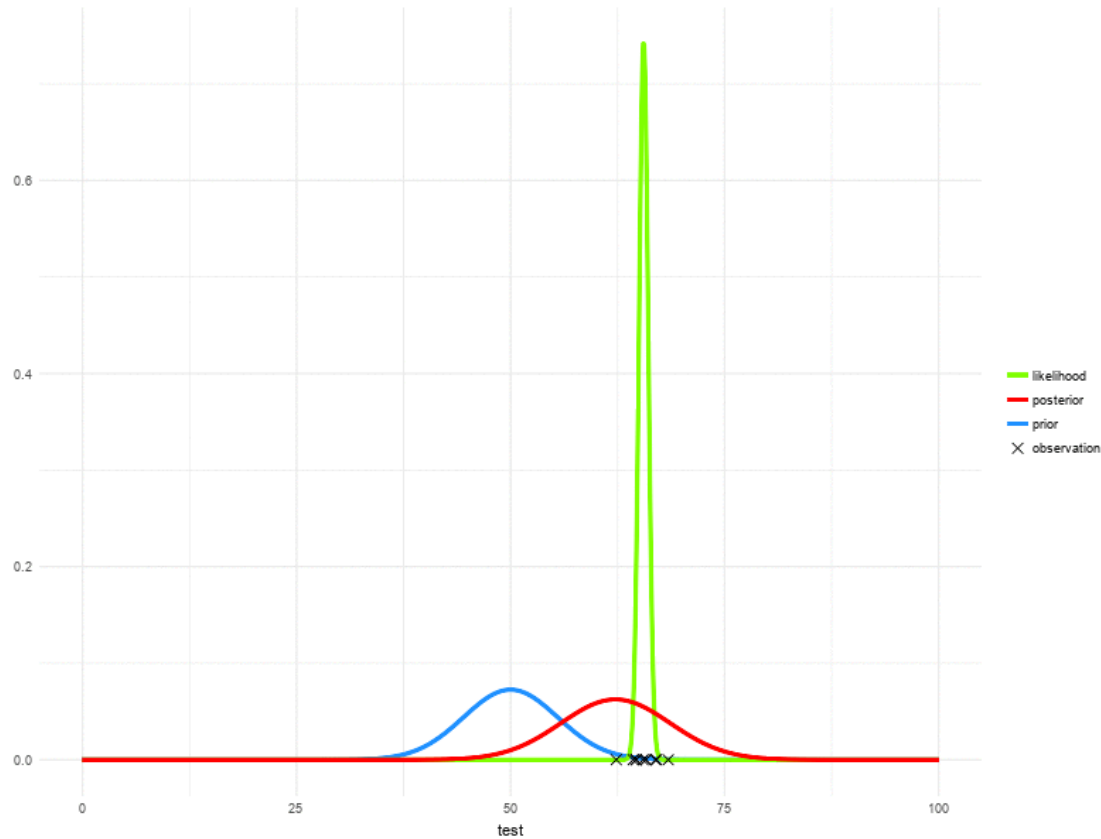
## Type of priors (visually)



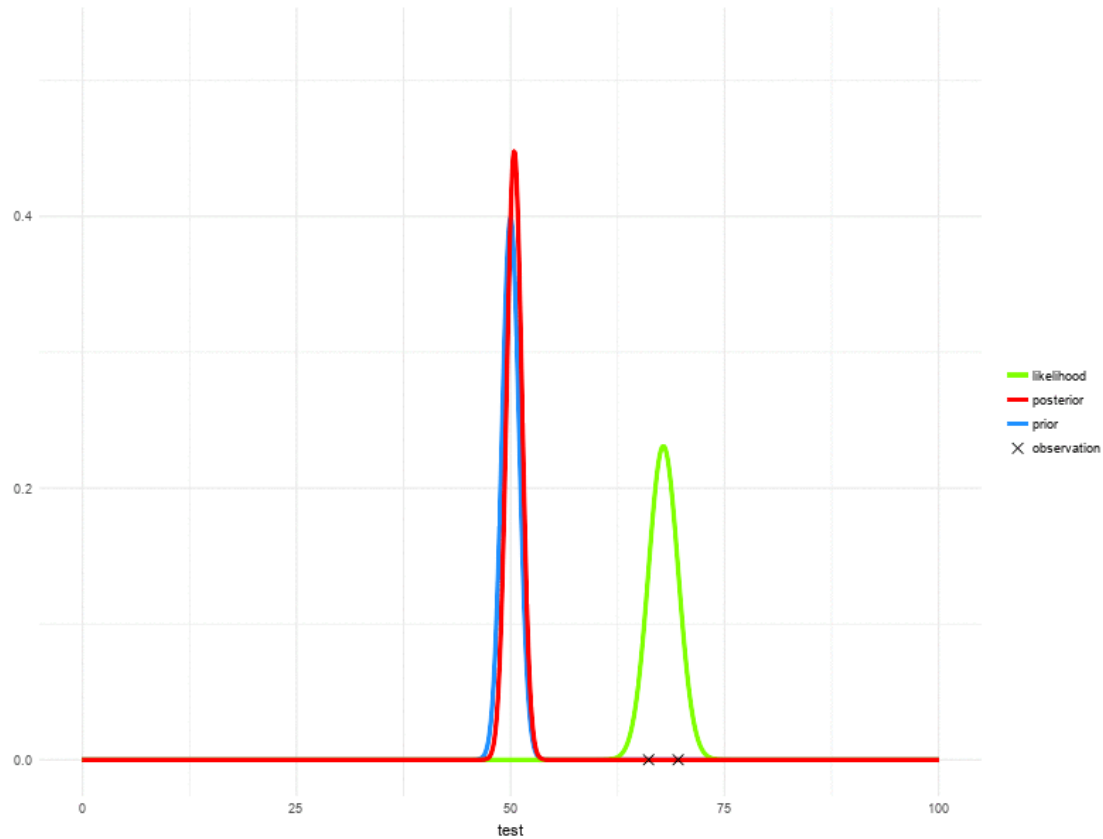
# Classical examples – Normal



# Classical examples – Increasing prior knowledge



# Classical examples – Increasing data



# Classical examples – Normal

- But what assumptions are made?
- What about the variance parameter in the model



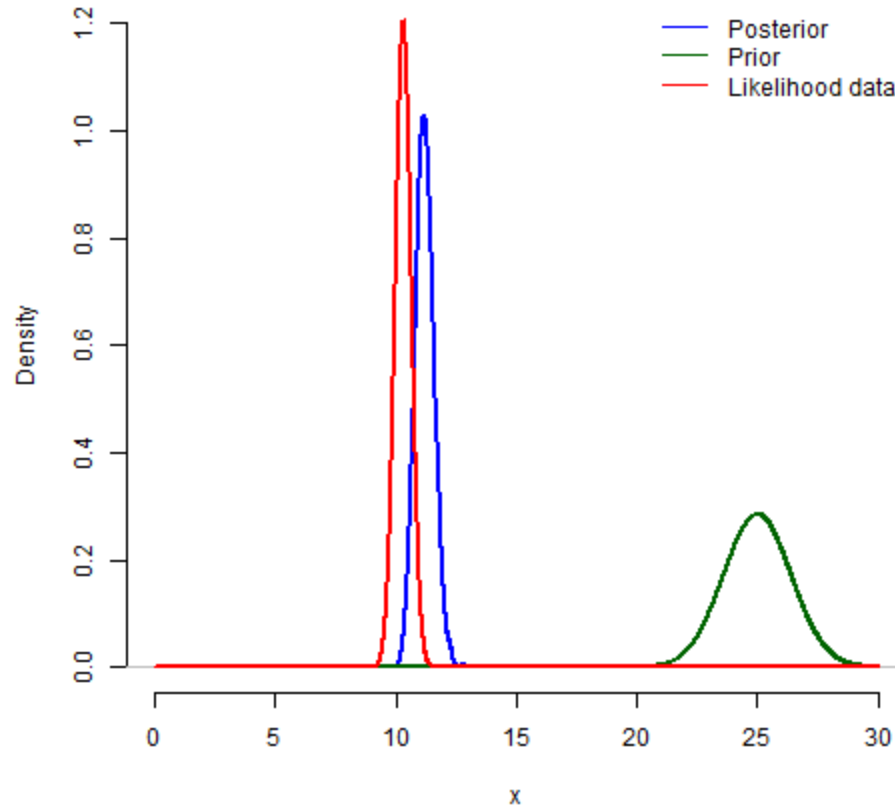
# Classical examples – Normal

- But what assumptions are made?
- What about the variance parameter in the model
  - **FIXED!**
  - **What happens if we let it go?**

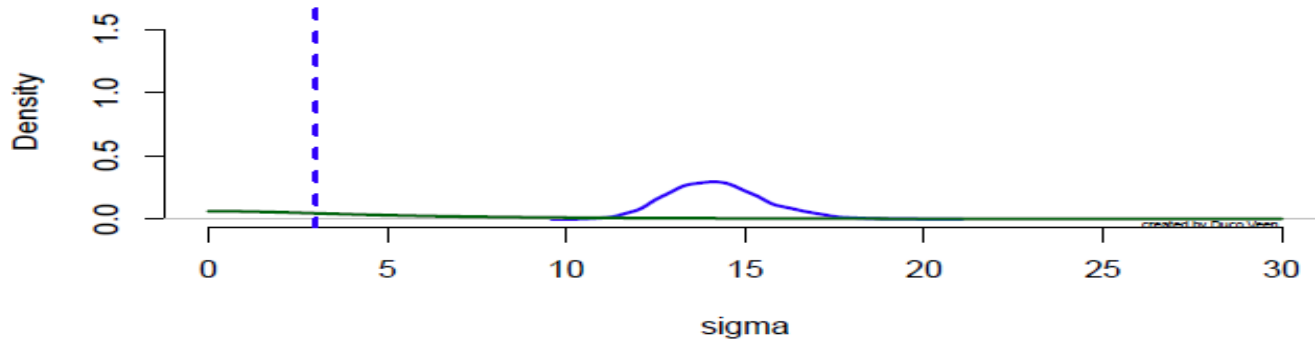
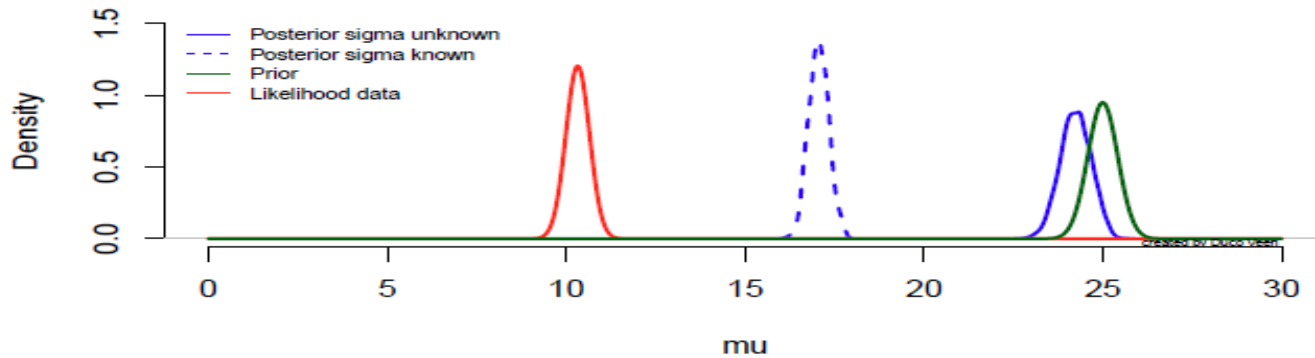




# Classical examples – With Sigma Unknown



# Classical examples – With Sigma (Un)known

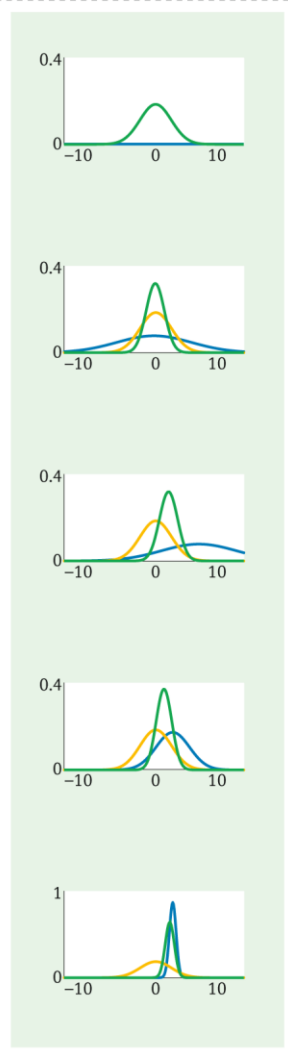
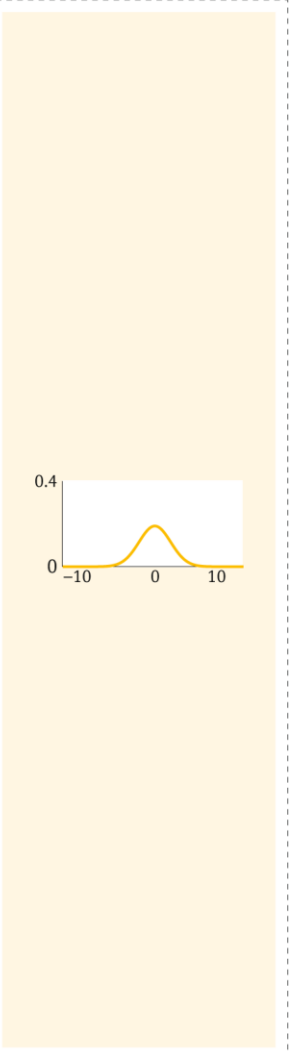
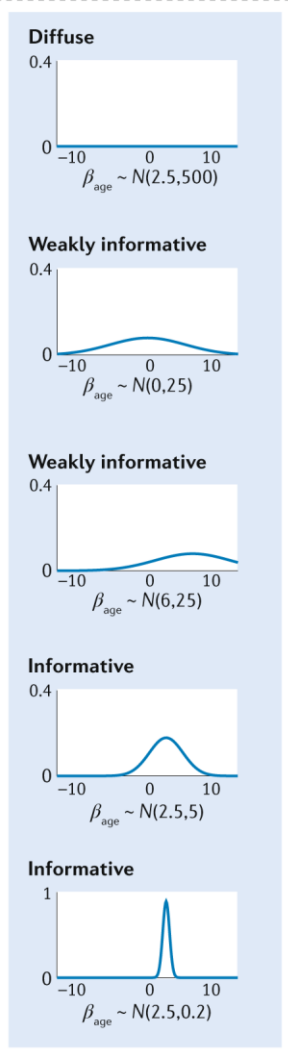




$P(\theta)$   
Prior

$P(y|\theta)$   
Likelihood

$P(\theta|y)$   
Posterior



[www.nature.com/articles/s43586-020-00001-2/](http://www.nature.com/articles/s43586-020-00001-2/)  
Universiteit Utrecht

# Types of priors

- Uninformative
- Weakly informative
- Informative
- Improper priors
- This is sometimes misleading:
  - Is  $[-\infty, \infty]$  uninformative?
  - Is 132 as likely as 2.4, really?
  - What about a transformed parameter space?





