External Validity of Risk Elicitation

Episode IV: Measurement Error

Paolo Crosetto, Antonio Filippin, Daniel Navarro Martinez, Xinghua Wang

IMT Seminar – Lucca – March 5th, 2025

[redacted]

Slovic (1962):

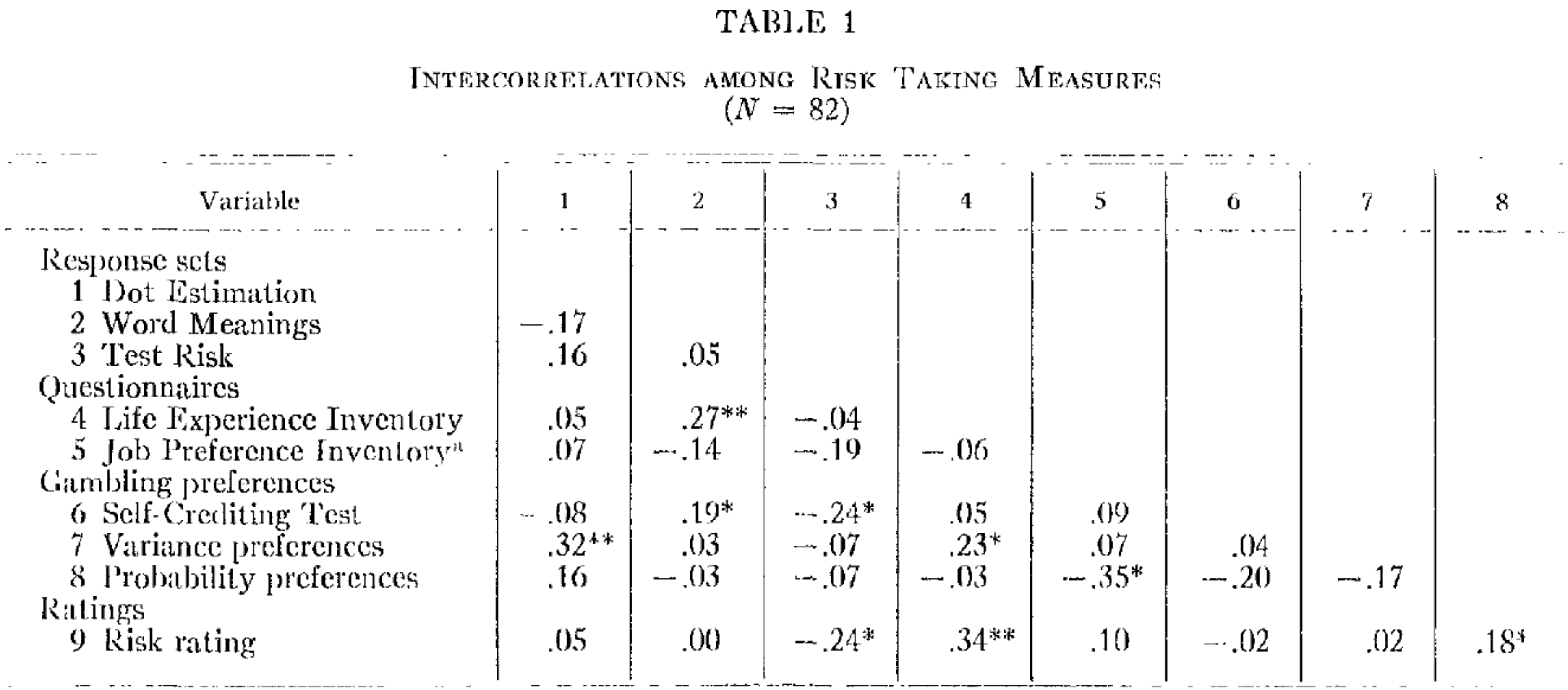


TABLE 1

INTERCORRELATIONS AMONG RISK TAKING MEASURES
($N = 82$)

Variable	1	2	3	4	5	6	7	8
Response sets								
1 Dot Estimation								
2 Word Meanings	−.17							
3 Test Risk	.16	.05						
Questionnaires								
4 Life Experience Inventory	.05	.27**	−.04					
5 Job Preference Inventory[a]	.07	−.14	−.19	−.06				
Gambling preferences								
6 Self-Crediting Test	−.08	.19*	−.24*	.05	.09			
7 Variance preferences	.32**	.03	−.07	.23*	.07	.04		
8 Probability preferences	.16	−.03	−.07	−.03	−.35*	−.20	−.17	
Ratings								
9 Risk rating	.05	.00	−.24*	.34**	.10	−.02	.02	.18*

> *"…future research must carefully consider the problem of adequately **defining** and **assessing** risk taking behavior."*

So, how are we doing?

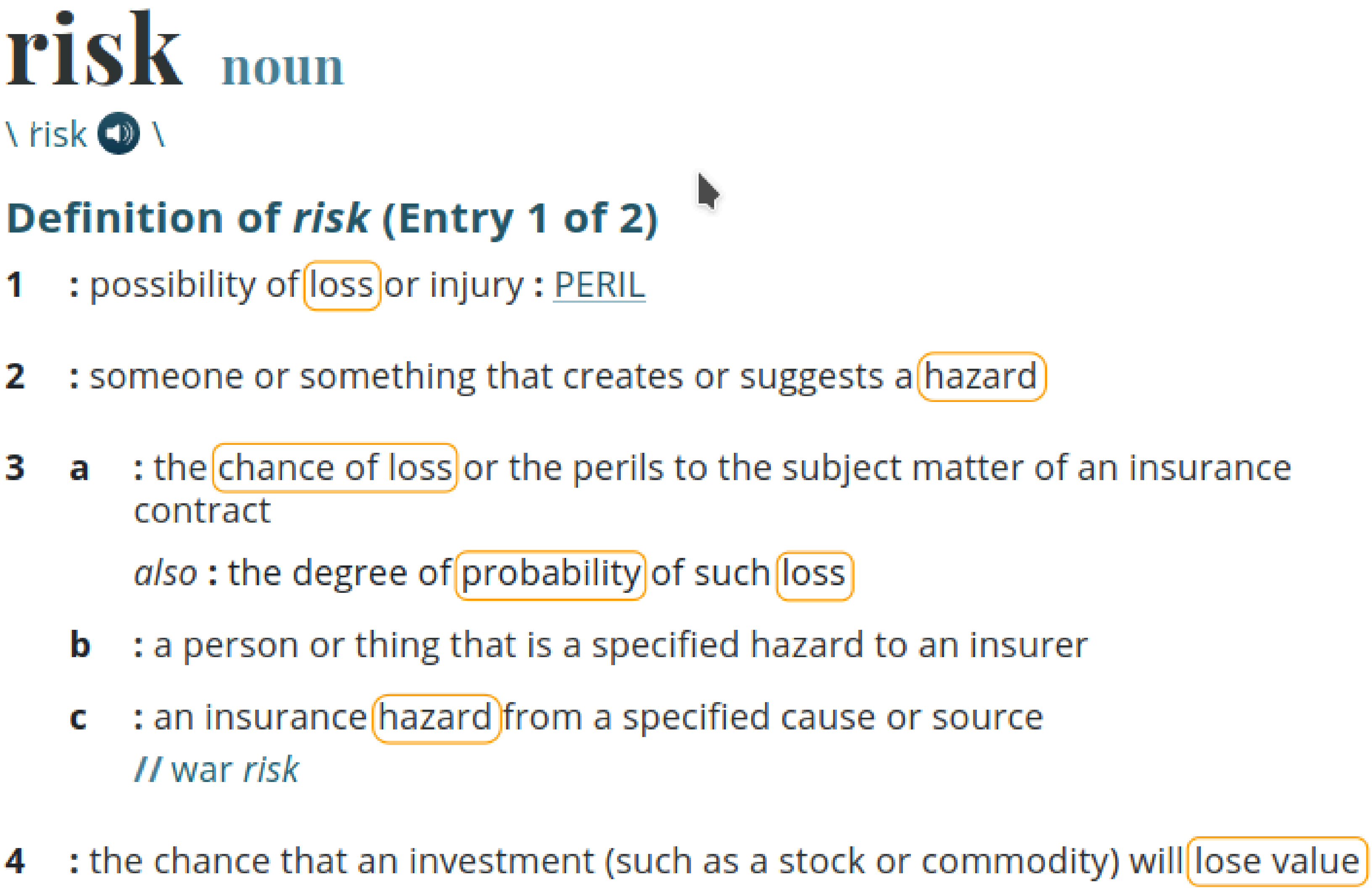
we live 60 years in Slovic’s future

This talk

- **Setting the stage**
  - the state of the art in the 2020s
- **Moving forward**
  - measurement error

Setting the stage

What is *risk*?



risk noun

\ ˈrisk \

Definition of *risk* (Entry 1 of 2)

1 : possibility of loss or injury : PERIL

2 : someone or something that creates or suggests a hazard

3 **a** : the chance of loss or the perils to the subject matter of an insurance contract

also : the degree of probability of such loss

b : a person or thing that is a specified hazard to an insurer

c : an insurance hazard from a specified cause or source

// war risk

4 : the chance that an investment (such as a stock or commodity) will lose value

Measuring risk attitudes

> A **difficult** task with **crucial** relevance

- directly *unobservable*: *latent* (⇒ requires a theory)
- should we..
  - *infer* from real world data or from *ad-hoc* choices
  - ask or **t**ask?
  - elicit by *description* or by *experience*?