## Types of priors

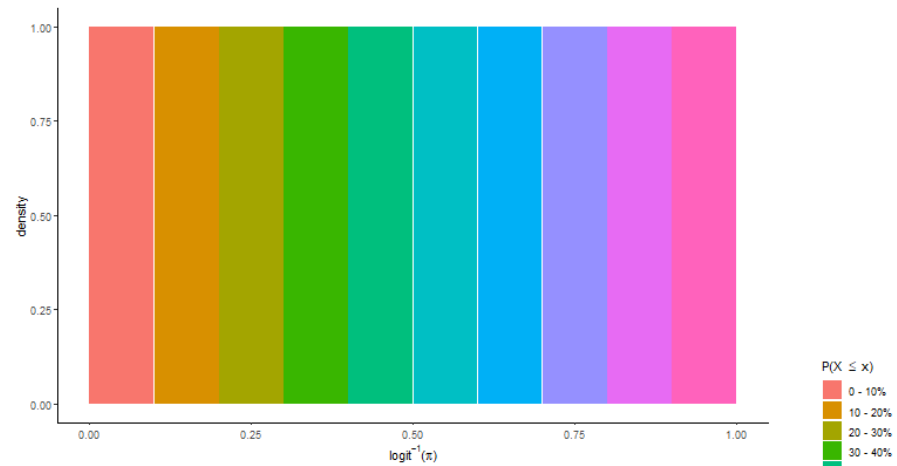
- Uninformative
  - Weakly informative
  - Informative
  - Improper priors
- 
- Informativeness can only be judged in comparison to the likelihood
  - Prior predictive checking can help to see the informativeness on the scale of the outcome
    - Especially helpful for large models

Suggested help source:

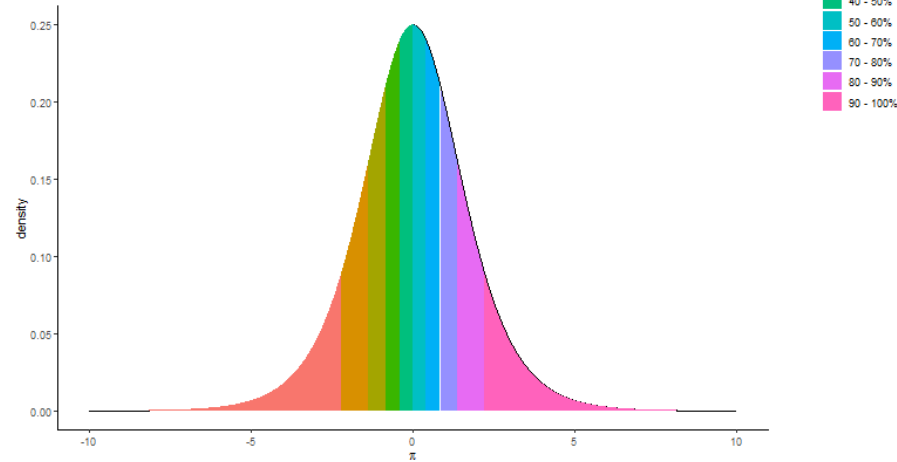
<https://github.com/stan-dev/stan/wiki/Prior-Choice-Recommendations>

# Prior on which scale?

➤ On inverse logit of parameter



➤ On parameter



## Prior on which scale?

- Jeffreys prior
  - It has the key feature that it is **invariant** under a change of coordinates for the parameter vector
  - Sometimes improper
- Whole field of study, reference priors for Objective Bayesian Inference
- Add other knowledge is sometimes considered Subjective Bayesian inference



# Reflection / Discussion

## Questions?





# Priors: Does it always matter?





# Sensitivity Analysis

- Common to check what would have happened with another prior?
- How influential are the priors?
- Posterior shrinkage



# What Took Them So Long? Explaining PhD Delays among Doctoral Candidates

Rens van de Schoot<sup>1,2\*</sup>, Mara A. Yerkes<sup>3,4</sup>, Jolien M. Mouw<sup>5</sup>, Hans Sonneveld<sup>6,7</sup>

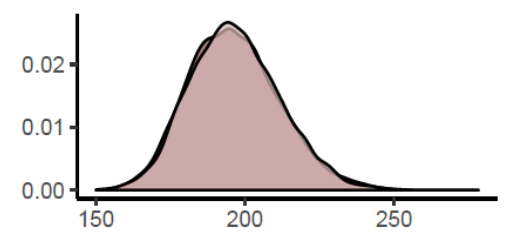
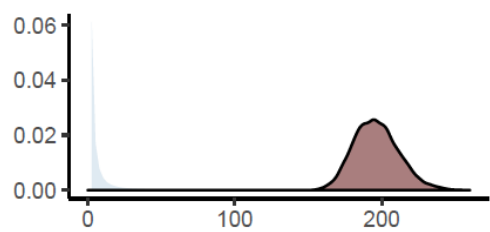
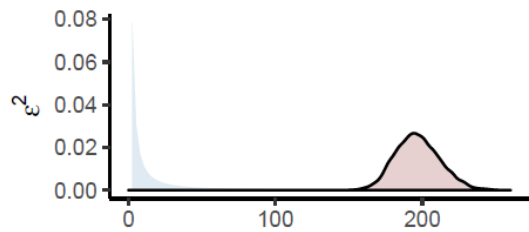
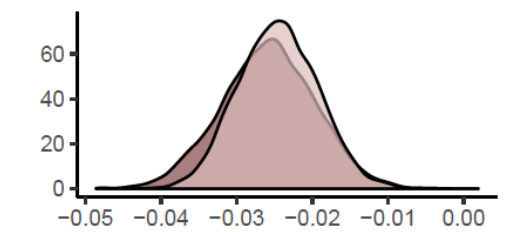
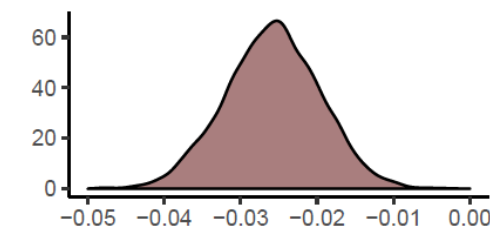
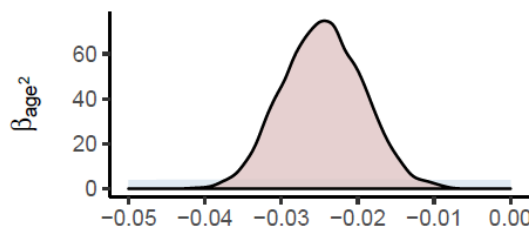
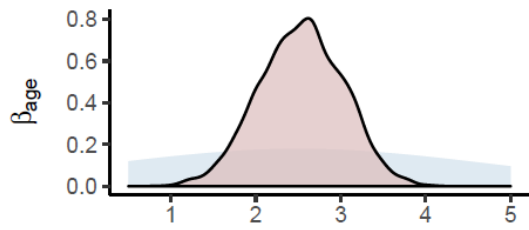
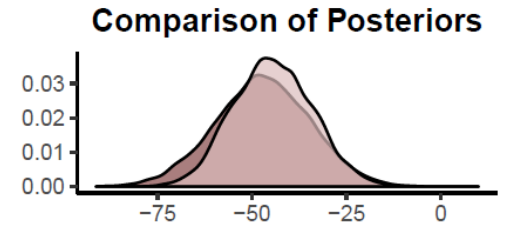
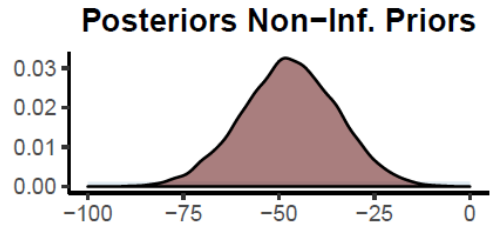
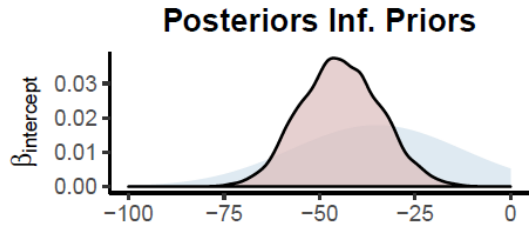
**1** Department of Methods and Statistics, Utrecht University, Utrecht, The Netherlands, **2** Optentia Research Focus Area, North-West University, Vanderbijlpark, South Africa, **3** Institute for Social Science Research, University of Queensland, Brisbane, Australia, **4** Erasmus University Rotterdam, Rotterdam, The Netherlands, **5** Education and Child Studies, Faculty of Social and Behavioural Sciences, Leiden University, Leiden, The Netherlands, **6** Netherlands Centre for Graduate and Research Schools, Utrecht, The Netherlands, **7** Tilburg Law School, Tilburg University, Tilburg, The Netherlands

## Abstract

A delay in PhD completion, while likely undesirable for PhD candidates, can also be detrimental to universities if and when PhD delay leads to attrition/termination. Termination of the PhD trajectory can lead to individual stress, a loss of valuable time and resources invested in the candidate and can also mean a loss of competitive advantage. Using data from two

- 333 PhD recipients in The Netherlands
- how long it had taken them to finish their PhD thesis
  - => 59.8 months
- difference between planned and actual project time in months
  - =>  $M = 9.97$ ,  $min / max = -31/91$ ,  $SD = 14.43$
- assume we are interested in the question whether age ( $M=31.68$ ,  $min/max=26/69$ ) of the PhD recipients is related to delay in their project.
- assume we expect this relation to be non-linear.





Density  
Posterior Prior

Density  
Posterior Prior

Density  
Inf. Prior N-Inf. Prior



de Klerk, M., Veen, D., Wijnen, F., & de Bree, E. (2019). A step forward: Bayesian hierarchical modelling as a tool in assessment of individual discrimination performance. *Infant Behavior and Development*, 57, 101345.





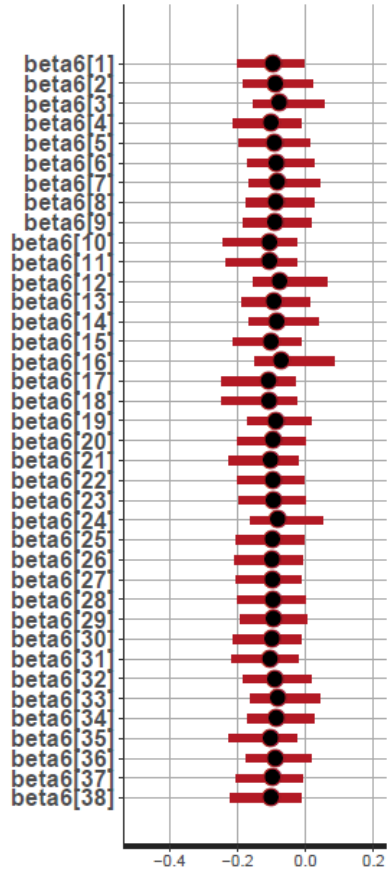
Remember from Tuesday?



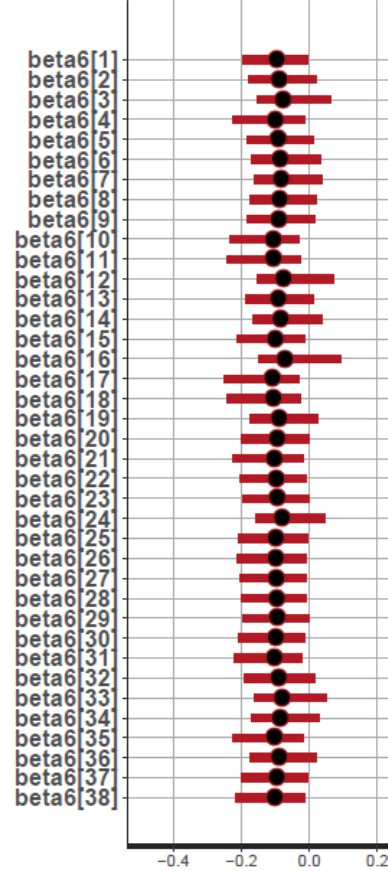
**Universiteit Utrecht**



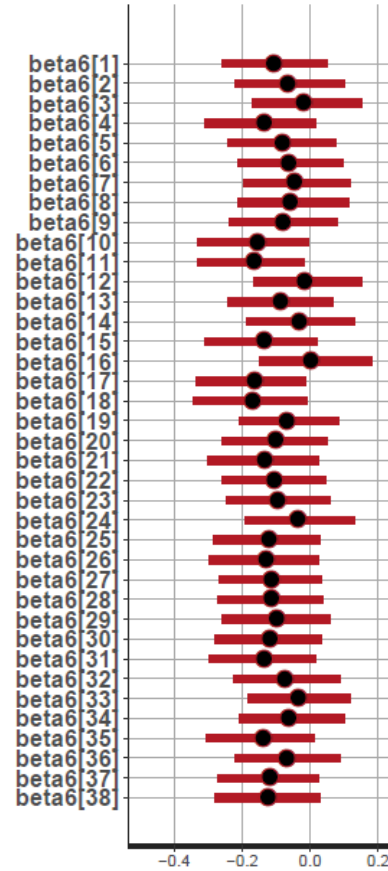
Cauchy(0,2.5)



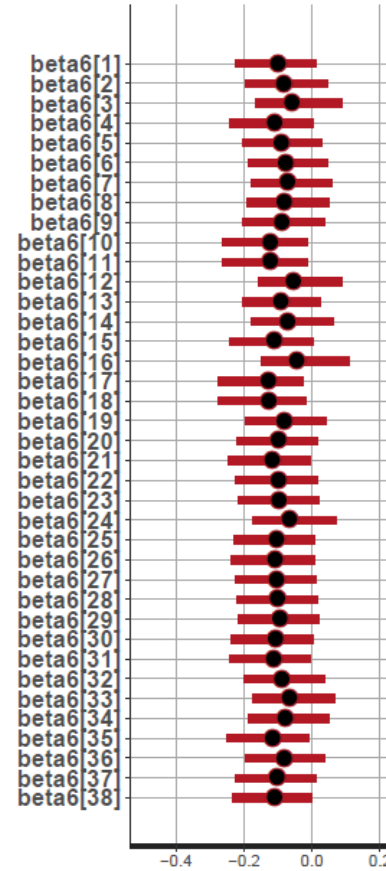
Uniform



IG(.5,.5)

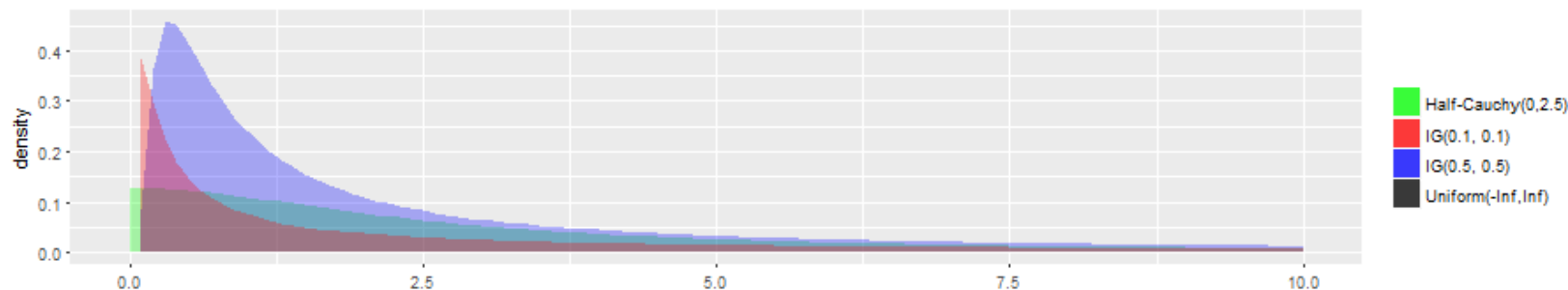


IG(.1,.1)

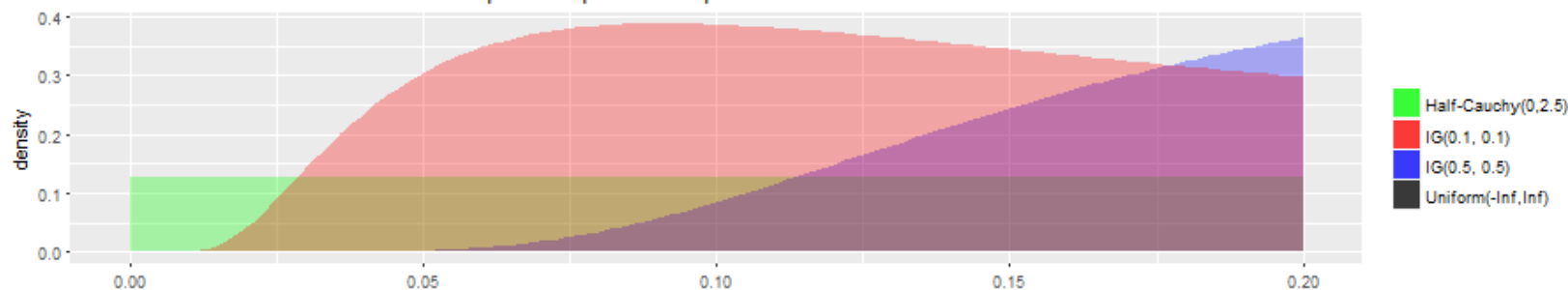




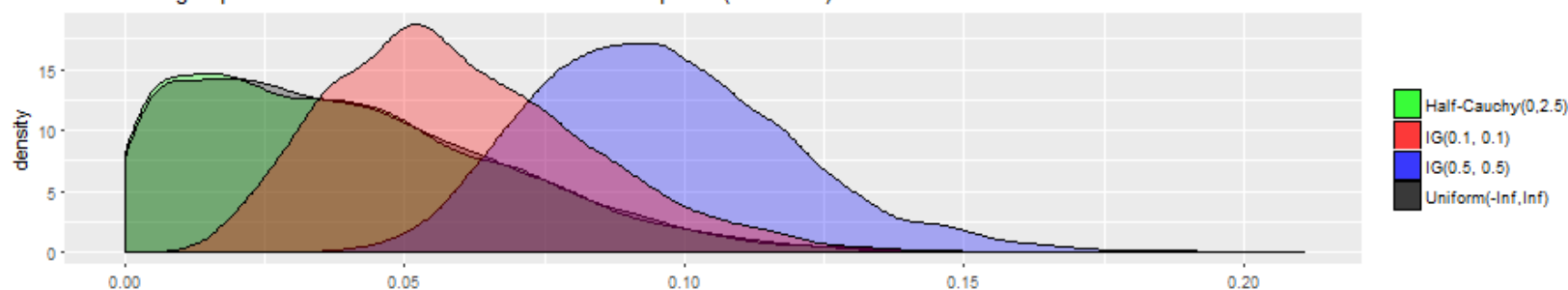
Priors on standard deviations



Priors on standard deviations zoomed at posterior parameter space



Posteriors group level variance condition effect for different priors (6 months)







Priors: Does it always matter?

Sometimes...