The state of the art: psychology

> Risk as **probability of harm.**

- **Questionnaires**:
  - directly ask, over different domains
  - tackle risk perception
- **Tasks**
  - putting the subject in a ‘risky’ situation
  - card/gambling tasks

The state of the art: economics

> risk formally defined as **uncertainty over outcomes**.

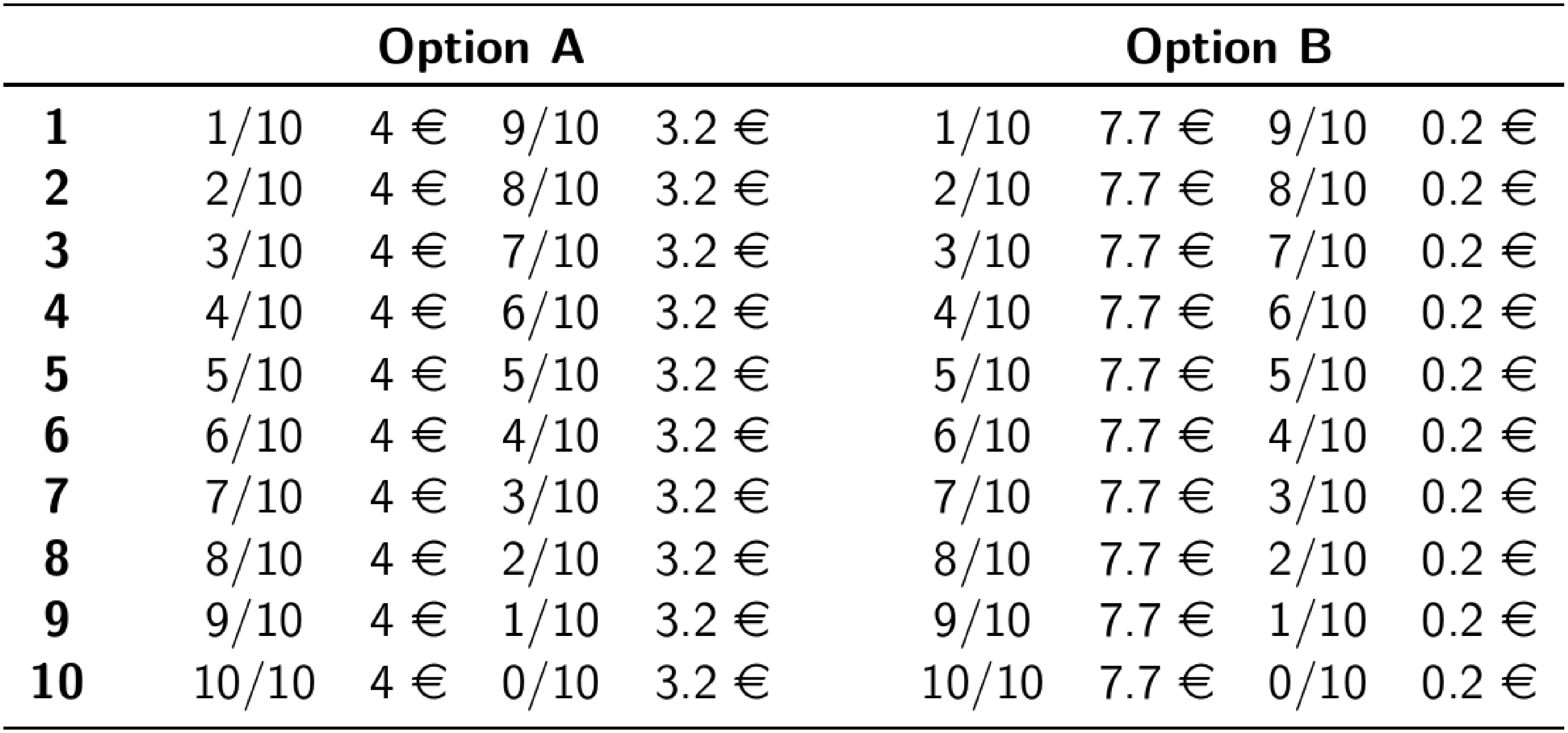
The lottery paradigm

- incentives – choice over lotteries
- strong theoretical underpinning
- different formats, cover stories, contexts
- estimation of utility functions (⇒ models)

> Metric of success: **internal validity** (task ⇒ theory)

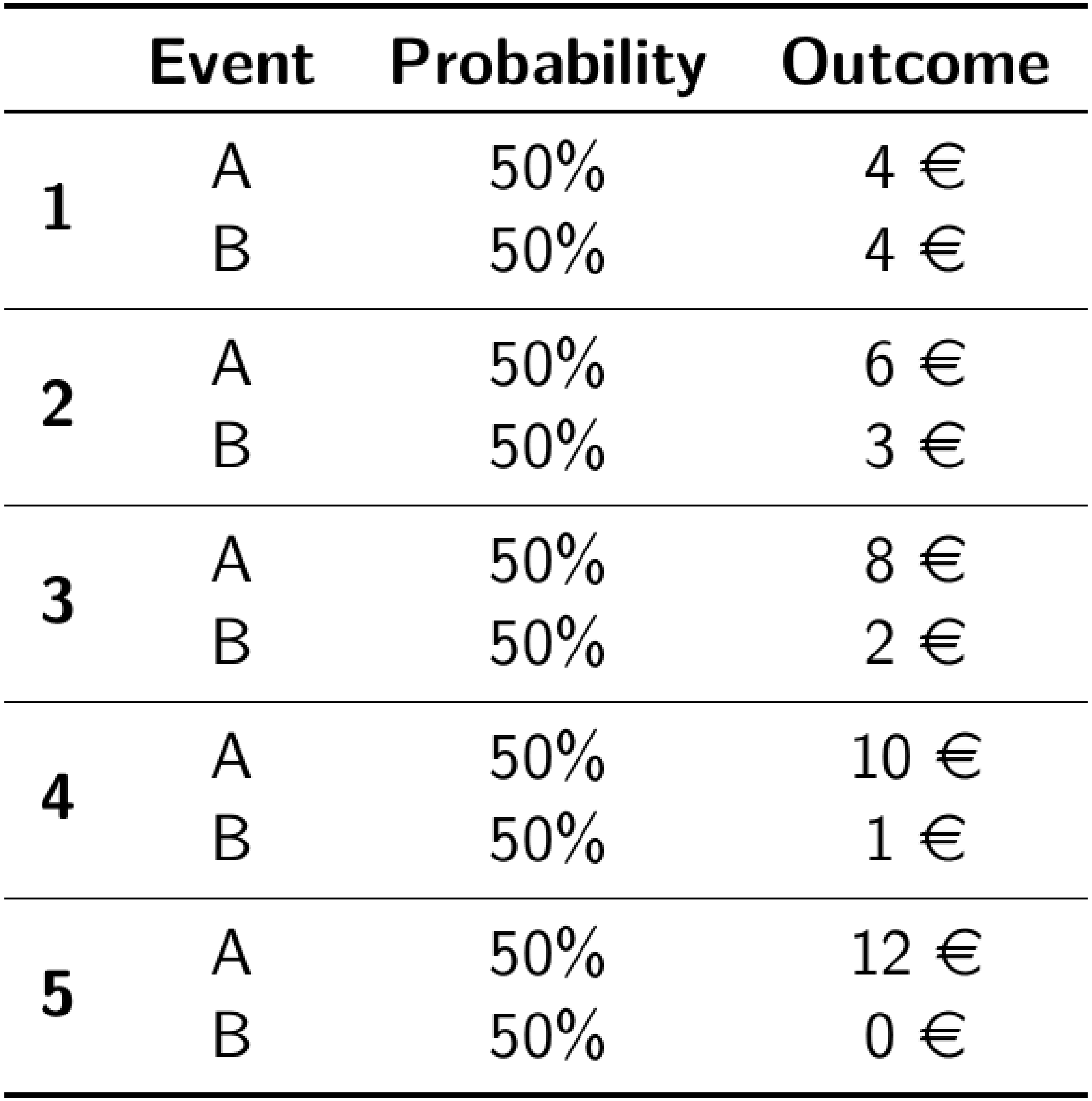
Tools: RETs

Holt and Laury



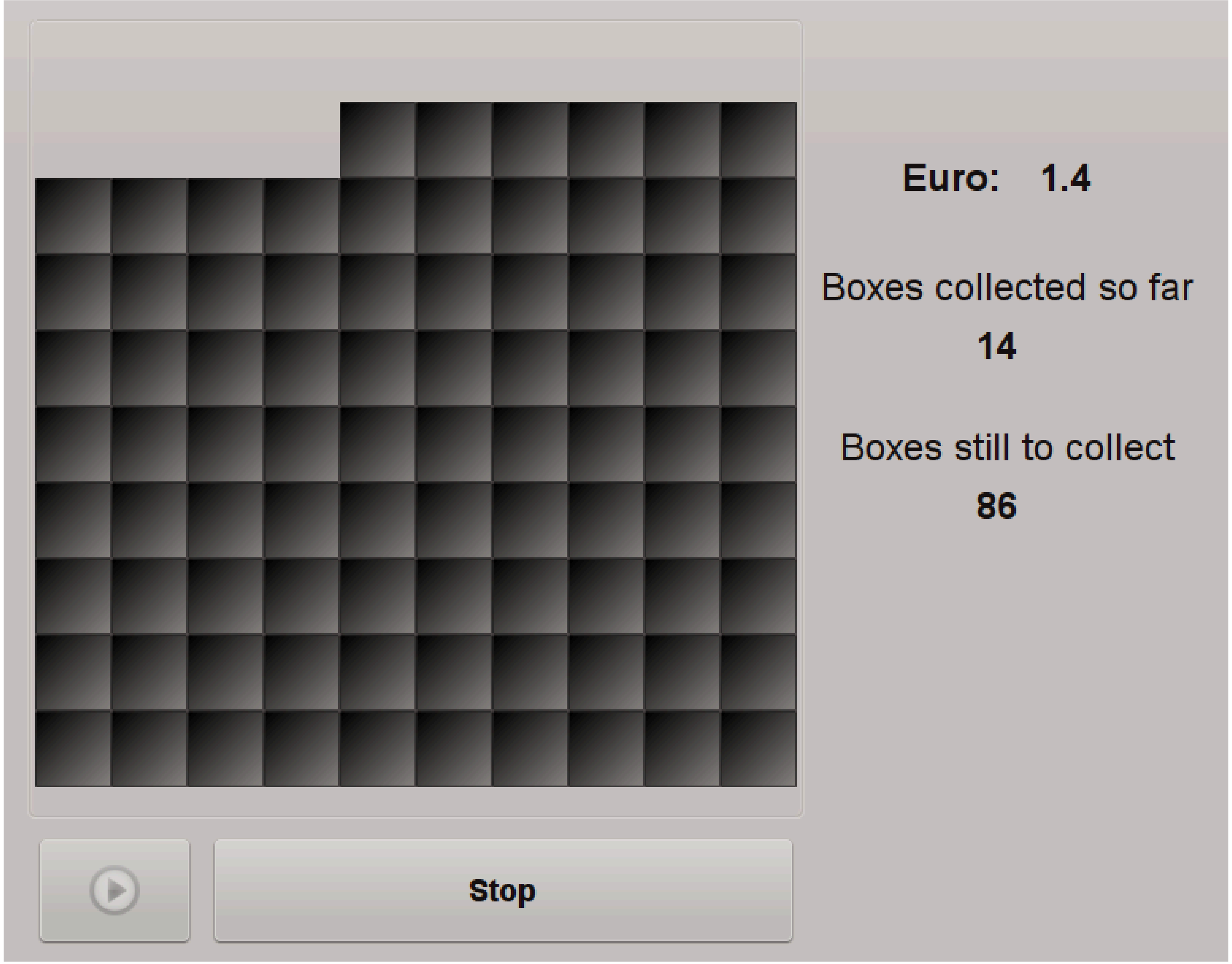
	Option A				Option B			
1	1/10	4 €	9/10	3.2 €	1/10	7.7 €	9/10	0.2 €
2	2/10	4 €	8/10	3.2 €	2/10	7.7 €	8/10	0.2 €
3	3/10	4 €	7/10	3.2 €	3/10	7.7 €	7/10	0.2 €
4	4/10	4 €	6/10	3.2 €	4/10	7.7 €	6/10	0.2 €
5	5/10	4 €	5/10	3.2 €	5/10	7.7 €	5/10	0.2 €
6	6/10	4 €	4/10	3.2 €	6/10	7.7 €	4/10	0.2 €
7	7/10	4 €	3/10	3.2 €	7/10	7.7 €	3/10	0.2 €
8	8/10	4 €	2/10	3.2 €	8/10	7.7 €	2/10	0.2 €
9	9/10	4 €	1/10	3.2 €	9/10	7.7 €	1/10	0.2 €
10	10/10	4 €	0/10	3.2 €	10/10	7.7 €	0/10	0.2 €

Binswanger / Eckel and Grossmann



	Event	Probability	Outcome
1	A	50%	4 €
	B	50%	4 €
2	A	50%	6 €
	B	50%	3 €
3	A	50%	8 €
	B	50%	2 €
4	A	50%	10 €
	B	50%	1 €
5	A	50%	12 €
	B	50%	0 €

Bomb Risk Elicitation Task



Euro: 1.4

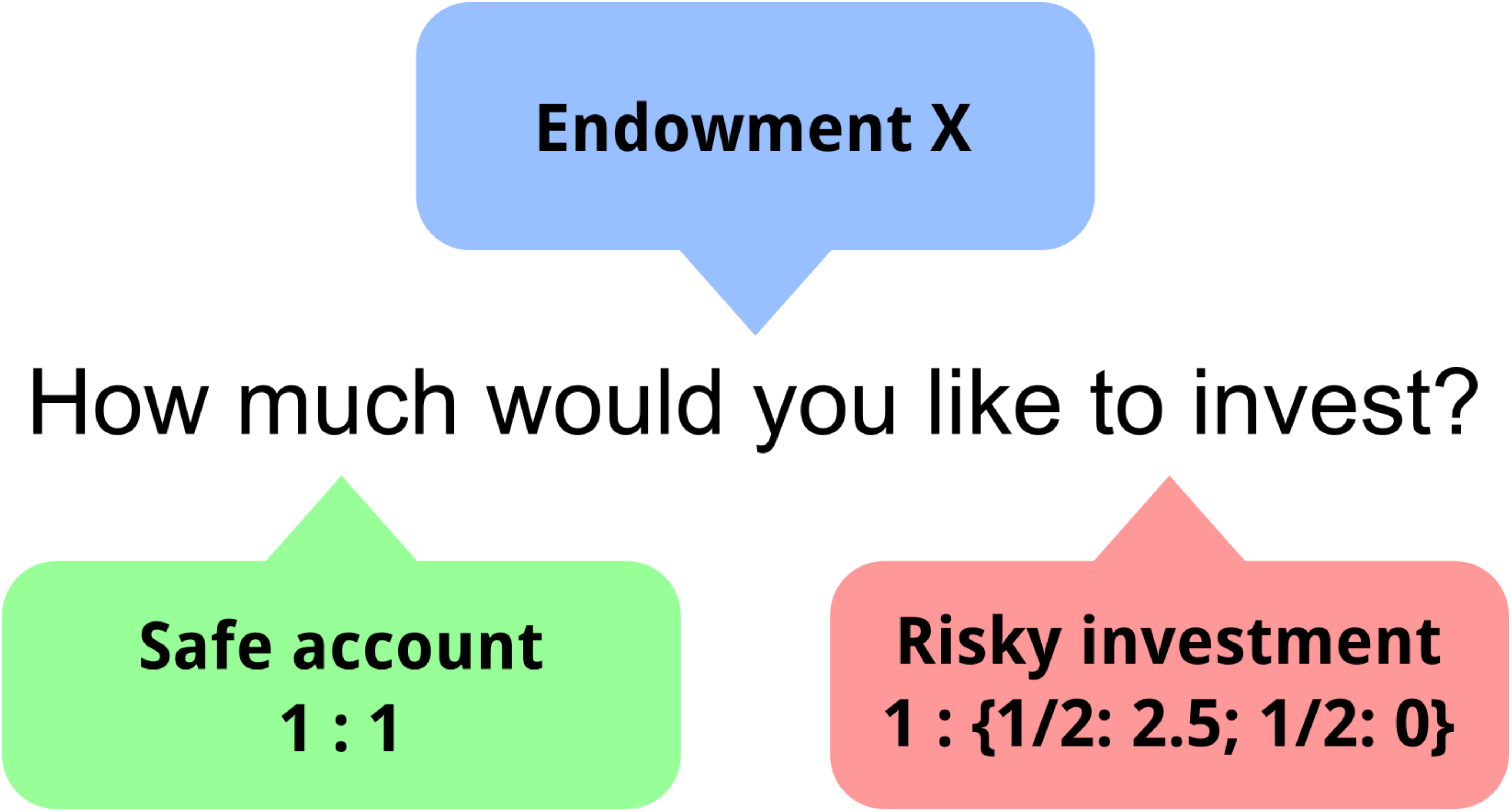
Boxes collected so far

14

Boxes still to collect

86

Investment Game (Gneezy & Potters)



Endowment X

How much would you like to invest?