# Priors: Experts and Literature



**Universiteit Utrecht**

## Types of priors – where can they come from

- Results of a previous publication as prior specification
- An expert, or a panel of experts
- Meta-analysis
- A pilot study
- Data-based priors can be derived based on a variety of methods including:
  - maximum likelihood
  - or sample statistics
  - Training data
  - Data splitting priors

Note that there are some arguments against using such “double-dipping” procedures where the sample data are used to derive priors and then used in estimation



## Types of priors – Let's take a closer look at

- **Results of a previous publication as prior specification**
- **An expert, or a panel of experts**



# 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?

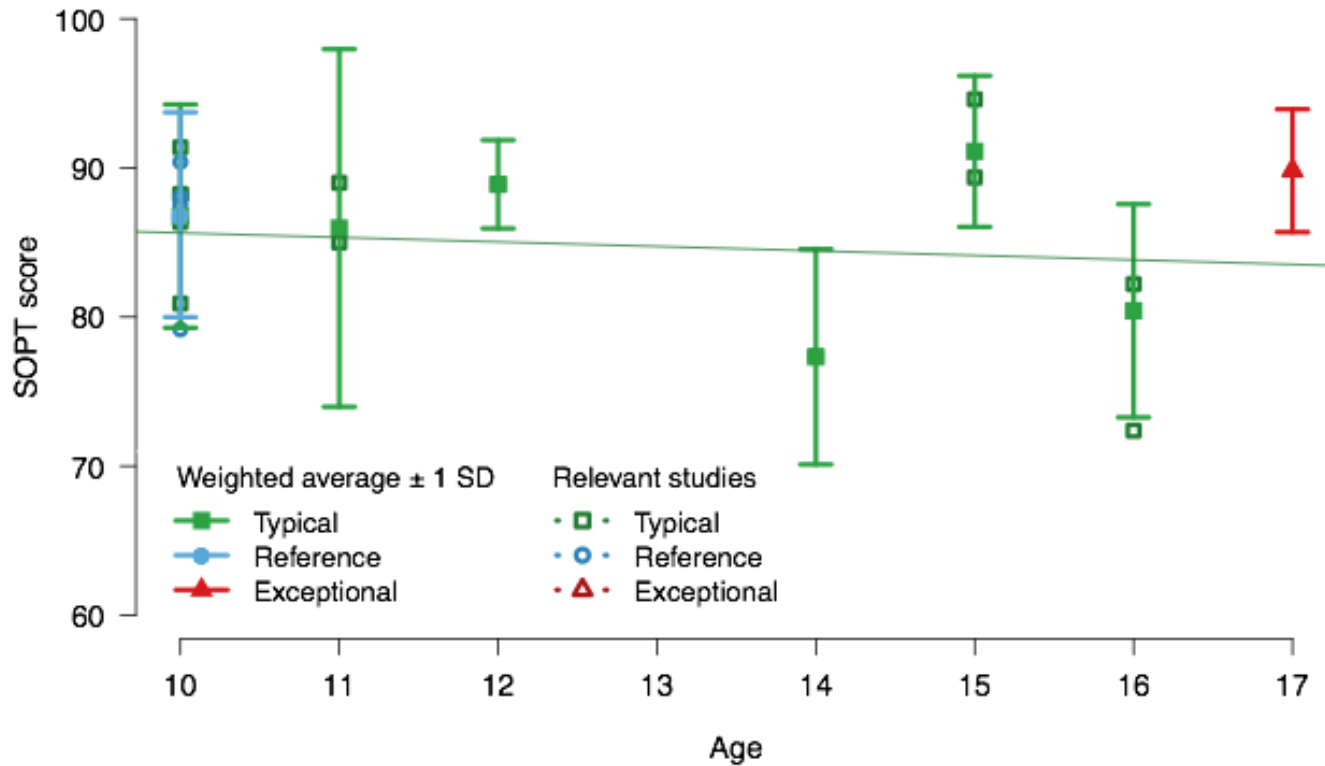


Beschrijving onderzoeksgroep	Representativiteit voor cluster 4 populatie (0-1)	Leeftijd	Verwacht % cannabis gebruikers	noot
US Children with ADHD (DSM-IV) and a mother with elevated depression level (ADHD and elevated mother depression are risk factors for CD)	.85	9.6	< 5%	
British typically developing children (no special educational needs) attending state primary schools	.01	10.1	< 1%	
Canadian adolescents that accepted a lab invitation	.01	12.5	< 1%	degen die blowen komen veel 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





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

 Susanne Schulz<sup>1\*</sup>,  Mariëlle Zondervan-Zwijenburg<sup>2</sup>,  Stefanie A. Nelemans<sup>1</sup>,  Duco Veen<sup>3,4</sup>,  Albertine J. Oldehinkel<sup>5</sup>,  Susan Branje<sup>1</sup> and  Wim Meeus<sup>1</sup>

<sup>1</sup> Youth and Family, Utrecht University, Utrecht, Netherlands

<sup>2</sup> Methodology and Statistics, Utrecht University, Utrecht, Netherlands

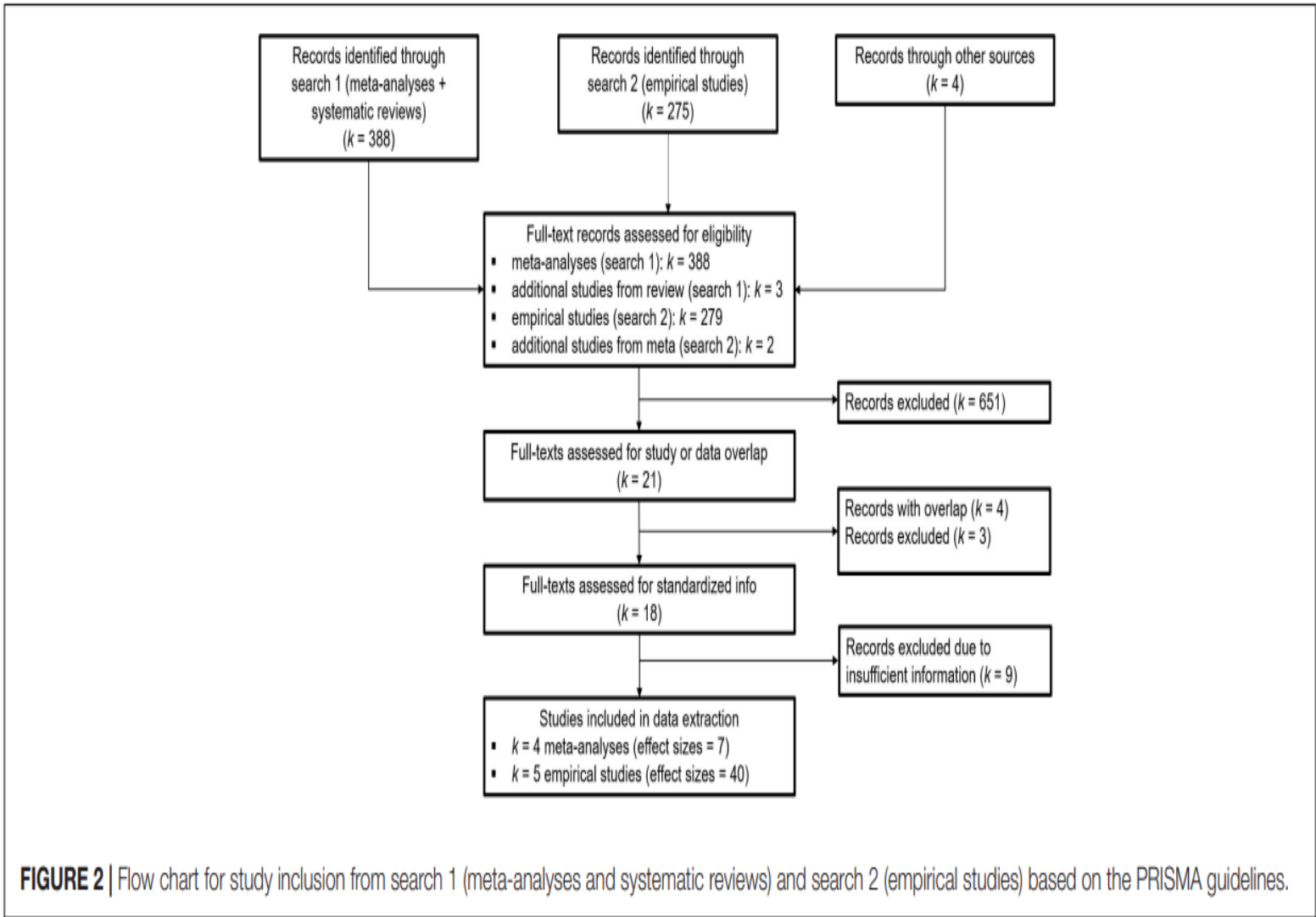
<sup>3</sup> Julius Global Health, Julius Center for Health Sciences and Primary Care, University Medical Center Utrecht, Utrecht, Netherlands

<sup>4</sup> Optentia Research Program, North-West University, Potchefstroom, South Africa

<sup>5</sup> 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.
<b>10 categories (standard 5)</b>	<b>100 (80)</b>	<b>Each study can score between 0 and 100 points (or between 0 and 80 points for T1 → T2 associations).</b>



**TABLE 1B** | Final scoring of all included studies.

Study	T1-T2	lag	cT1	obs	Age	M <sub>dep+anx (or anx)</sub>	A <sub>dep+anx (or anx)</sub>	cov <sub>s</sub>	cov <sub>i</sub>	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<sub>s</sub>, controlling for parental symptoms; cov<sub>i</sub>, 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>	$N(-0.06, 0.0704)^{0.3750}$	$N(-0.06, 0.19)$	$N(-0.06, 0.19)$	



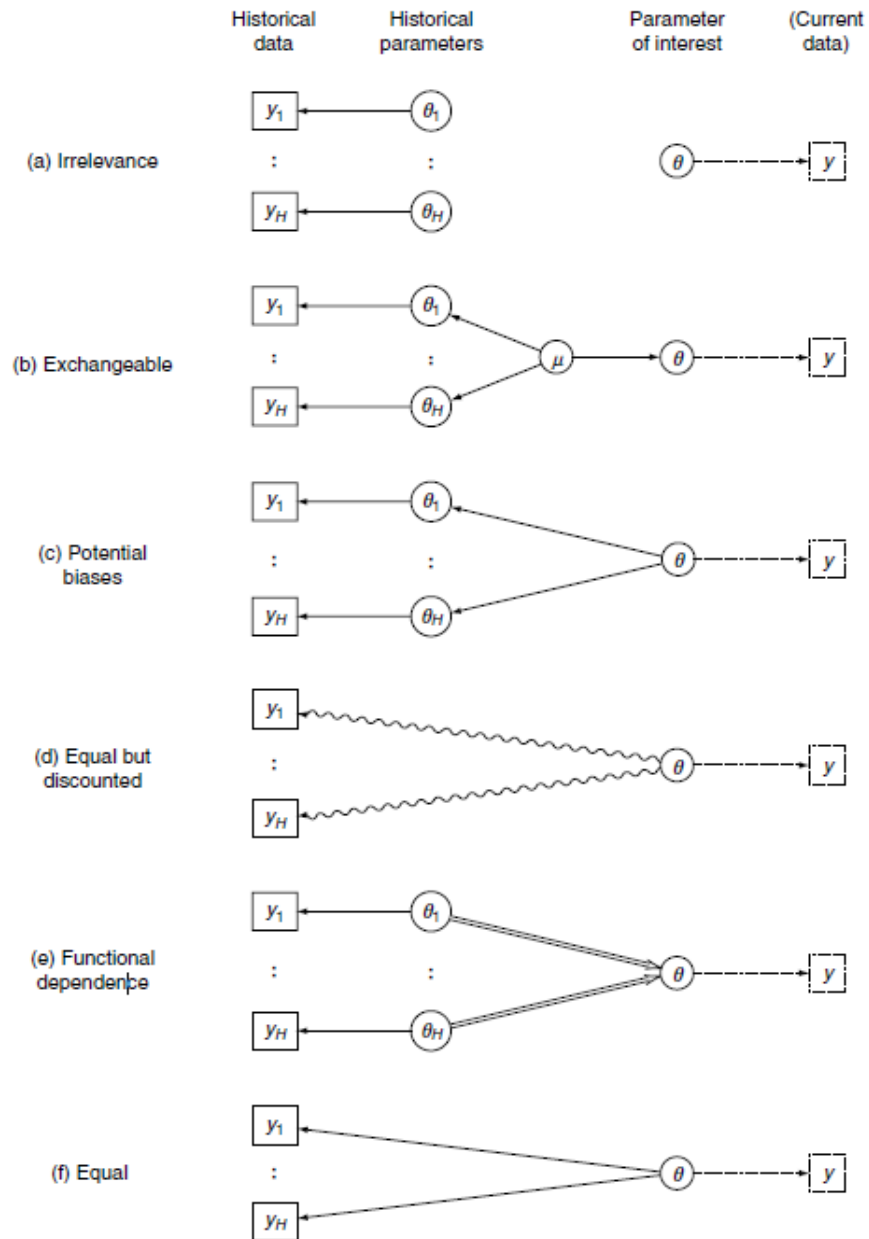


Figure 5.4 Different assumptions relating parameters underlying historical data to the parameter of current interest: single arrows represent a distribution, double arrows represent logical functions, and wavy arrows represent discounting.

Spiegelhalter, D. J.,  
Abrams, K. R., &  
Myles, J. P.  
(2004). *Bayesian  
approaches to clinical  
trials and health-care  
evaluation* (Vol. 13).  
John Wiley & Sons.



# Priors based on expert knowledge



# 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





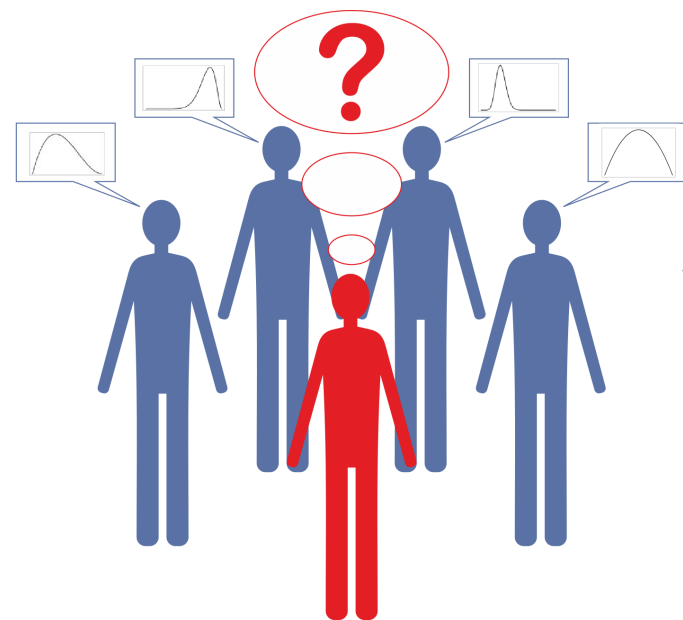


## Expert elicitation - What is it?

*“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





## Reasons for elicitation of expert judgement

- Experts offer unique information
- 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?

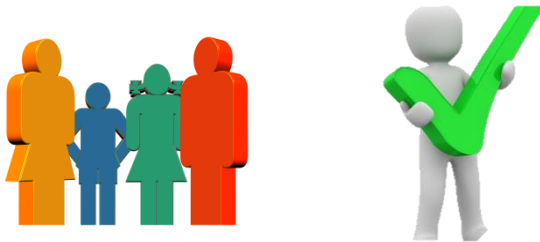




# Expert elicitation – What to do?

- Direct vs. Indirect
  - quantile elicitation
  - predicting data

- Group vs. Individual





# Uncertain Judgements

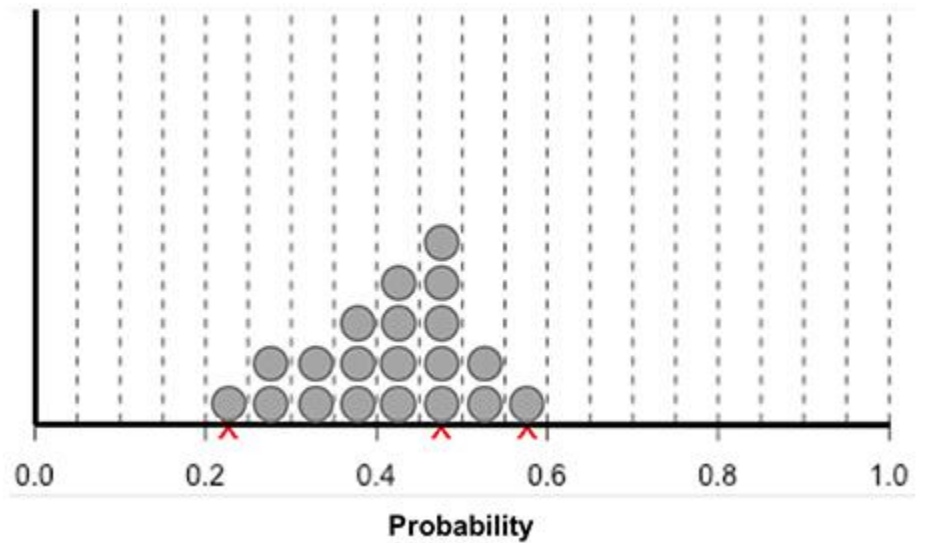
Eliciting Experts' Probabilities



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



# Correlation Cognitive potential and Academic performance

- Four behavioral scientists give judgments
- Correlation between cognitive potential and academic performance
- Two separate populations (all problematic)
  - youth with autism spectrum disorder (ASD)
  - youth with diagnoses other than ASD.