Safe account
1 : 1

Risky investment
1 : {1/2: 2.5; 1/2: 0}

Certainty Equivalent MPL

A	B	
100%	50%	50%
0		
10		
20		
30		
40		
50	100	0
60		
70		
80		
90		
100		

Questionnaire: SOEP

How likely are you to take risks in general, one a scale from 0 (not taking any risks) to 10 (taking many risks)?

(with further additional questions by domain, as health, driving, sports…)

Questionnaire: DOSPERT

Domain **Spe**cific **R**isk **T**aking Scale

- 6 domains: investing, gambling, health/safety, recreational, ethical, and social
- 1 to 7 scale: *how risky do you think X is?*
- 1 to 7 scale: *how likely are you to engage in X?*

Examples:

- Riding a motorcycle without a helmet.
- Investing 10% of your annual income in a moderate growth diversified fund.

What do we know?

METARET

> A meta-analysis of Risk elicitation tasks

- **elicited** risk attitudes: tasks and questionnaires
- **convergent** validity: correlations among measures
- **predictive** validity: correlation task $\iff$ questionnaires

Explore the data!

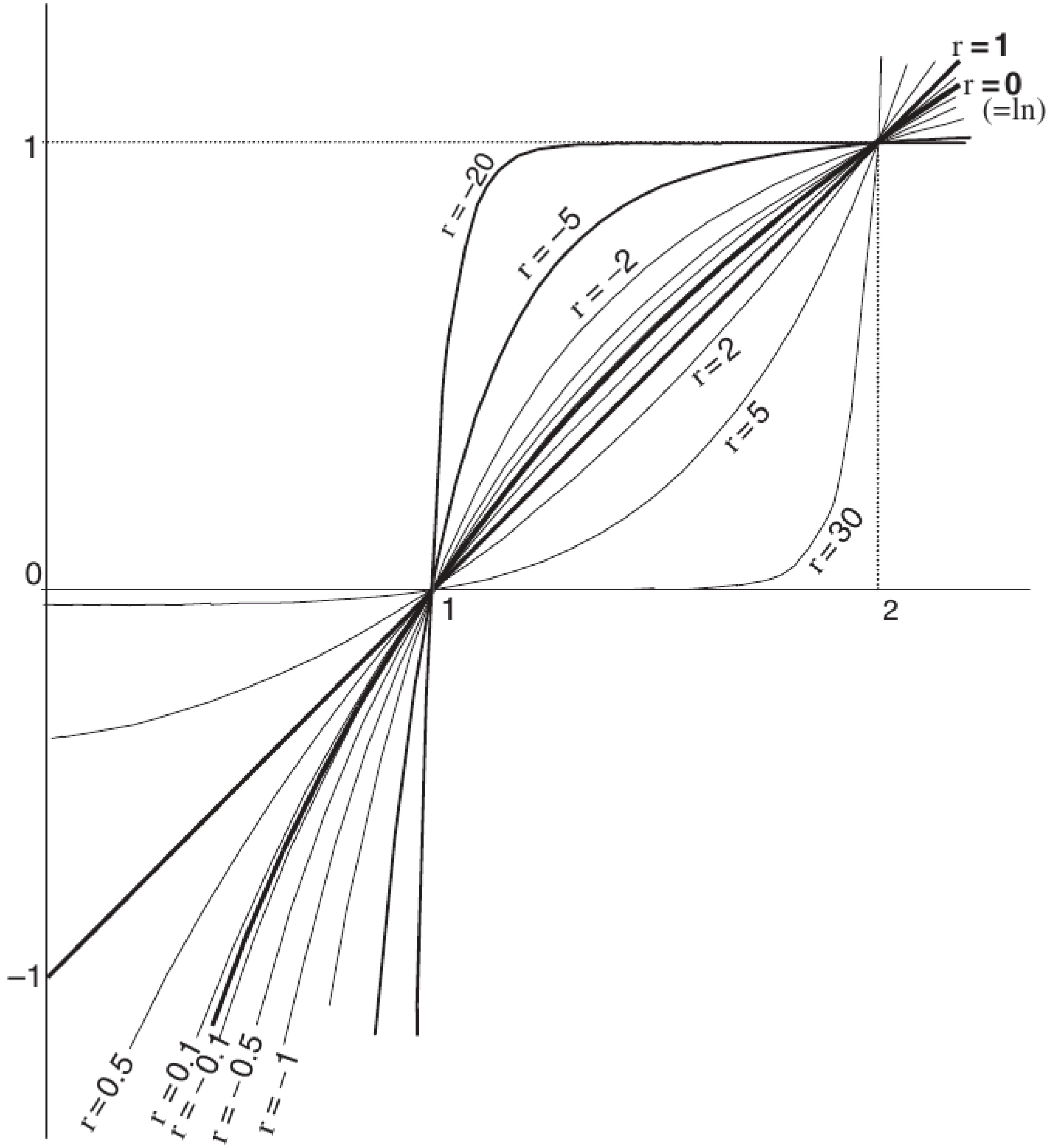
> Live data exploration on a shiny app

Assumptions: CRRA (à la Wakker)