< 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<sup>1\*</sup>, Wenneke van de Schoot-Hubbeek<sup>1</sup>, Kimberley Lek<sup>1</sup>, Herbert Hoijtink<sup>1,2</sup> and Rens van de Schoot<sup>1,3</sup>

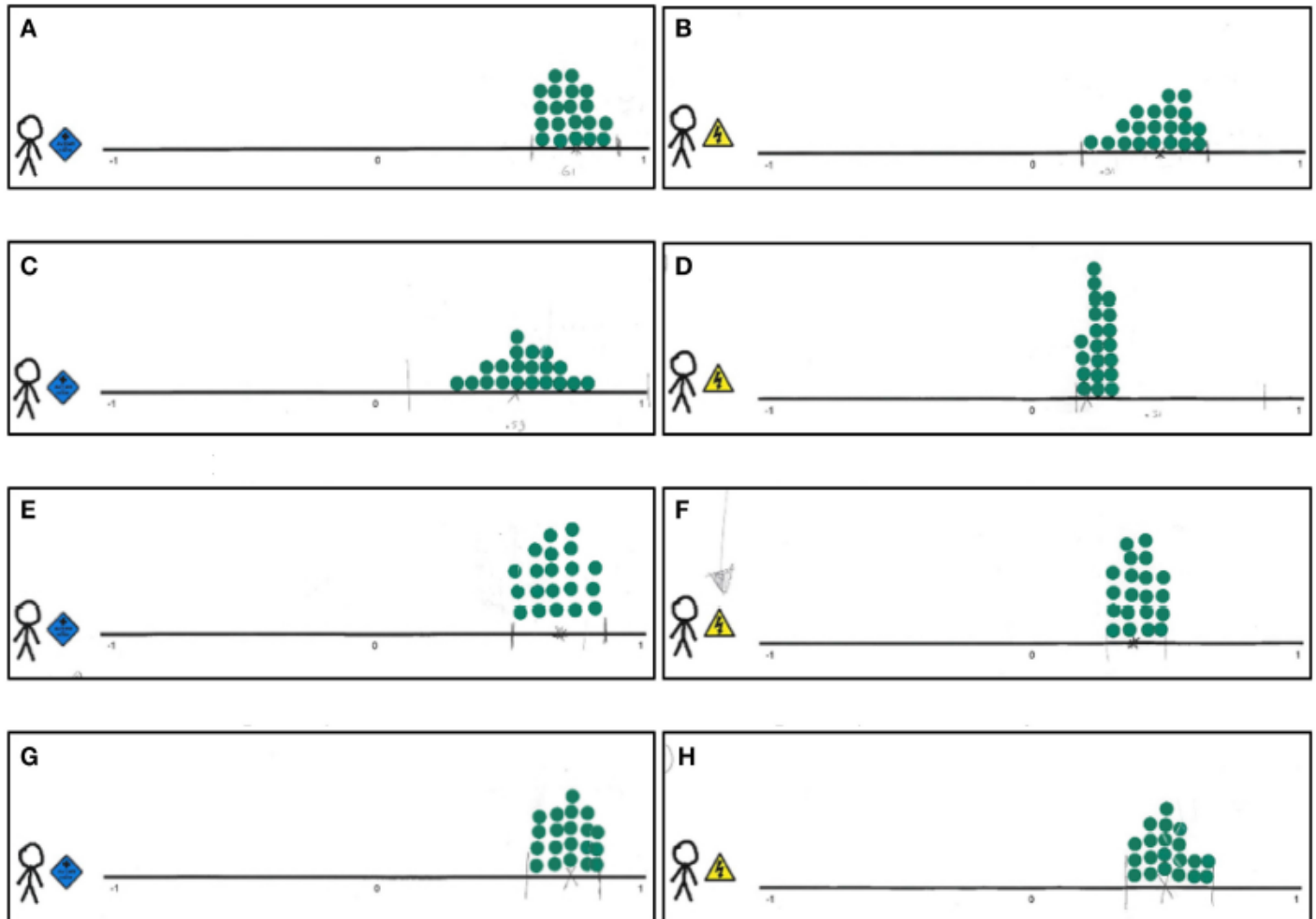
<sup>1</sup>Department of Methods and Statistics, Utrecht University, Utrecht, Netherlands

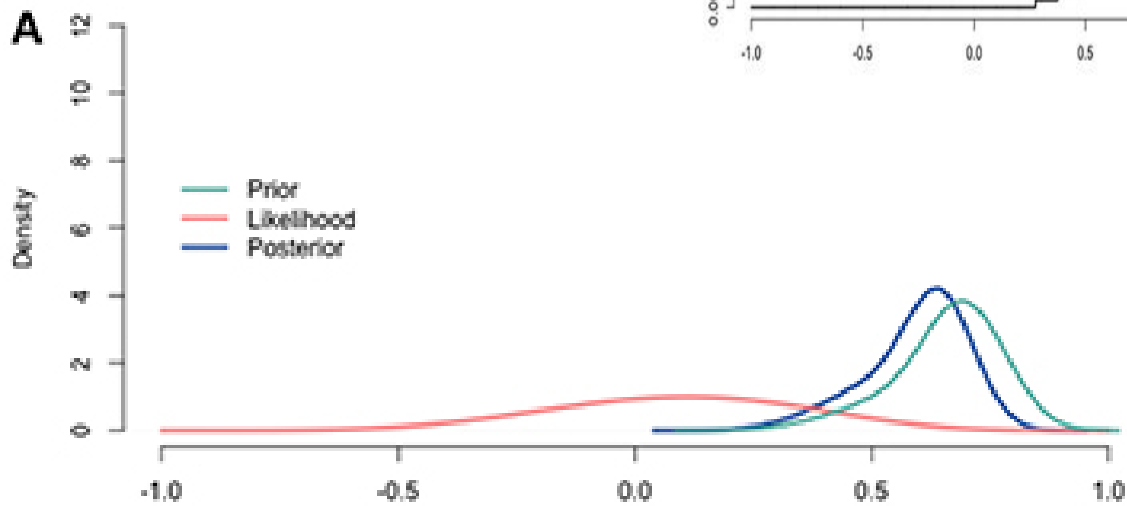
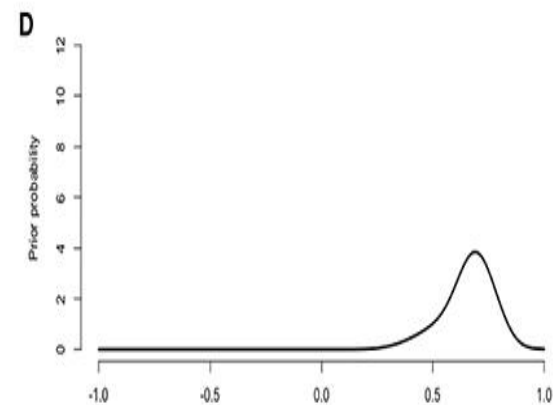
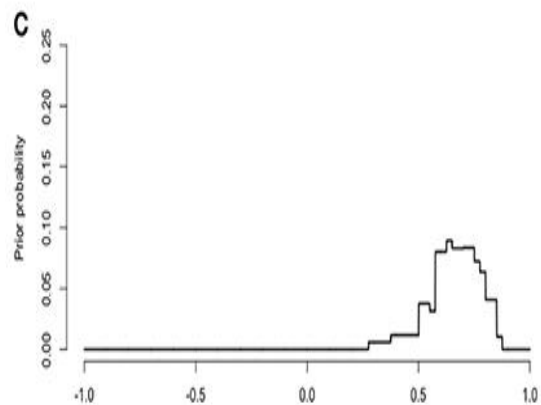
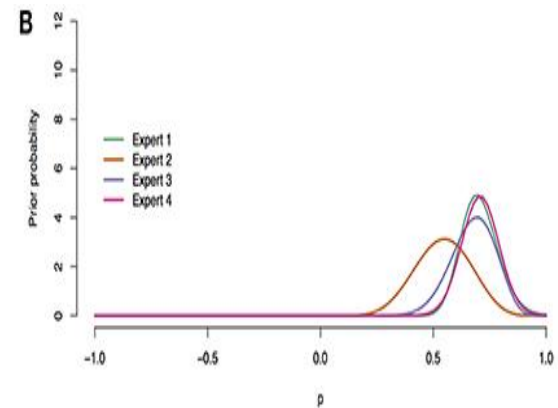
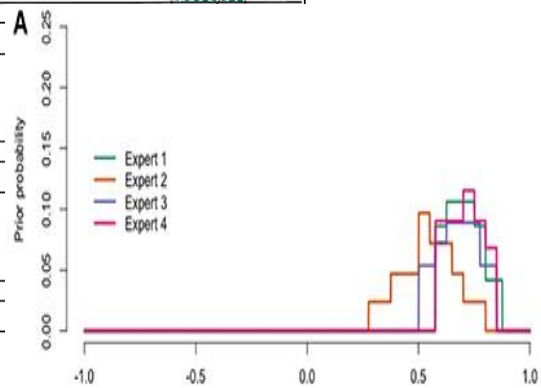
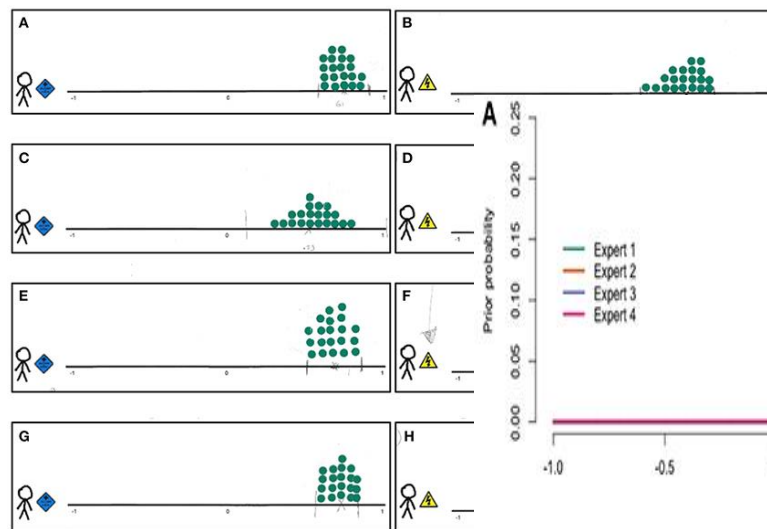
<sup>2</sup>Schreuder College Location Villerneuvestraat, Horizon Jeugdzorg en Childerwijs (Horizon Youth Care and Education), Rotterdam, Netherlands

<sup>3</sup>CITO 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 of a trial roulette quest: a concordance probability elicitation procedure in terms of means that the elicited distribution







# Improving elicitation quality

- Avoid triggering of heuristics and biases
- Employ face-to-face elicitation
- Training experts and facilitators



# 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)



# Expert elicitation – Digitizing for feedback

- What is out there? – Systematic review (2016)
  - MATCH (Morris et al., 2014)
  - Based on SHELF (Oakley, 2016)
  - Single use elicitation programs
  
- What do we think works well with our experts?
  - Direct – indirect?





# Expert elicitation – Digitizing for feedback

- Experts had difficulty with concept of hyperparameters with uncertainty
- Cut elicitation into smaller steps
- Combine direct and indirect



# Expert elicitation – Five-step method

- 1) Elicit location parameter using trial roulette – direct elicitation
- 2) Provide feedback
- 3) Elicit scale and shape parameters
- 4) Provide feedback
- 5) Use elicited distribution

# Five-step method – Steps 1 & 2

Five-Step Method Elicitation App

Disclaimer

Agree and continue to shiny App ▾

ID

Duco

Sales results

Sales 1 ▾

Number of sales

18

Minimum sales value

2

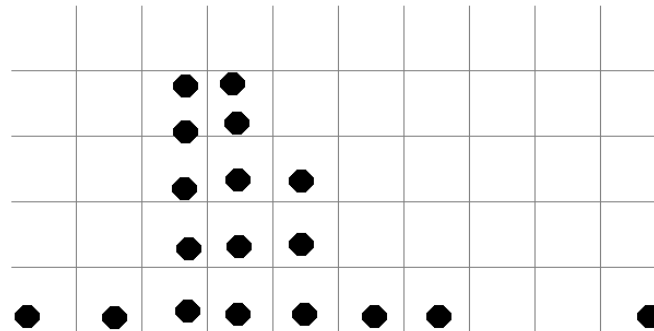
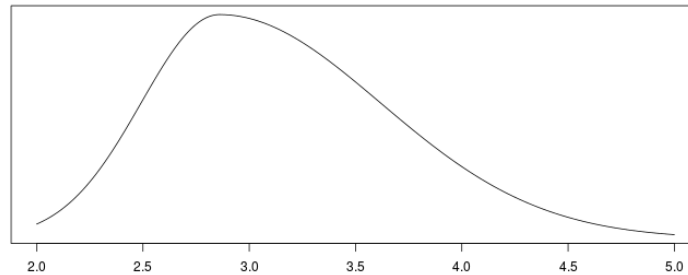
Maximum sales value

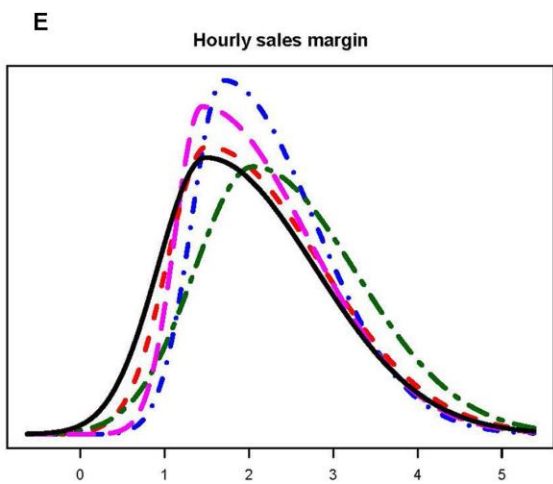
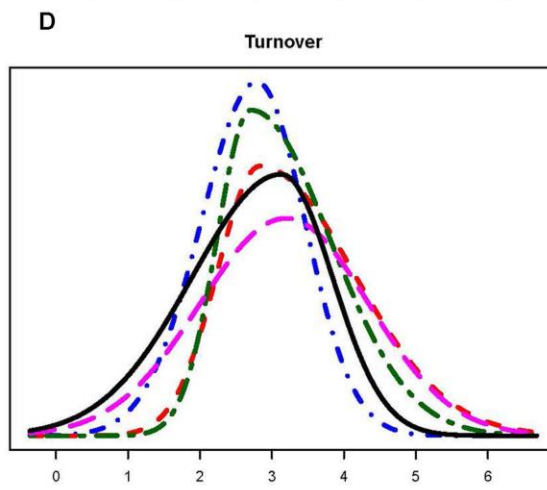
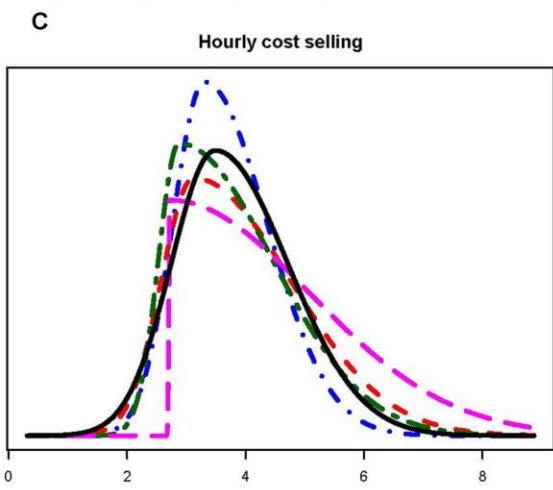
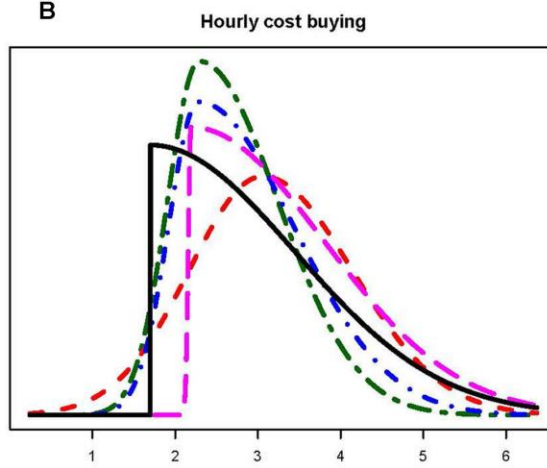
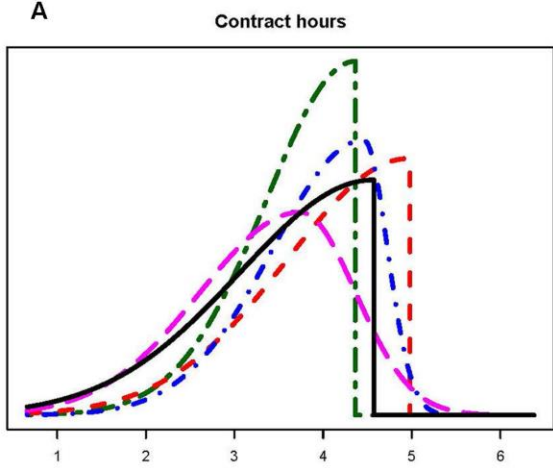
5

Reset drawing

Undo

Submit





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



# Five-step method – Steps 3 & 4

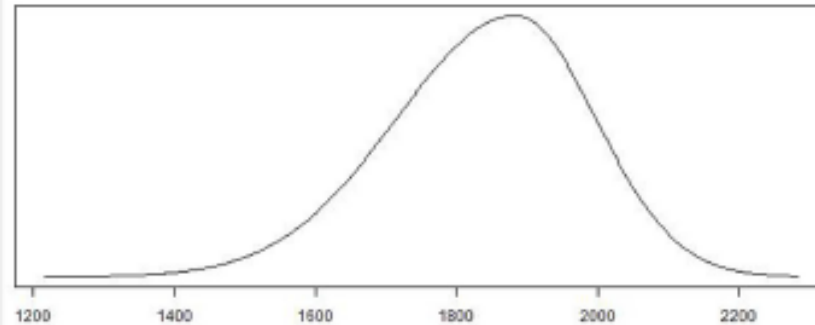
ID

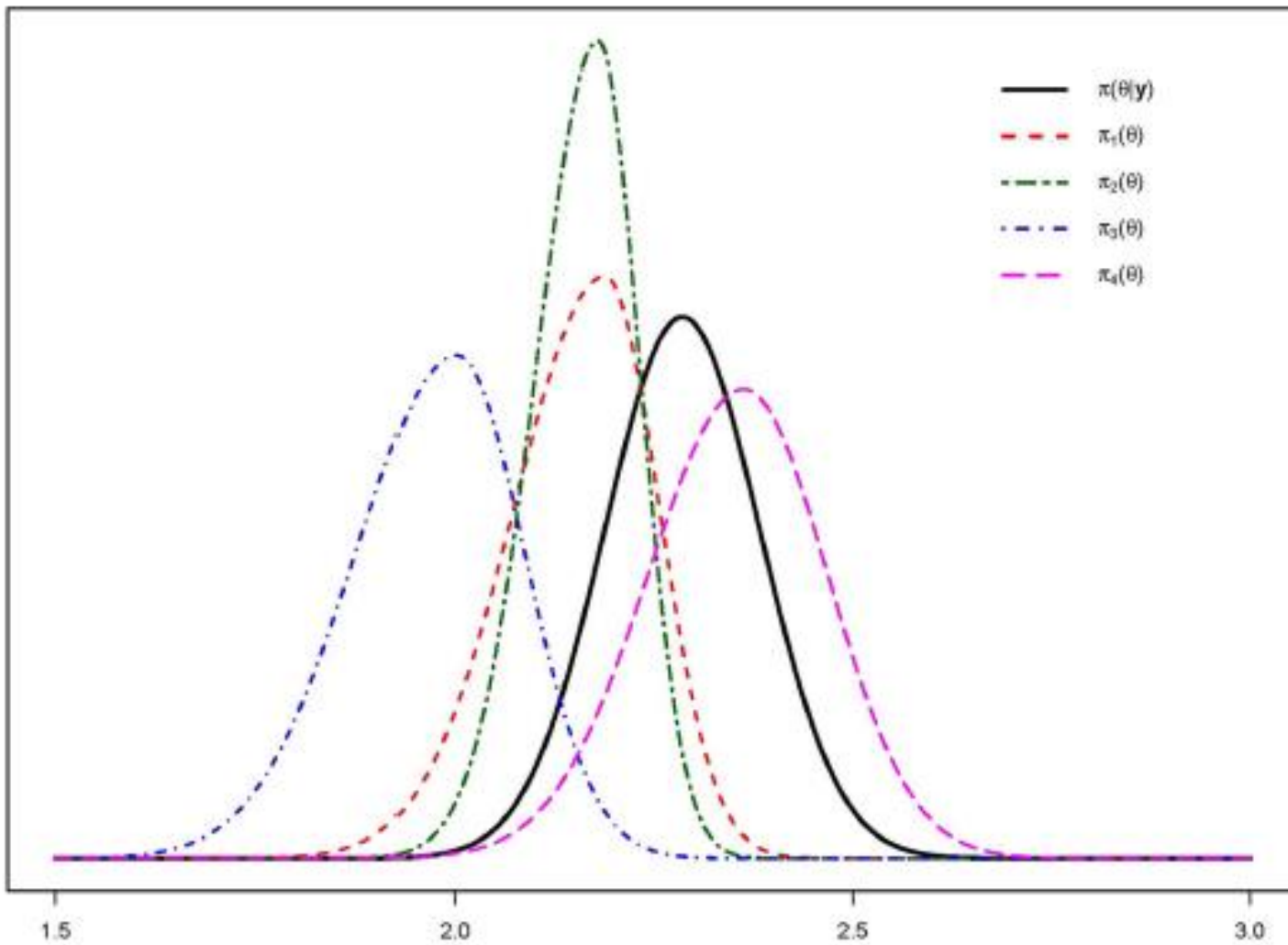
Sales results

Total

Reasonable lowerbound

Reasonable upperbound





# Five-step method – Steps 5

- Use elicited distributions





## Five-step method – Steps 5

- Use elicited distributions

**But what about quality control?**





# General reflections - Quality

- calibration questions can be 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 - Quality

- 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





## 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)



## EXERCISE FIVE-STEP METHOD

[https://utrechtuniversity.github.io/BayesianEstimation/content/friday/exercise\\_elicit\\_expert\\_judgement.html](https://utrechtuniversity.github.io/BayesianEstimation/content/friday/exercise_elicit_expert_judgement.html)

<https://utrecht-university.shinyapps.io/elicitation/>

<https://github.com/VeenDuco/Five-Step-Method-Shinyapp>



# Contrasting experts' beliefs and data



**Universiteit Utrecht**



## 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?





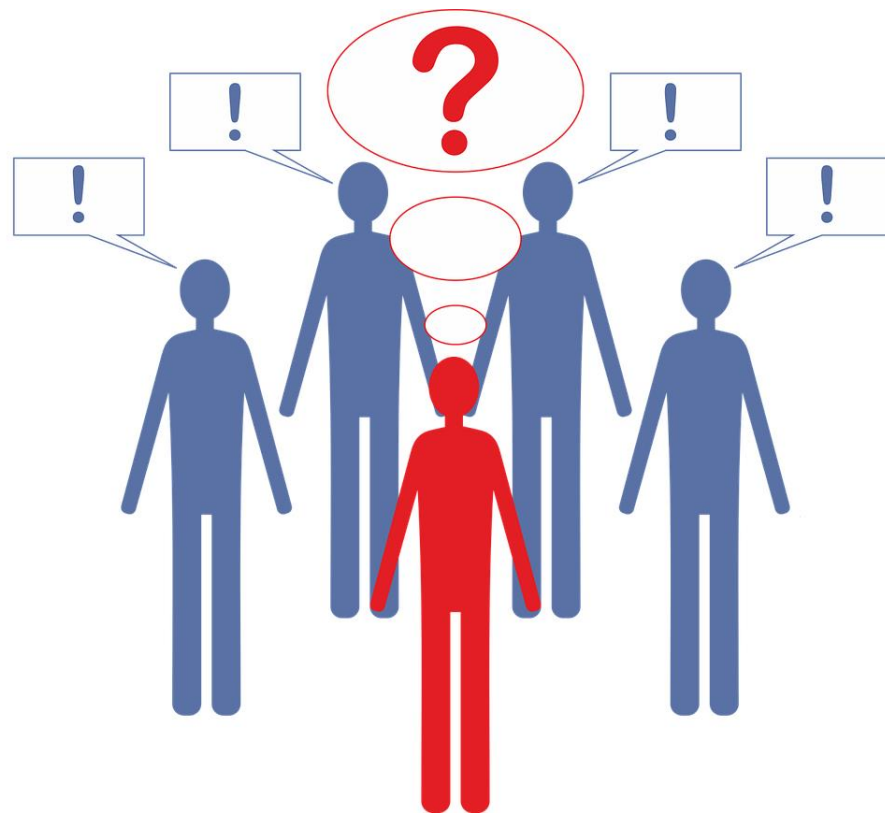
## Expert elicitation – Quality control

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



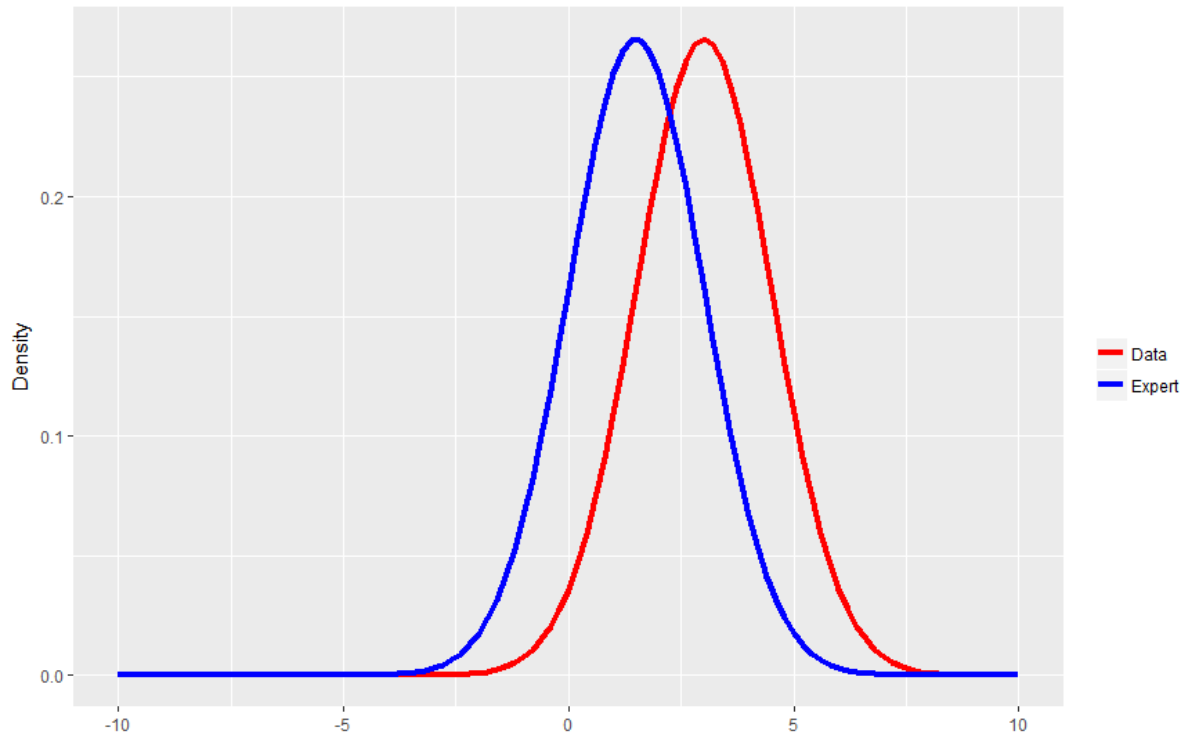


# Quality Control

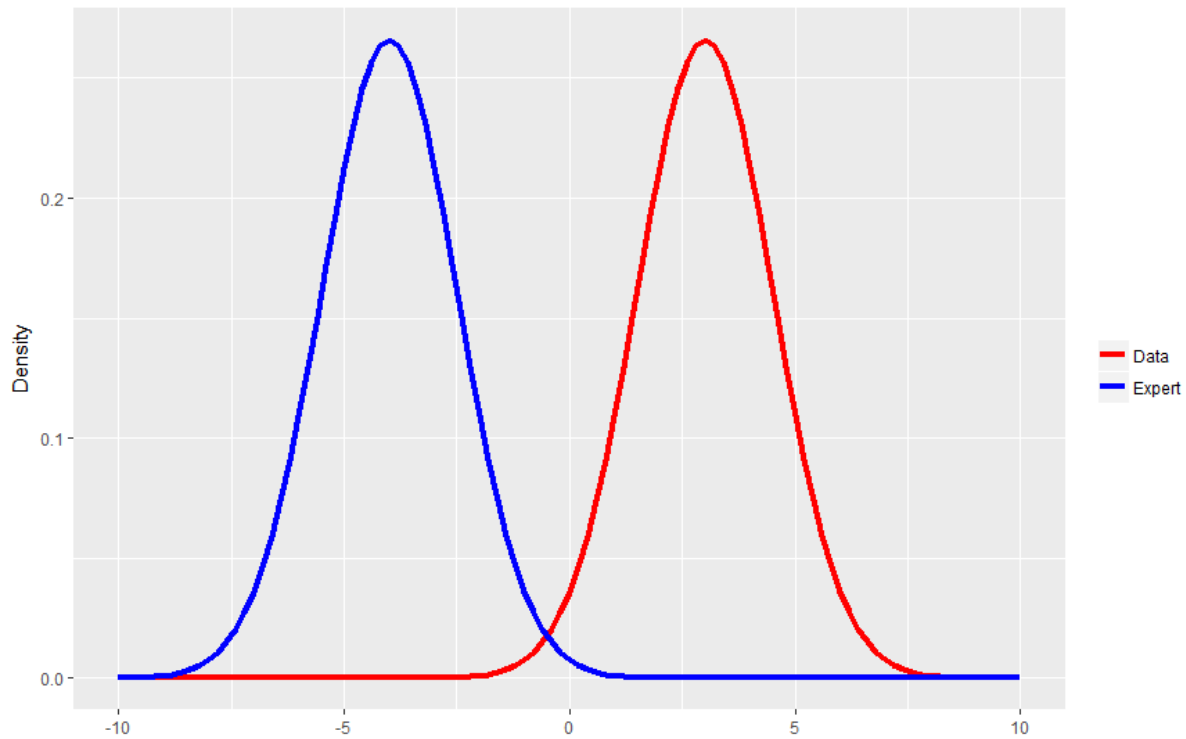




# Prior-data Agreement



# Prior-data Disagreement





## 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



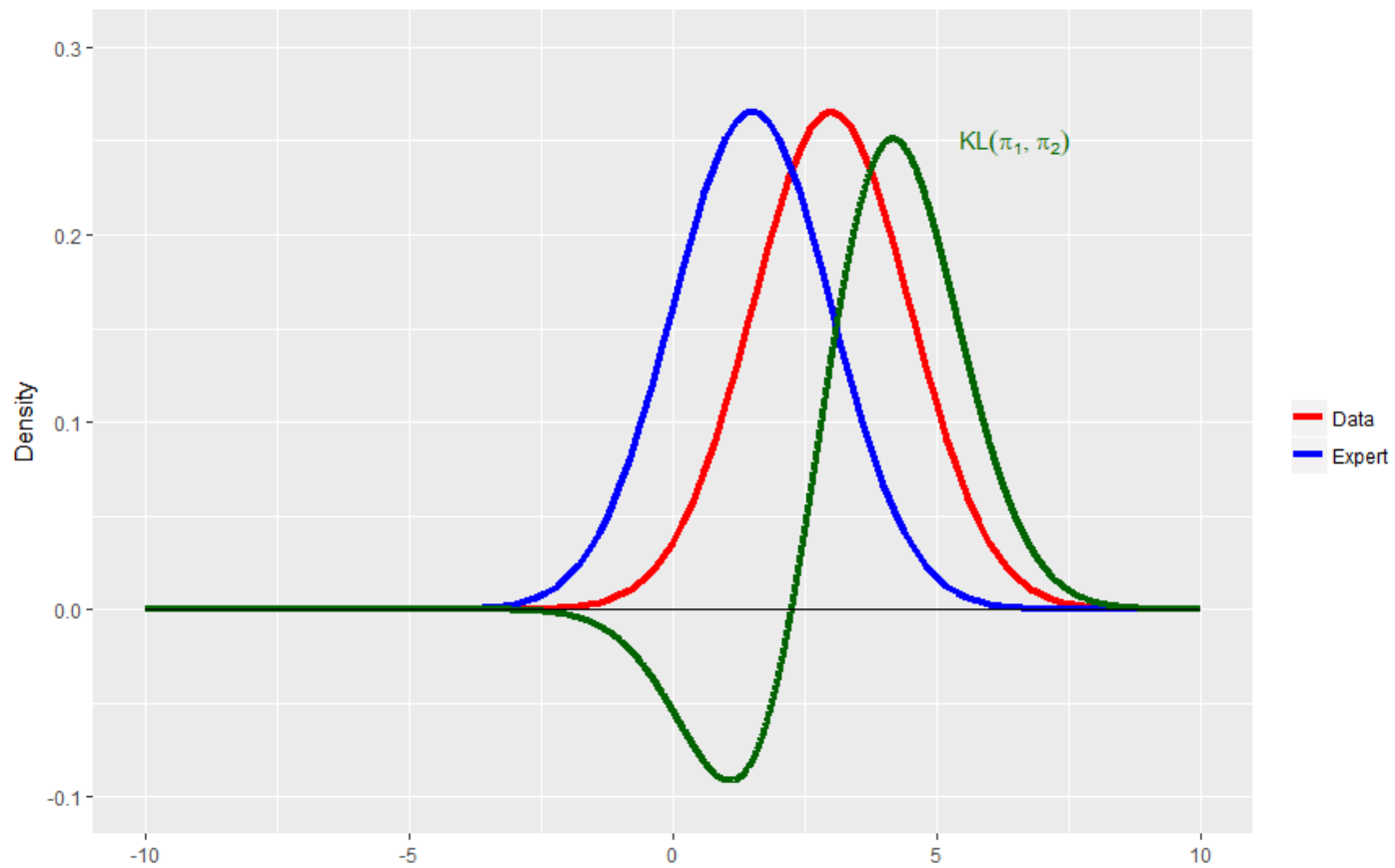
# Data Agreement Criterion<sup>1</sup>

- Ratio of two Kullback-Leibler divergences<sup>2</sup>

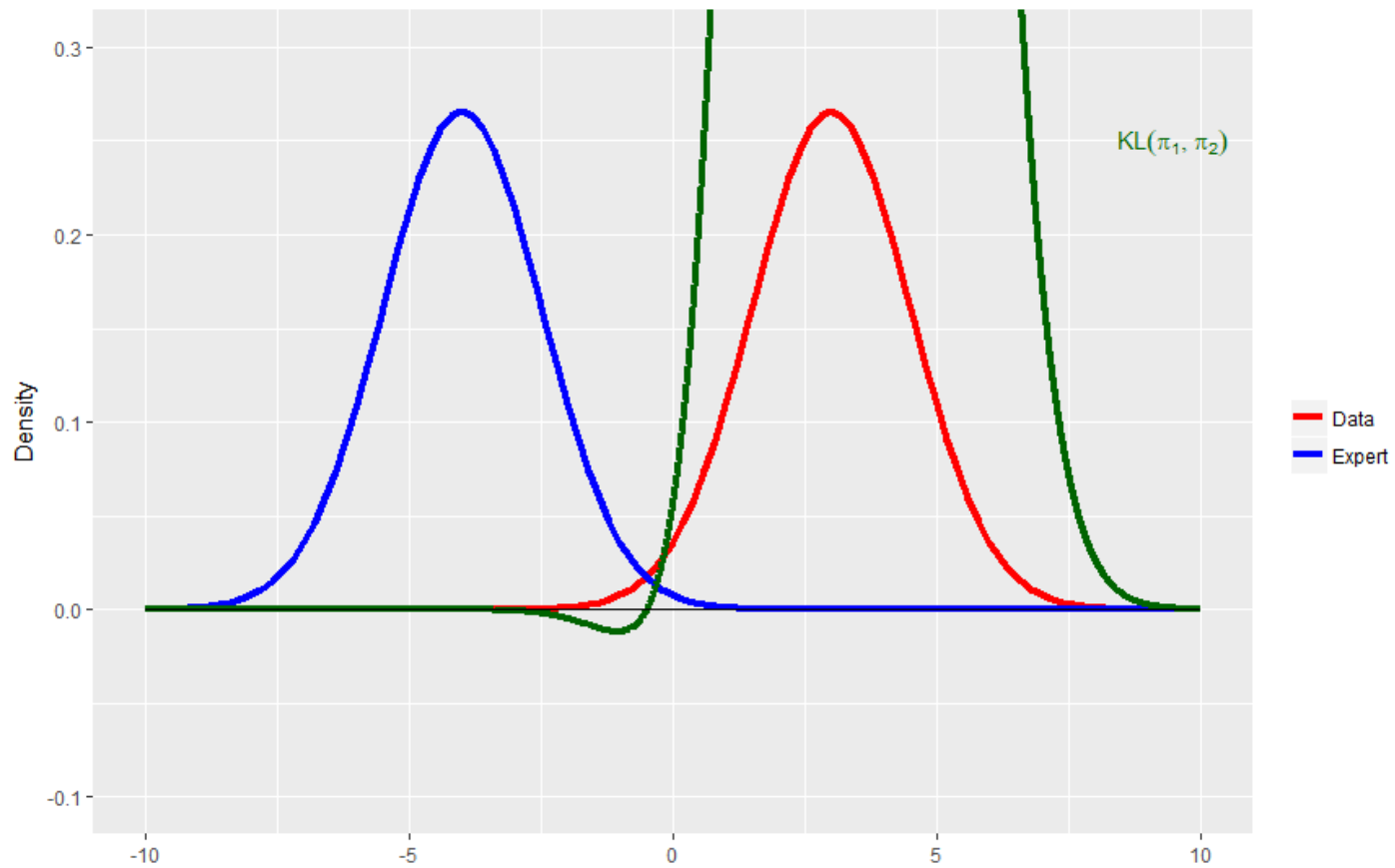
$$KL(\pi_1 || \pi_2) = \int_{\Theta} \pi_1(\theta) \log \frac{\pi_1(\theta)}{\pi_2(\theta)} d\theta$$