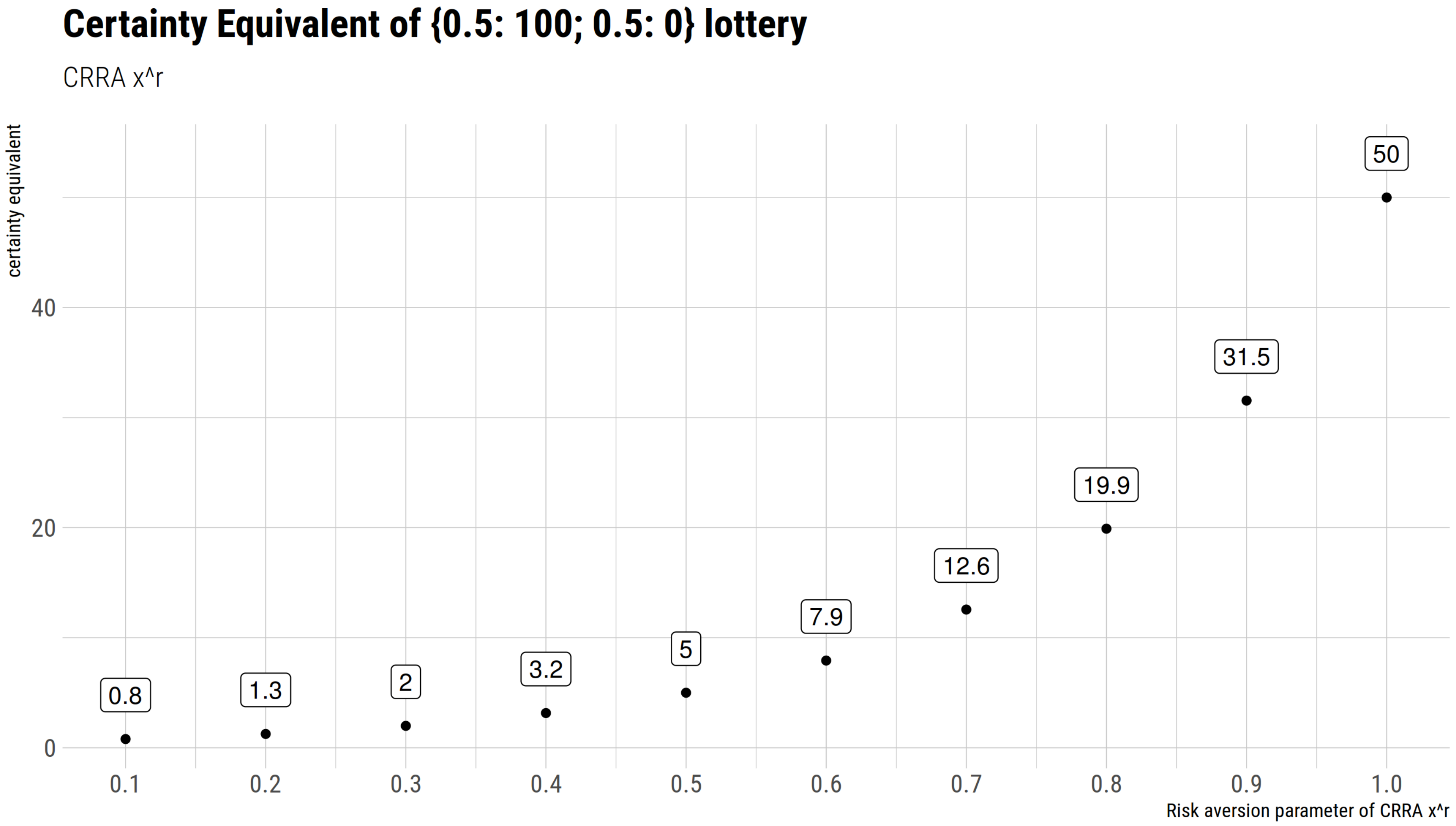
$$u(x) = x^r$$

- simple
- captures risk aversion
- makes different tasks comparable

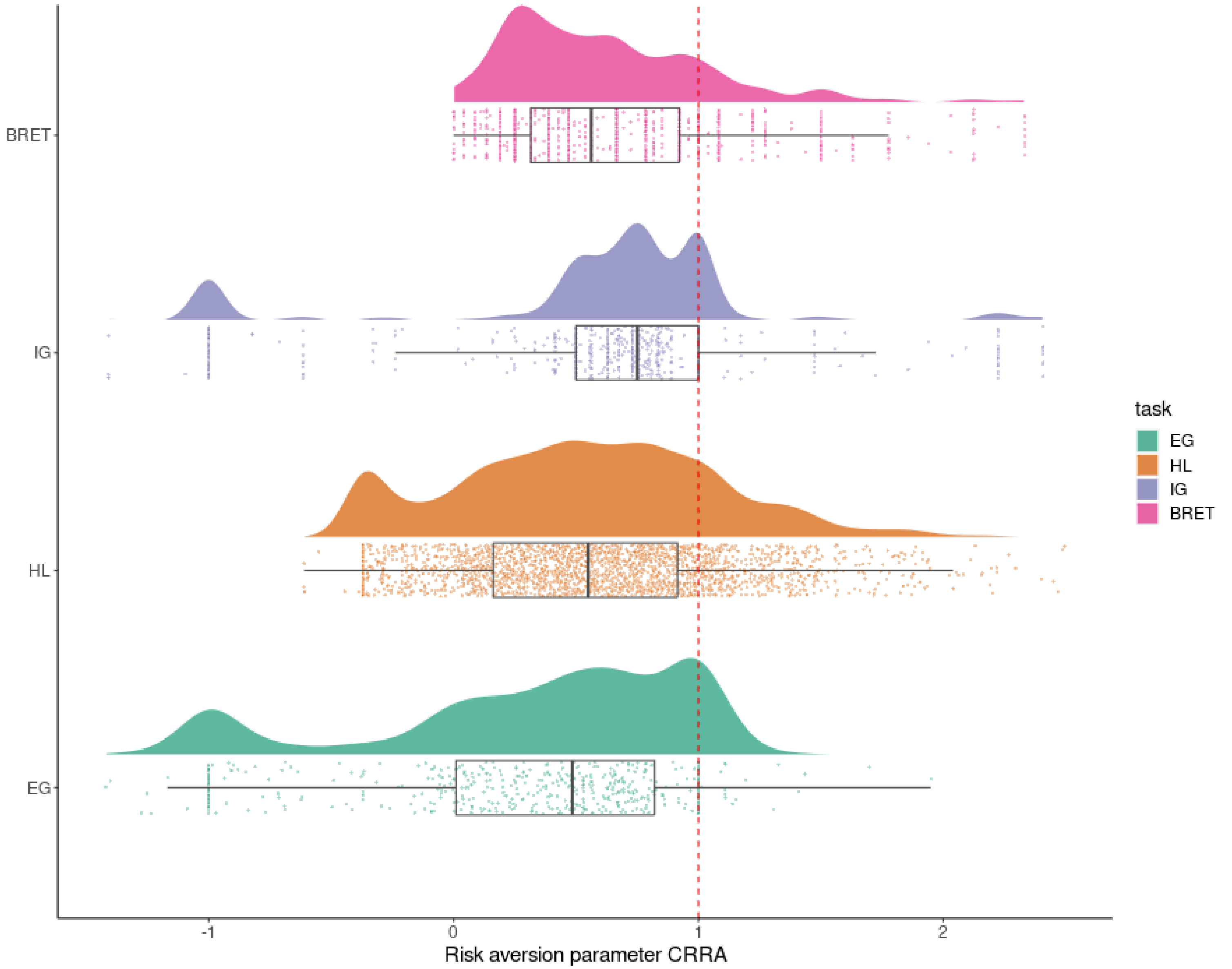
CRRA



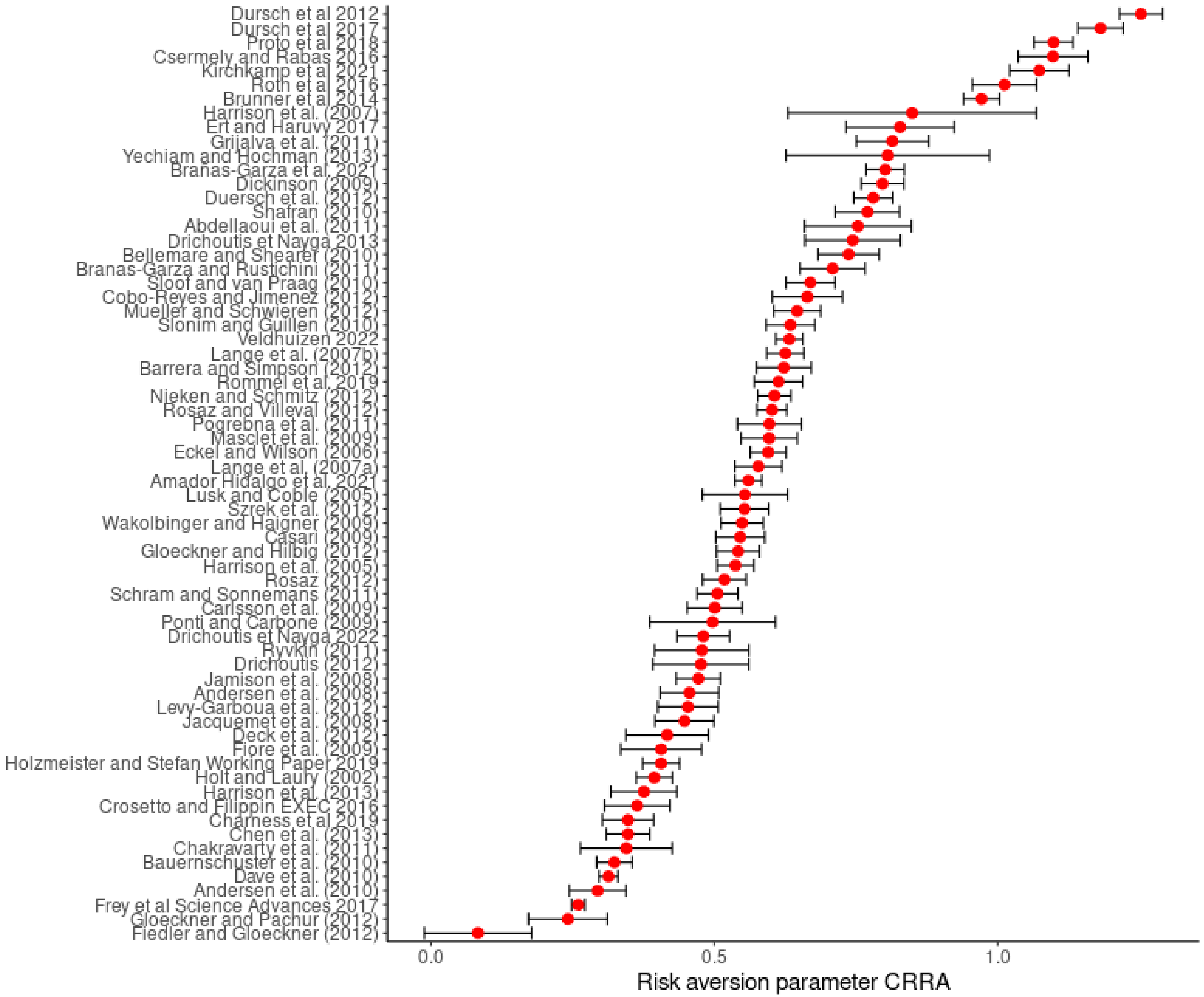
How big are the differences?