# Kullback-Leibler Divergence



# Kullback-Leibler Divergence



# Data Agreement Criterion

$$\text{DAC} = \frac{KL[\pi^J(\theta | \mathbf{y}) || \pi(\theta)]}{KL[\pi^J(\theta | \mathbf{y}) || \pi^J(\theta)]}$$





# Data Agreement Criterion

$$\text{DAC} = \frac{KL[\pi^J(\theta | \mathbf{y}) || \pi(\theta)]}{KL[\pi^J(\theta | \mathbf{y}) || \pi^J(\theta)]}$$



# Data Agreement Criterion

$$\text{DAC} = \frac{KL[\pi^J(\theta | \mathbf{y}) || \pi(\theta)]}{KL[\pi^J(\theta | \mathbf{y}) || \pi^J(\theta)]}$$





## 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

- Ratio larger than 1
  - prior-data conflict
  - The candidate prior resembles the data less 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

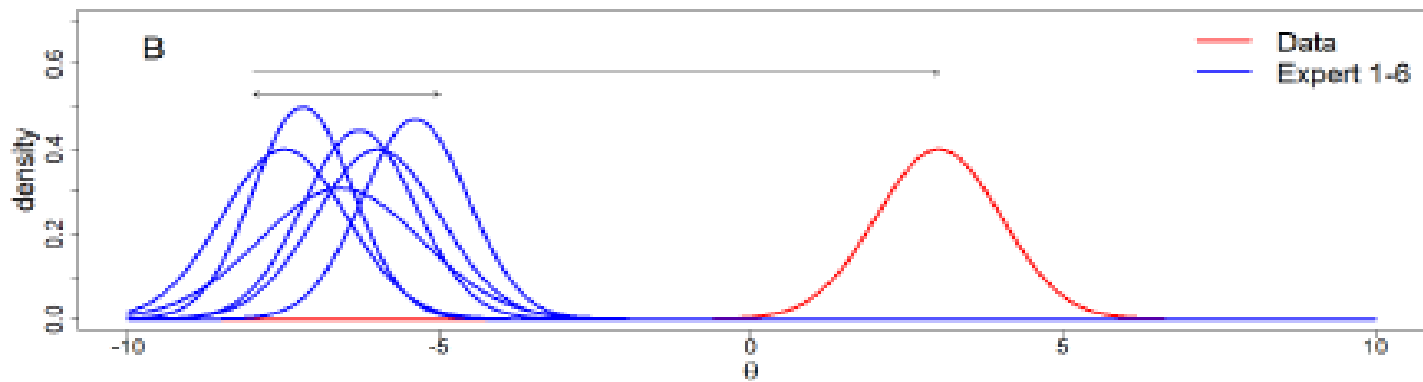
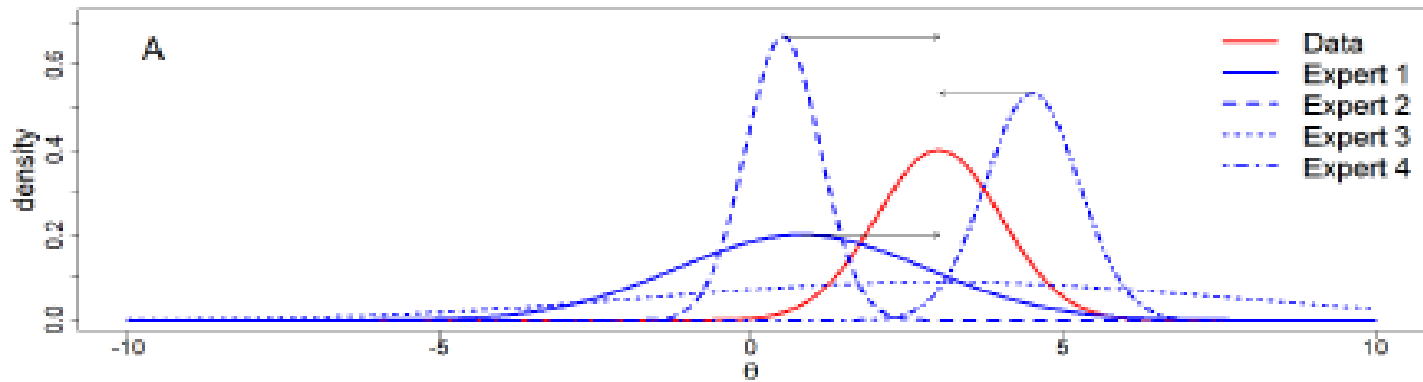
- When we have multiple experts
  - We can compare their KL divergences directly or all to the benchmark
  - Always look at data visually too







# 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

Sounds familiar from the exercise?



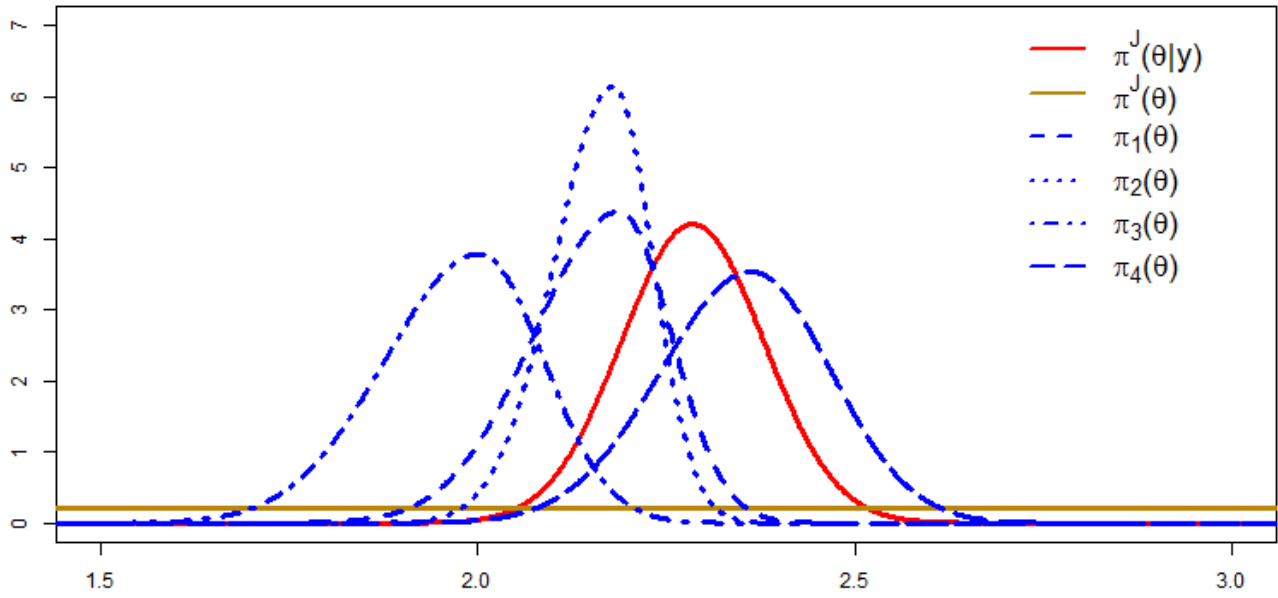
**Universiteit Utrecht**



## Experts in a financial institution

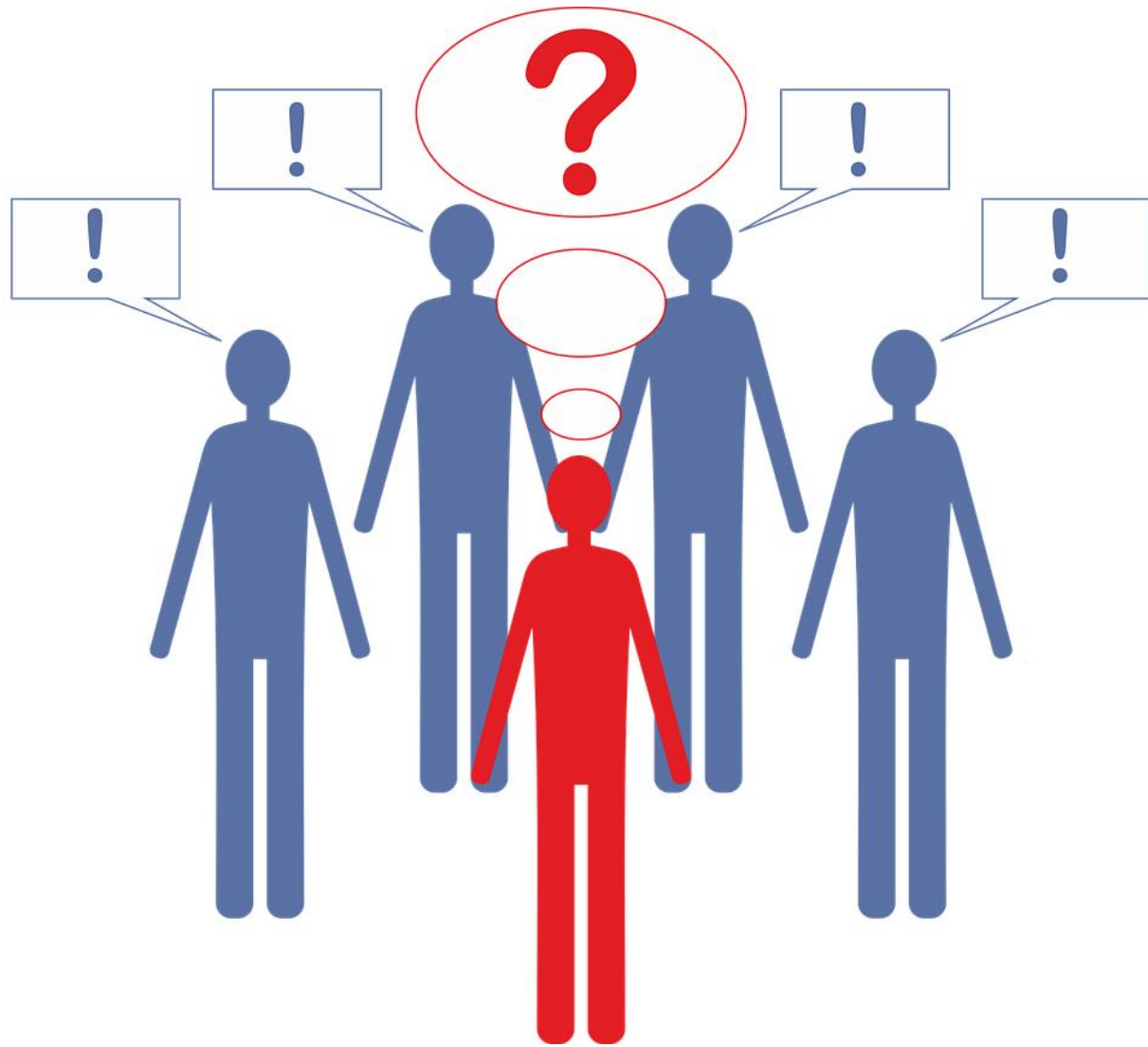
- We compared their prior beliefs to the actual realization of that quarter
  - Benchmark used was uniform prior ranging from 0 (no turnover) up to a large value that could not reasonably be attained.



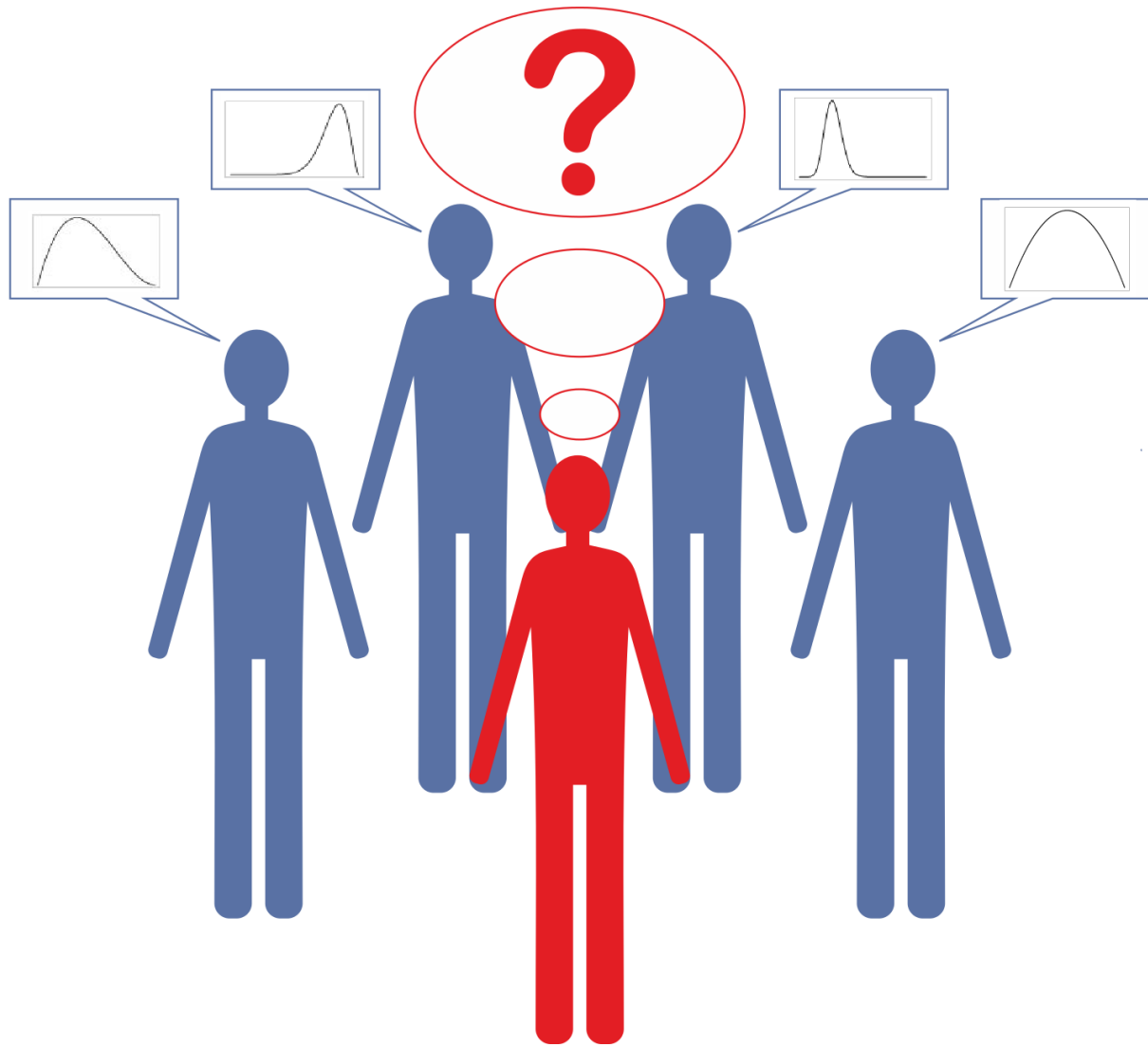


	KL divergence	DAC <sub>d</sub>	Ranking
Expert 1	<b>1.43</b>	<b>0.56</b>	<b>2</b>
Expert 2	<b>2.86</b>	<b>1.12</b>	<b>3</b>
Expert 3	<b>5.76</b>	<b>2.26</b>	<b>4</b>
Expert 4	<b>0.19</b>	<b>0.07</b>	<b>1</b>
Benchmark	<b>2.55</b>	-	-



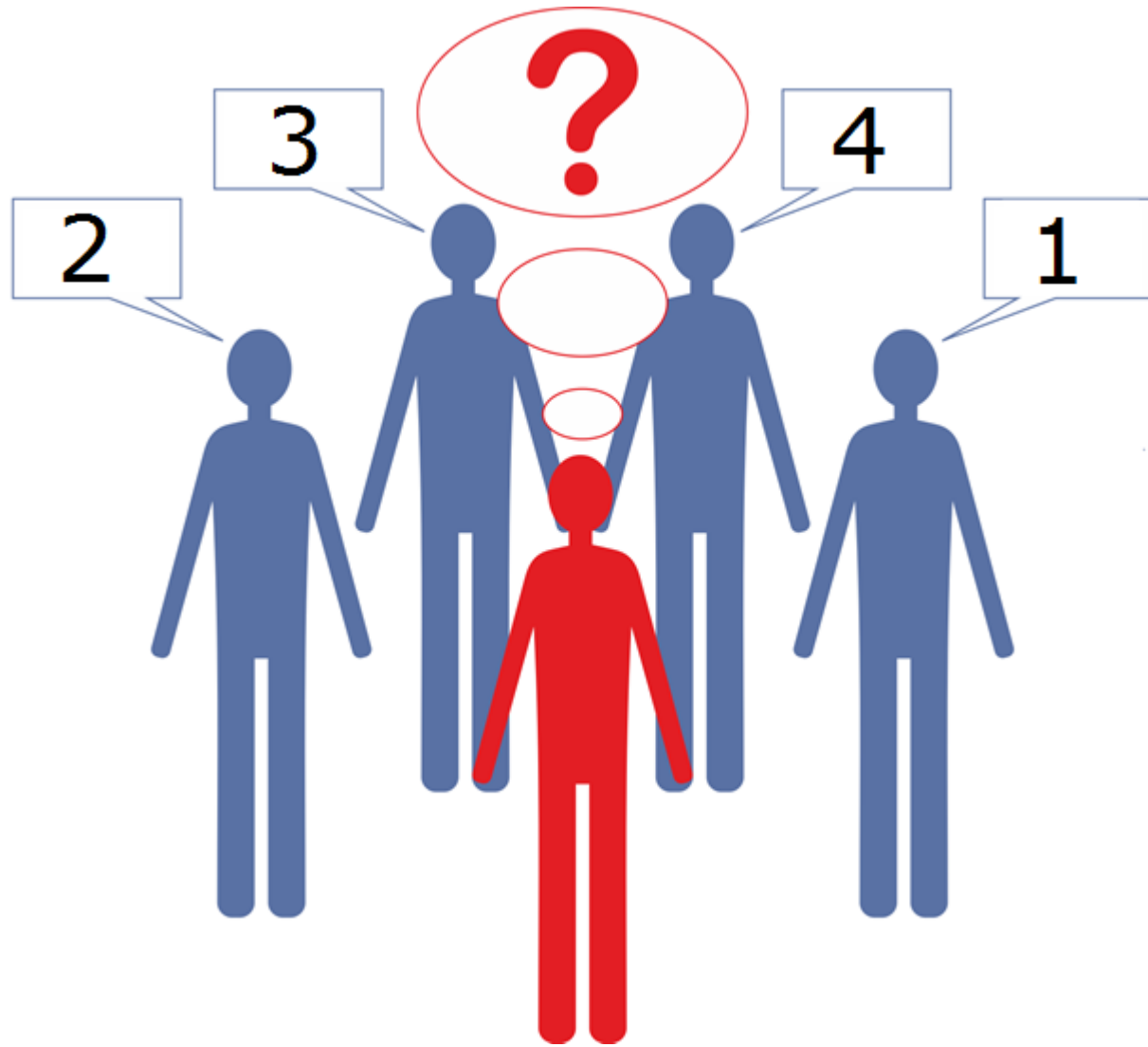


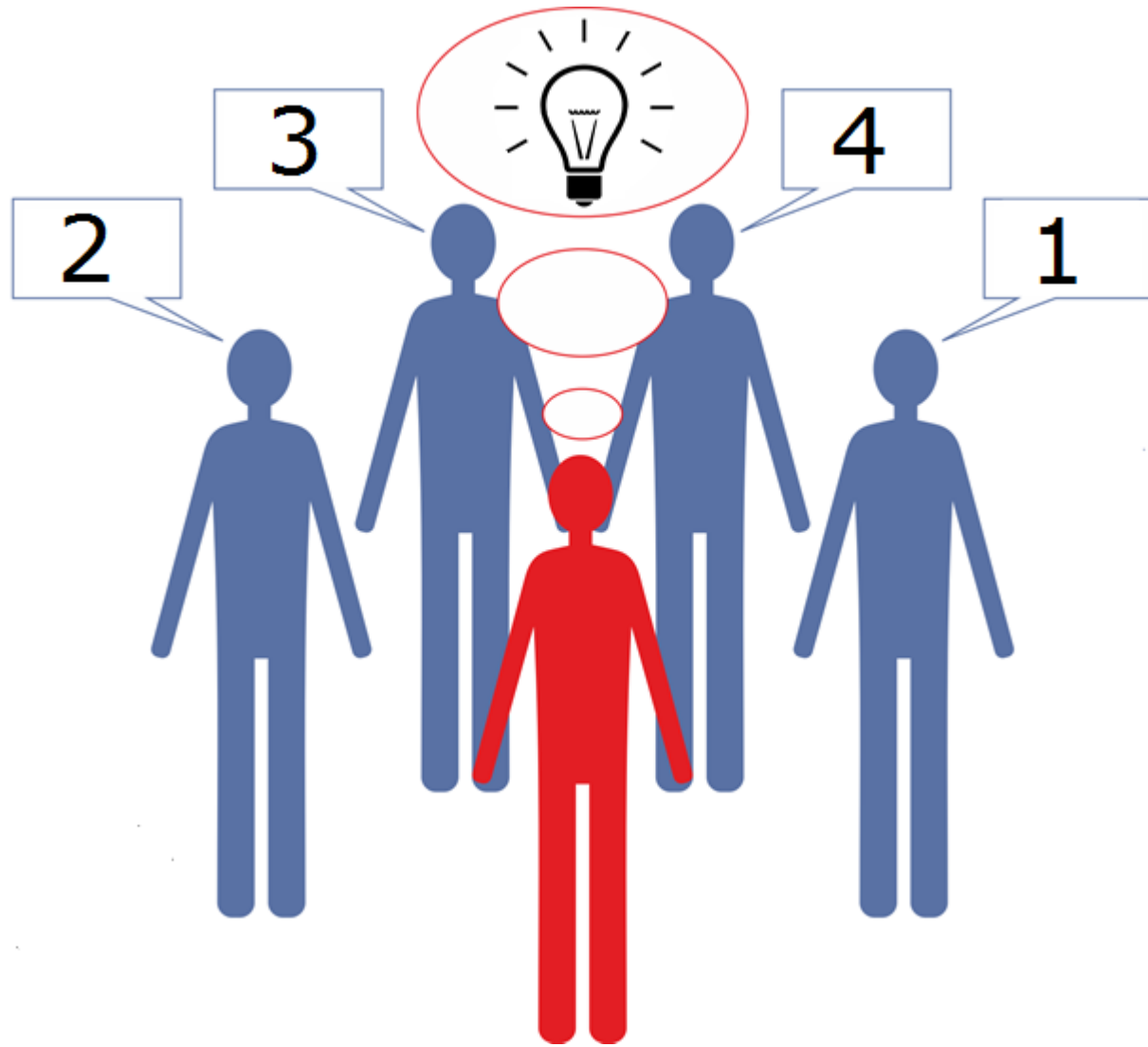
**Universiteit Utrecht**



**Universiteit Utrecht**







Can this be done for more complicated models?