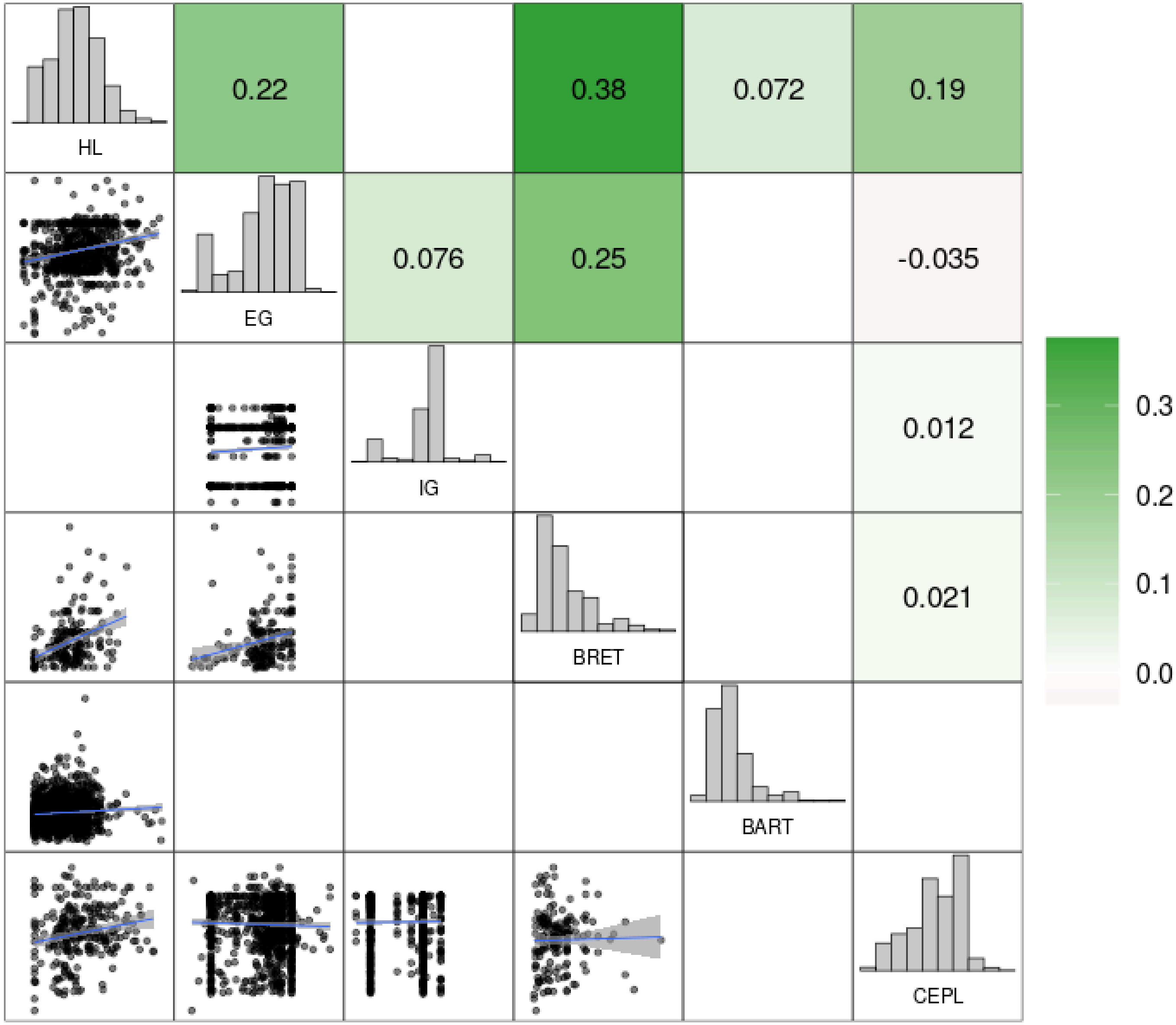
Low consistency *across* tasks



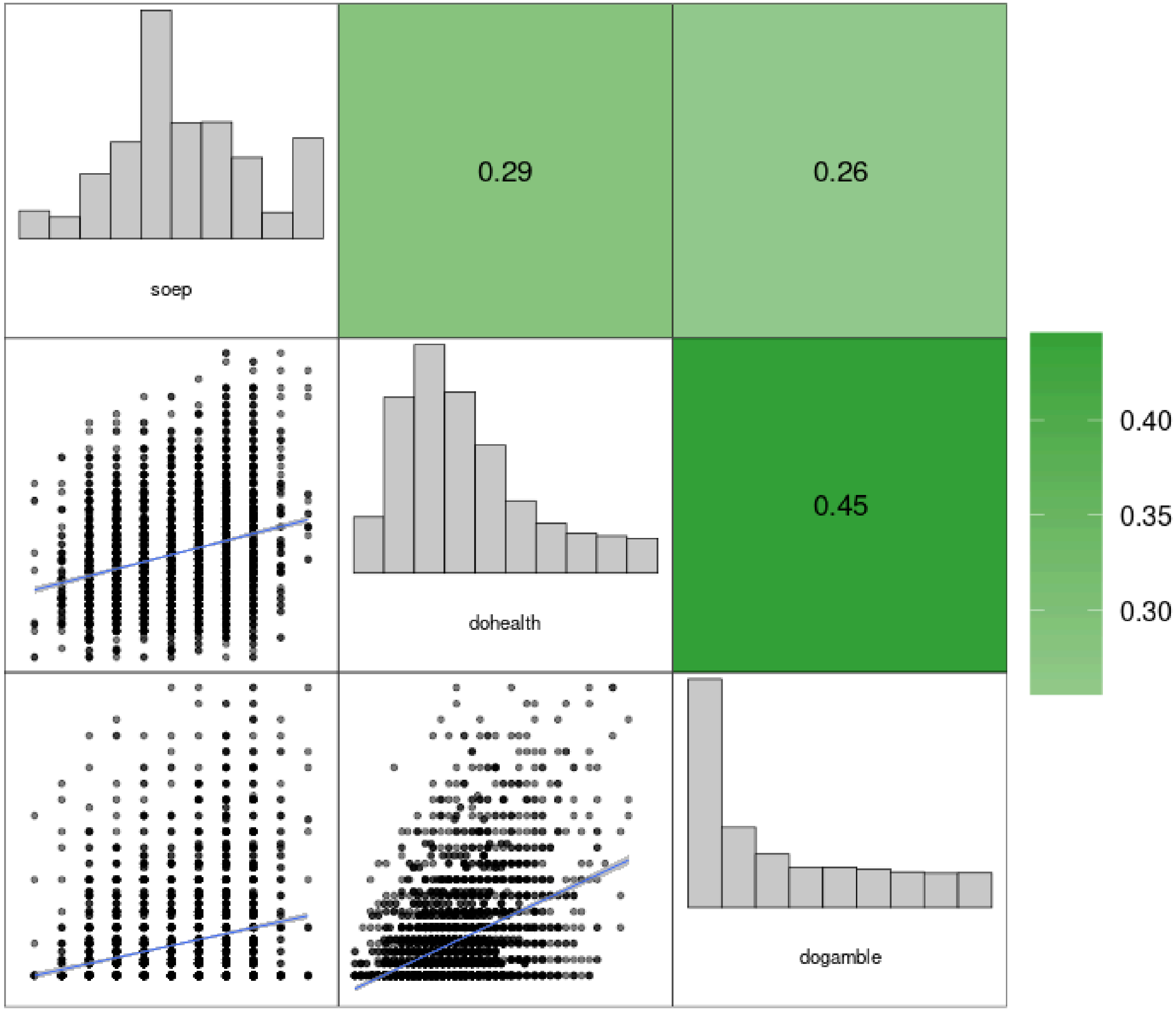
Low consistency *within* tasks



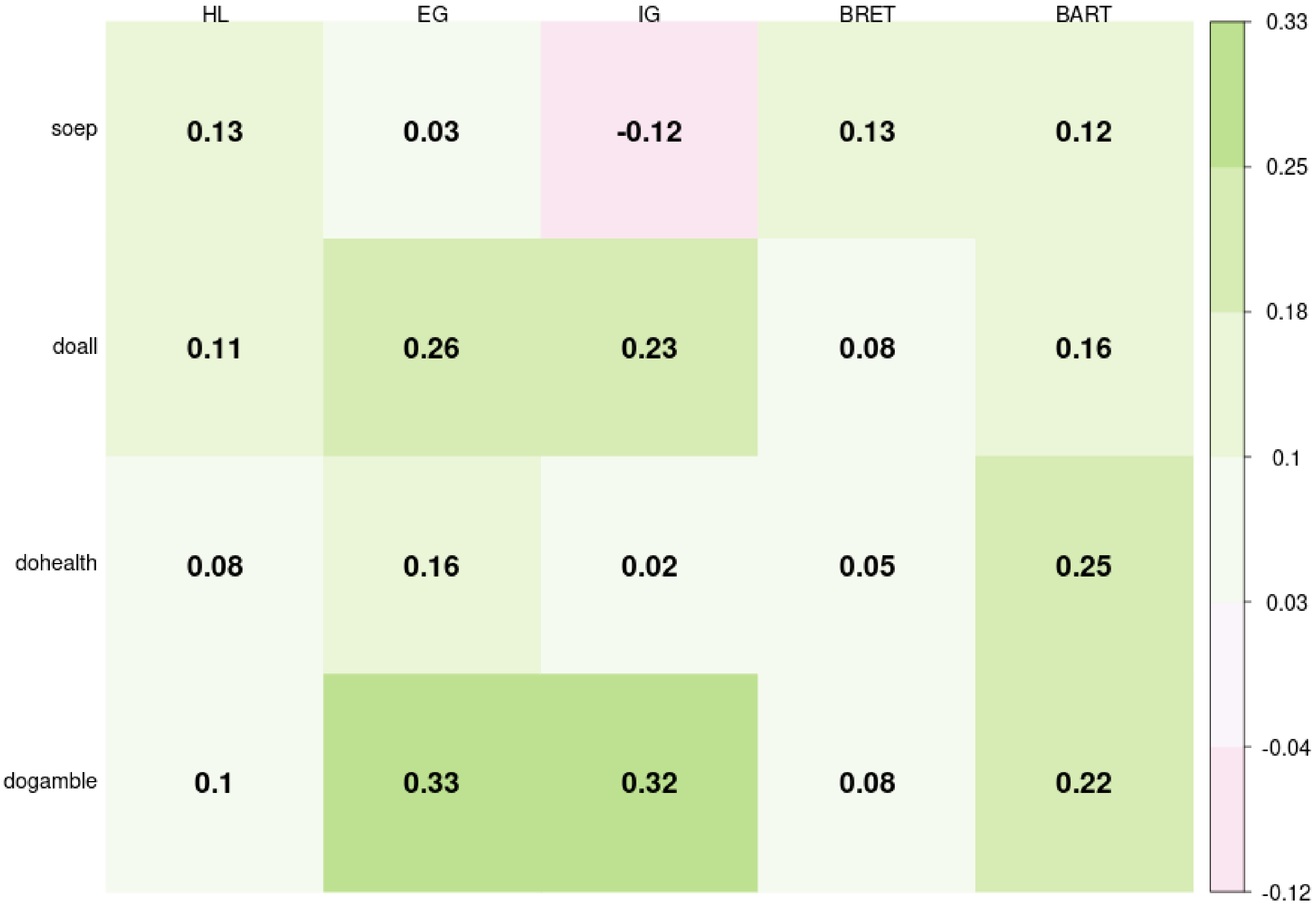
Convergence: tasks



Convergence: questionnaires



Tasks ⟺ Questionnaires



	HL	EG	IG	BRET	BART
soep	0.13	0.03	-0.12	0.13	0.12
doall	0.11	0.26	0.23	0.08	0.16
dohealth	0.08	0.16	0.02	0.05	0.25
dogamble	0.1	0.33	0.32	0.08	0.22

Summing up…

- *“…future research must carefully consider the problem of adequately* ***defining*** *and* ***assessing*** *risk taking behavior.”*

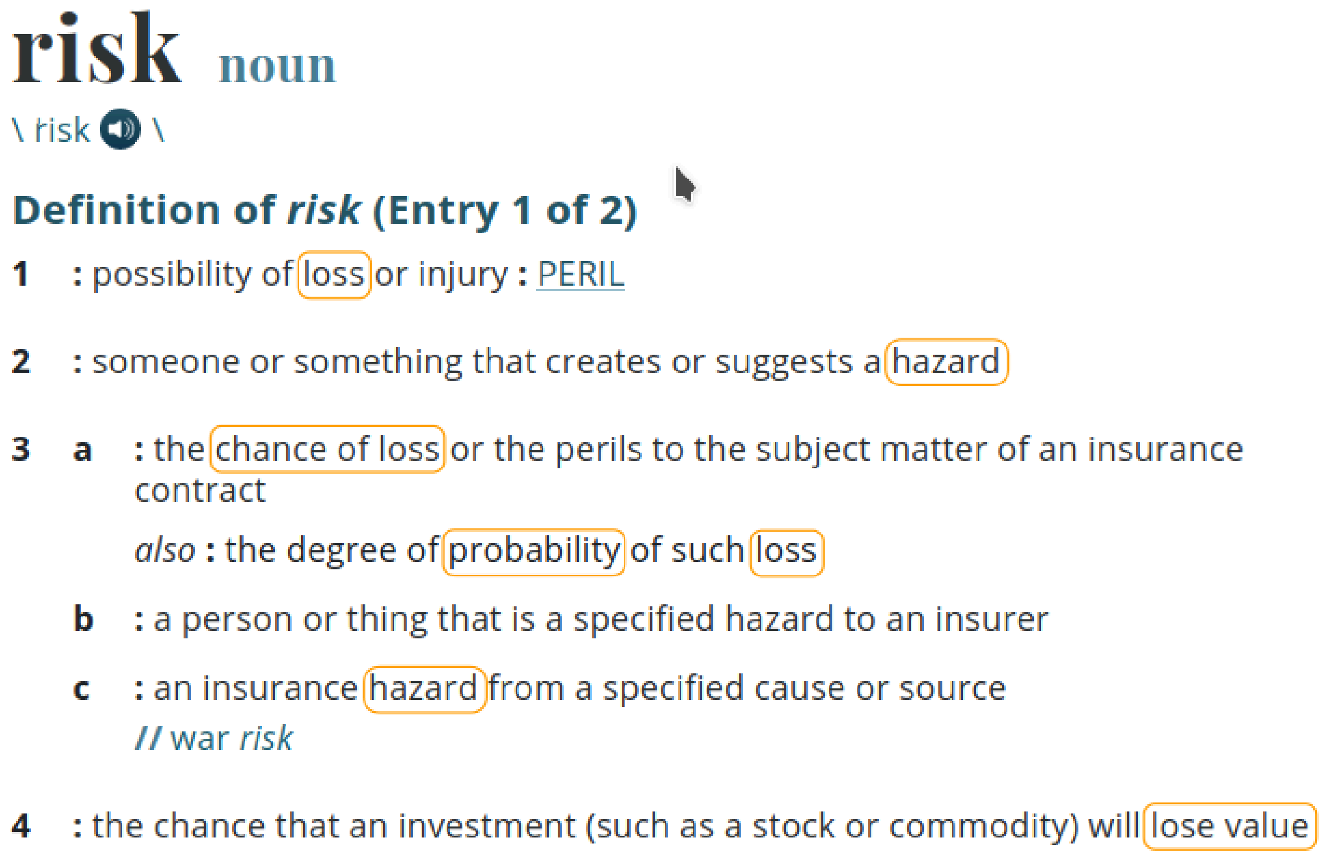
- **exactly as in 1962**

I am your father.

The road ahead

Is it *perception*?

> $\textit{observed} = \textit{attitude} + \textit{perception} \Rightarrow$ biased inference



risk noun

\ ˈrisk \

Definition of *risk* (Entry 1 of 2)

1 : possibility of loss or injury : PERIL

2 : someone or something that creates or suggests a hazard

3 **a** : the chance of loss or the perils to the subject matter of an insurance contract

also : the degree of probability of such loss

b : a person or thing that is a specified hazard to an insurer

c : an insurance hazard from a specified cause or source

// war *risk*

4 : the chance that an investment (such as a stock or commodity) will lose value

Is it *noise*?

> *noisy* preference + one-shot choices ⇒ noisy data

- fuzzy preferences
  - i.i.d. ⇒ measurement error
  - *not* i.i.d. ⇒ task-specific bias (*Crosetto & Filippin 2015*)
- other potential reasons for noise
  - cognitive limits ⇒ limited understanding
  - context-dependence

Measurement error

(with Antonio FIlippin, Daniel Navarro Martinez, Xinghua Wang)

Some references to get there

> Measurement error: $\hat{X} = X + \varepsilon,\ \ \varepsilon\,\text{i.i.d.}$

- *(1) Gillen et al 2019* – ME ⇒ false positives + techniques
- *(2) Galizzi & Navarro Martinez 2019* – social preference games have low external validity
- *(3) Navarro Martinez & Wang 2022* – applying (1) to (2) increases external validity
- *(4) this paper* – applying (3) to the risk elicitation puzzle

Goal

> Does aggregating **multiple** measures **over time** to reduce measurement error help solving the *risk elicitation puzzle*?