**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





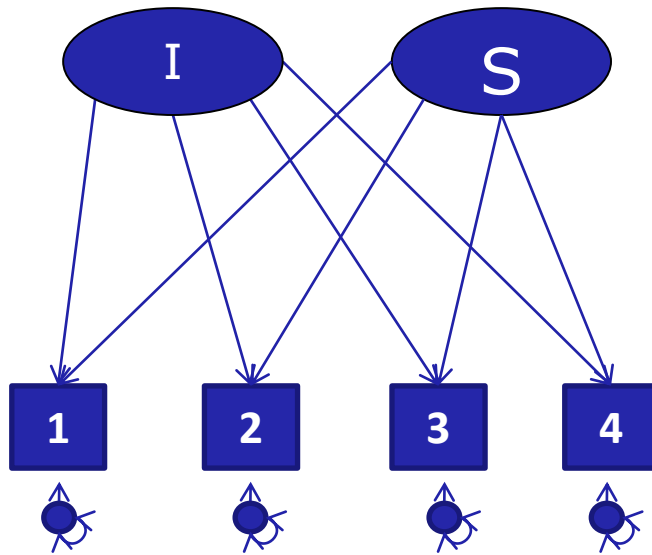
## Expert elicitation

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

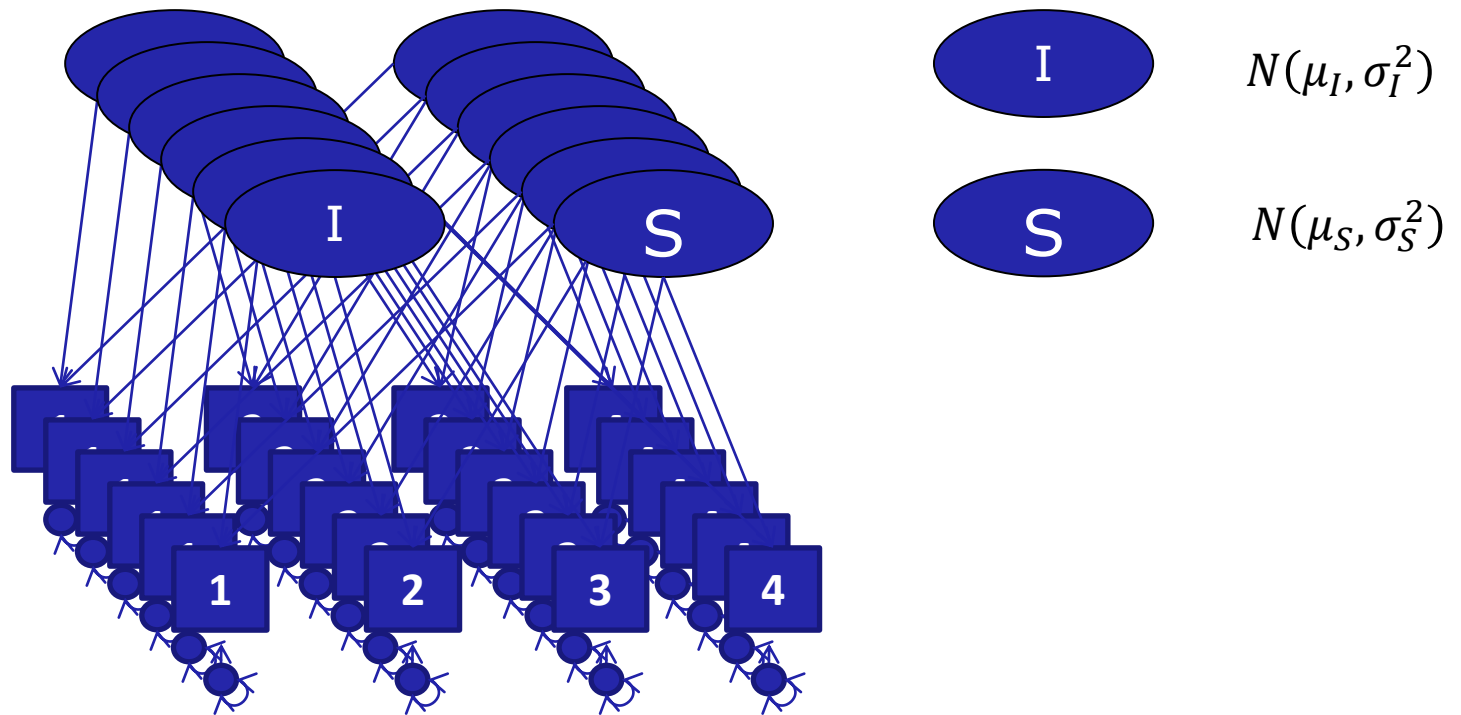




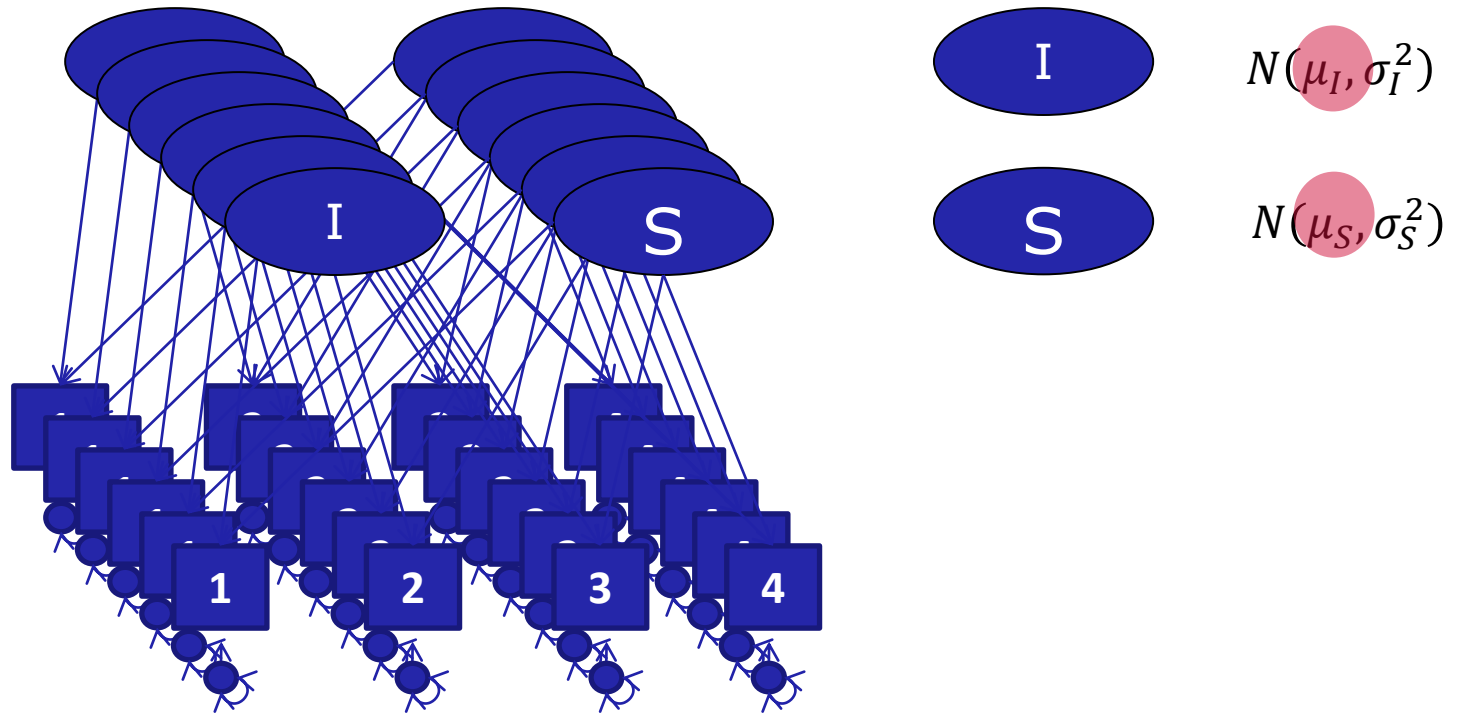
## Model per child



# Hierarchical model



# Hierarchical model

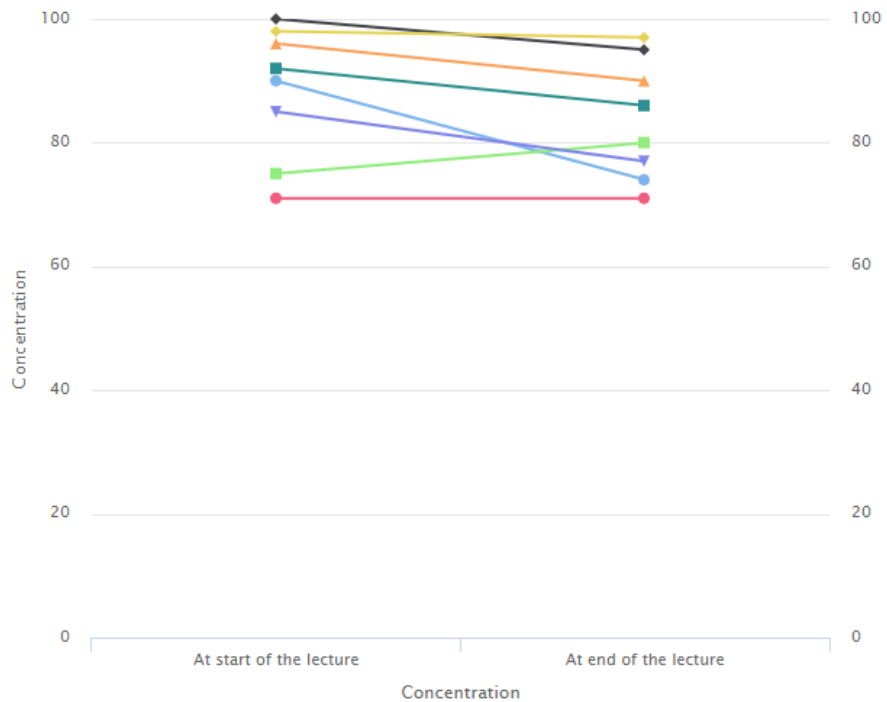




Add Additional Line

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

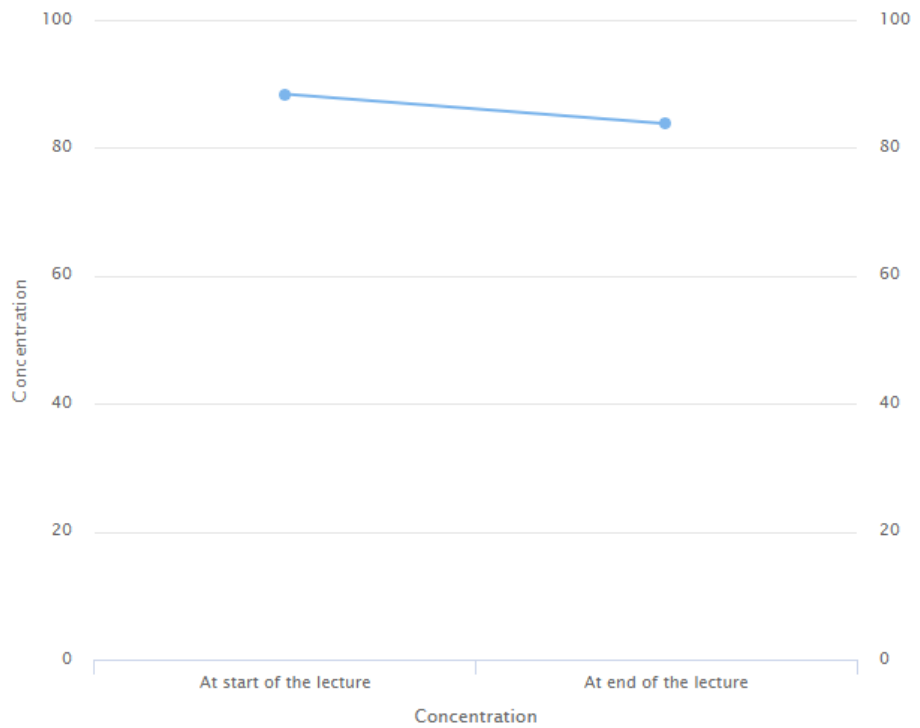




Add Additional Line

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

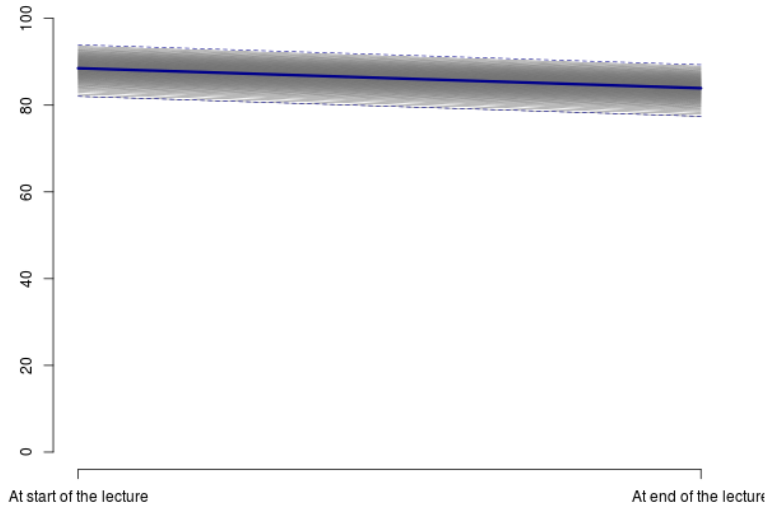
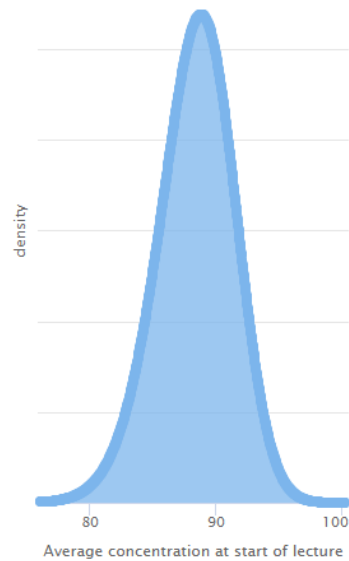
88,4

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

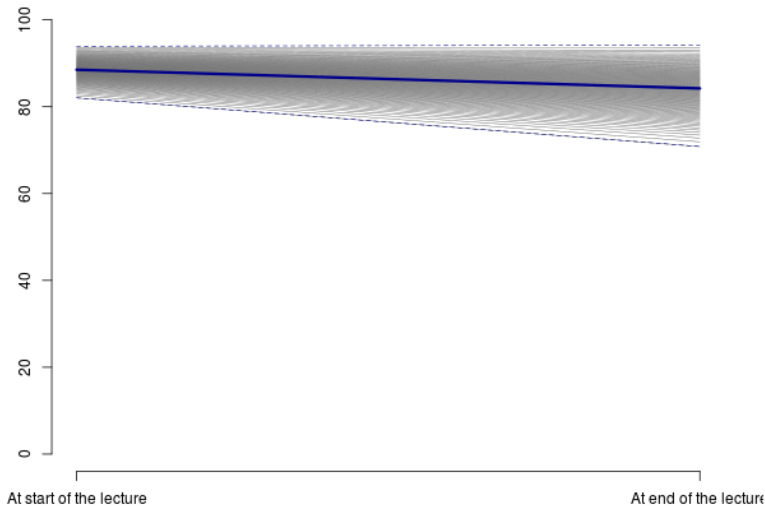
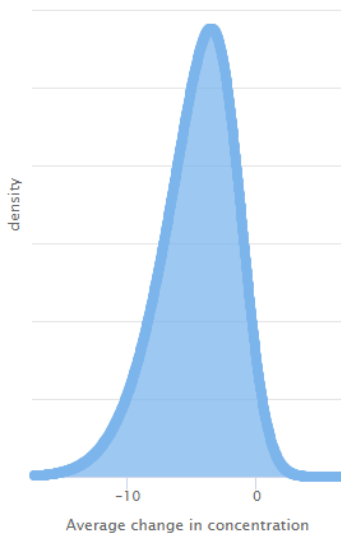
-4,6

Reasonable upperbound average change in concentration

-15

Fit distribution

Show implications

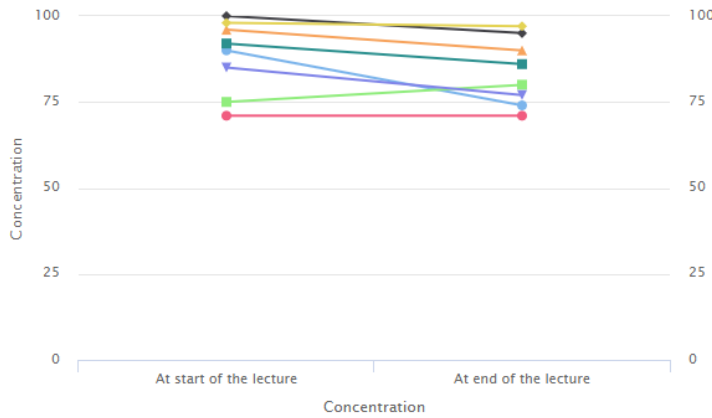


concentration				
2.5%	25%	50%	75%	97.5%
-11.2	-6.4	-4.3	-2.5	0.3
95% CI		50% CI		
[-11.2, 0.3]		[-6.4, -2.5]		

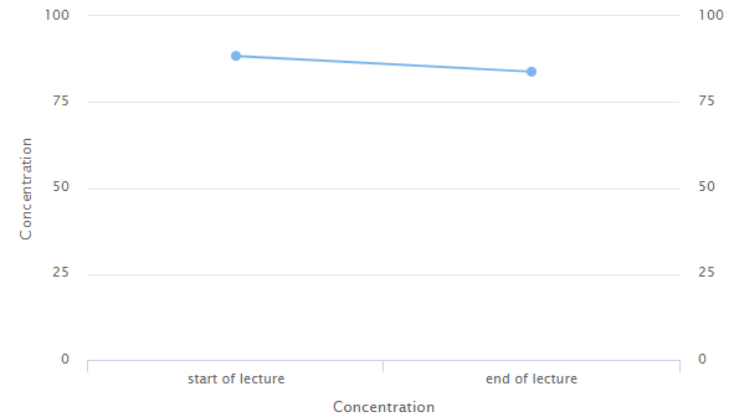




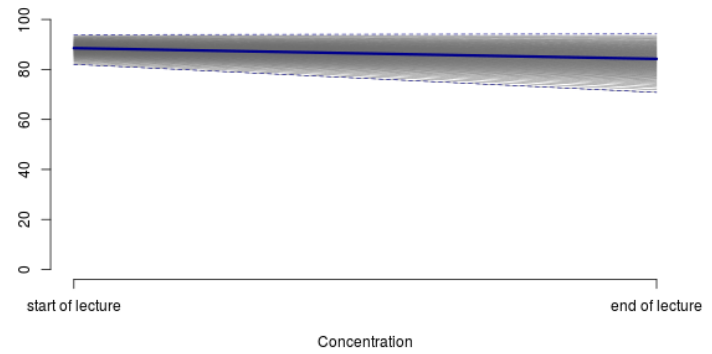
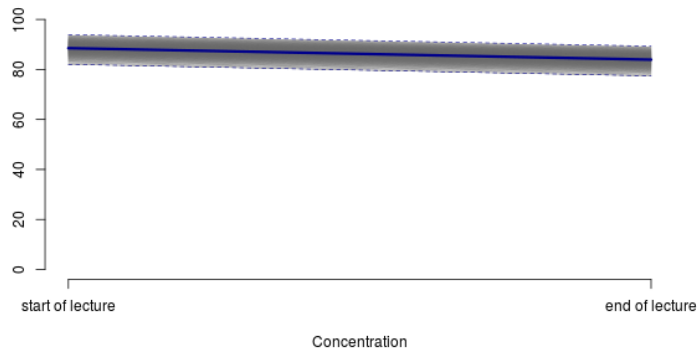
In this tab we provide a final summary of how we interpret your elicited beliefs and you can either agree to this or we go back to the relevant section of the procedure to adapt your input and our interpretation of your beliefs.



These are the concentration levels for your imagined individual children



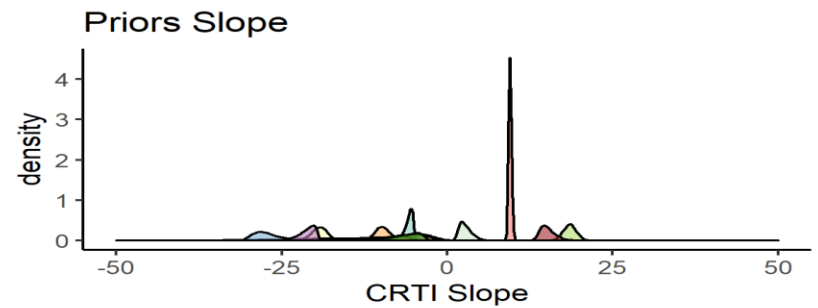
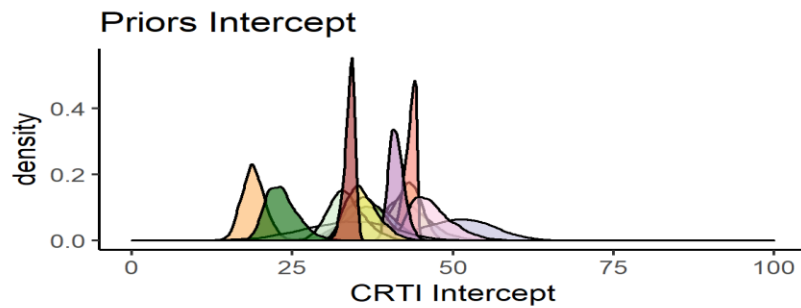
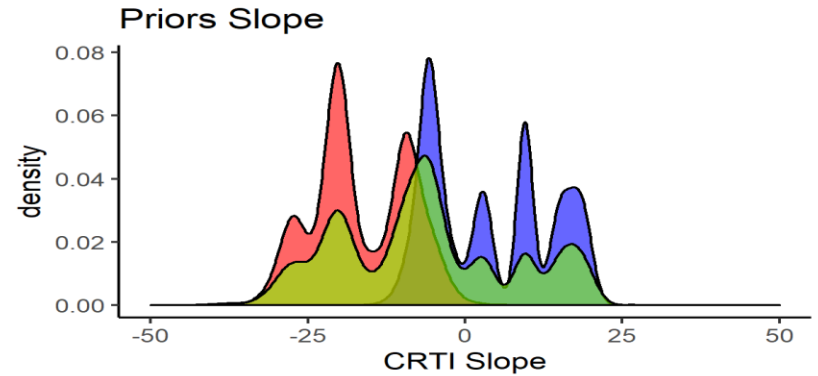
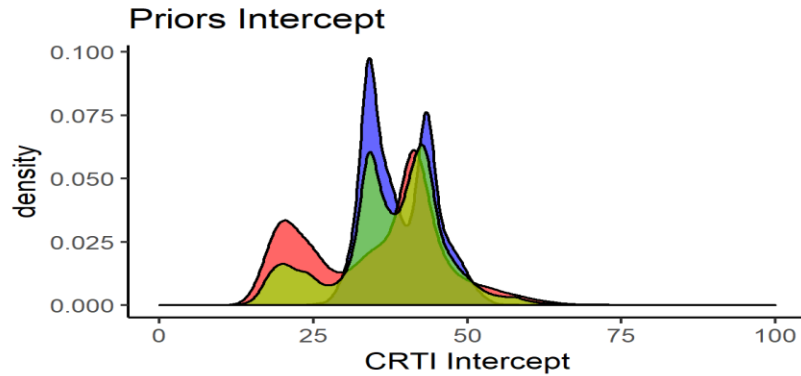
This is our interpretation of your beliefs regarding the average concentration levels at the start and the end of the lecture.







■ All experts ■ Nurses ■ Psychologists



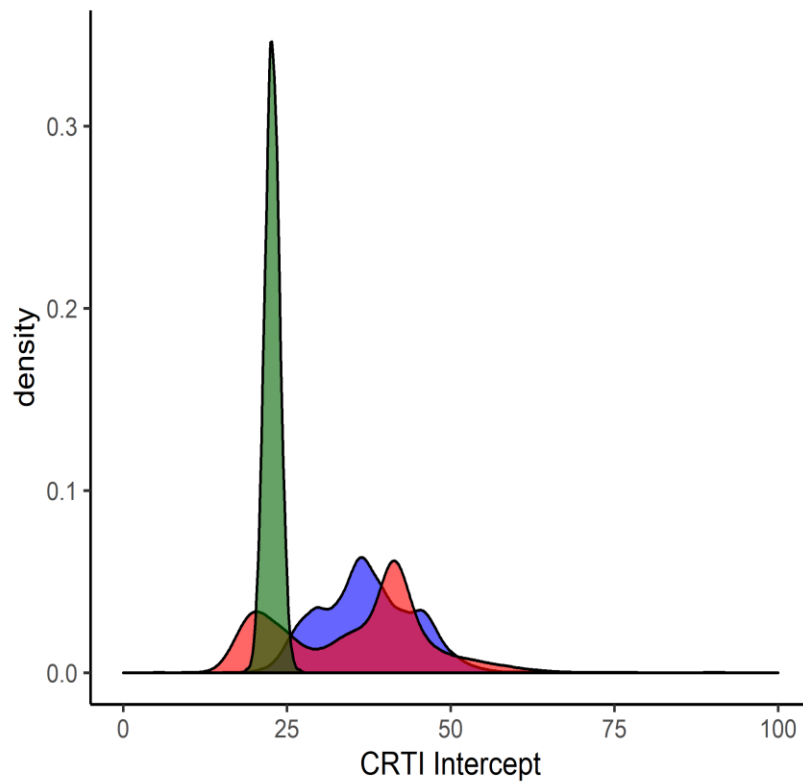
■ expert 1	■ expert 12	■ expert 2	■ expert 5	■ expert 8
■ expert 10	■ expert 13	■ expert 3	■ expert 6	■ expert 9
■ expert 11	■ expert 14	■ expert 4	■ expert 7	



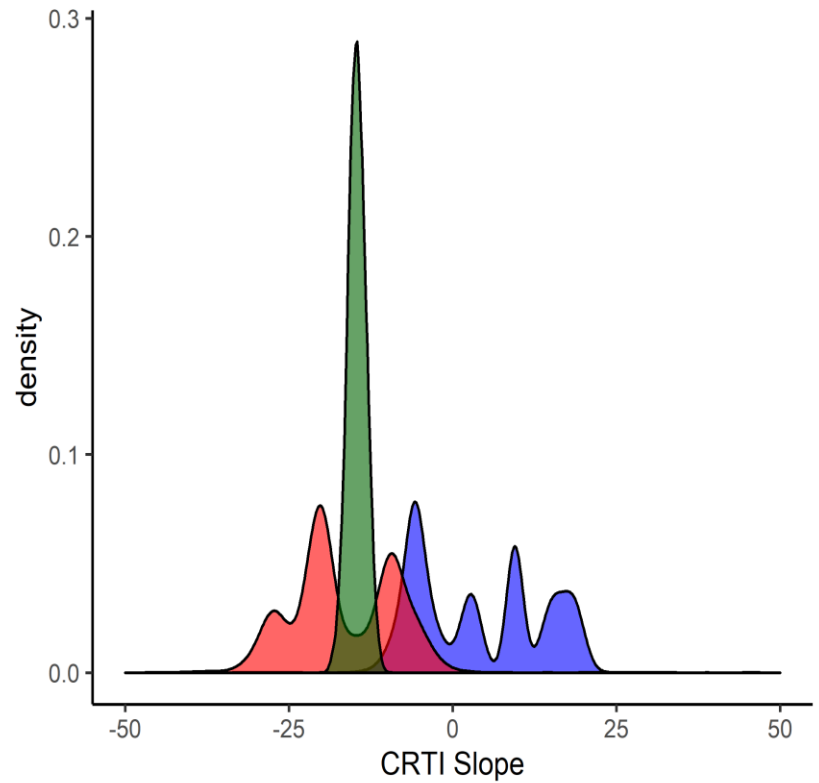


■ Nurses ■ Psychologists ■ Reference Posteriors

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



# 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....