Experimental design

Setup

> Follow subjects for **2 weeks** taking **repeated** measures of

- Questionnaires;
- Risk Elicitation Tasks;
- Daily Reconstruction Method (*Kahneman et al. 2004*)

Timeline

- **Day 0**: lab session, instructions, all tasks, all questionnaires
- **Days 1-14**
  - every day: Daily Reconstruction Method
  - every odd day: Tasks
  - every even day: Questionnaires

> **Payment**: one random tasks in day 0 + show-up fee; 1 task per odd day + 1.5€ per DRM

Validity checks

> We test different type of psychometric validity

- *Test-retest reliability* for tasks and questionnaires
- *Convergent validity* among tasks
- *Convergent validity* among questionnaires
- *Convergent validity* tasks <> questionnaires
- *External validity* task > DRM and questionnaires > DRM

Questionnaires

> Focus on the most widely used

- **SOEP** (aggregate and by domain, 1 question)
- **DOSPERT** (over domains, 30 questions)

Tasks

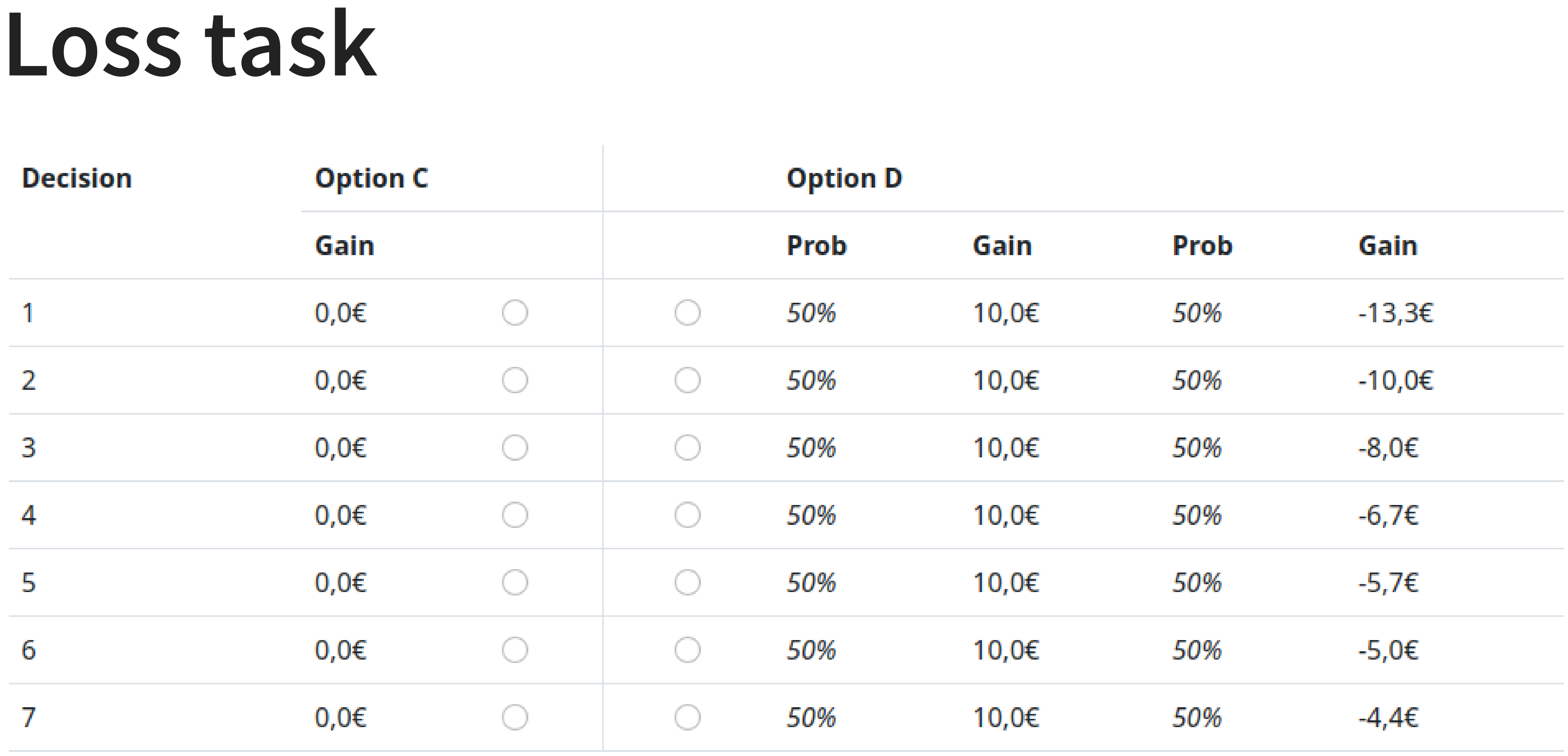
> Easy, intuitive, fine-grained + **loss aversion**

- **BRET** (intuitive, fine-grained, “game”)
- **HL** (more complex, “standard”)
- **Investment game** (intutive, fine-grained, investment)
- **Loss** (we need a measure of loss aversion)

Loss task

> Estimating λ

- estimate r from first three tasks
- assume $r^+ = r^-$
- provide a price list that identifies λ



Loss task

Decision	Option C			Option D			
	Gain			Prob	Gain	Prob	Gain
1	0,0€	○	○	*50%*	10,0€	*50%*	-13,3€
2	0,0€	○	○	*50%*	10,0€	*50%*	-10,0€
3	0,0€	○	○	*50%*	10,0€	*50%*	-8,0€
4	0,0€	○	○	*50%*	10,0€	*50%*	-6,7€
5	0,0€	○	○	*50%*	10,0€	*50%*	-5,7€
6	0,0€	○	○	*50%*	10,0€	*50%*	-5,0€
7	0,0€	○	○	*50%*	10,0€	*50%*	-4,4€

DRM

> List of all the *active* decisions under risk of the day

Filled from 6pm to midnight, every day:

- **Domains**: health, safety, recreation, driving, financial, ethical, social
- **Do vs Avoid**: *not* taking risk is a decision under risk!

DRM

> List of all the *active* decisions under risk of the day

For each activity:

- category
- importance of positive (1..10) and negative (1..-10) consequences
- likelihood of positive and negative consequences (0..100%)
- perception of the risk avoided or taken when deciding (-10..0..10)

Privacy and credibility concerns

> We need our subjects to tell us the **truth**. That’s not easy.

- subjects know only *one* person will read their raw descriptions
- that person does not and will never know their identity
- data cleaned from any identifying element by that person
- then shipped to 4 **external judges** for rating
- judge ratings, not subject’s description, will be used & released

External judges

> We hire 4 “judges” to rate the overall risk taking by subjects

- for each activity, they fill the same questions as subjects
  - is the activity risky?
  - category
  - risk avoided/taken (-10..0..10)
- judges are paid lump-sum for their work
- 4 PhD students in economics, across France

(pre-registered) Hypotheses

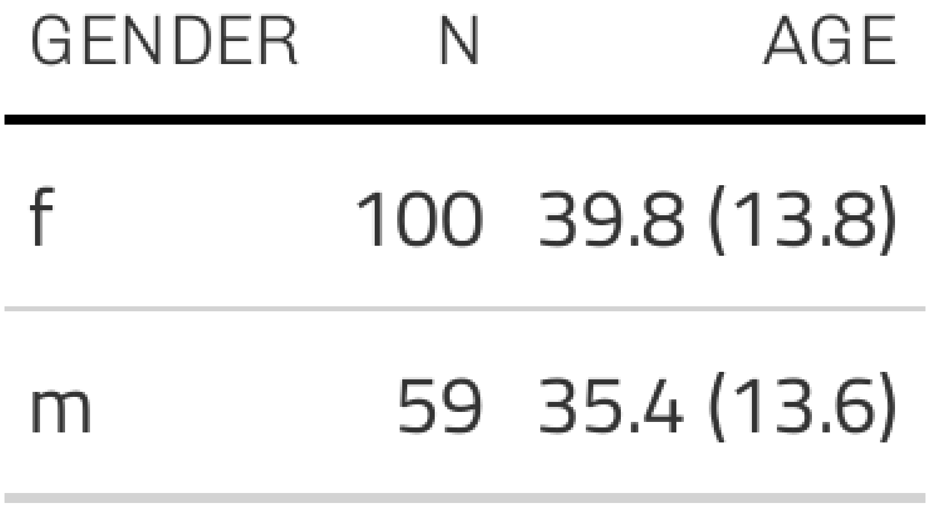
> Measurement error **plays** a role: as we aggregate *more*…

- Test-retest reliability **up**
- Between-tasks convergence **up**
- Between-questionnaires convergence **up**
- Task-questionnaire convergence **up**
- External validity (against DRM) **up**

Results

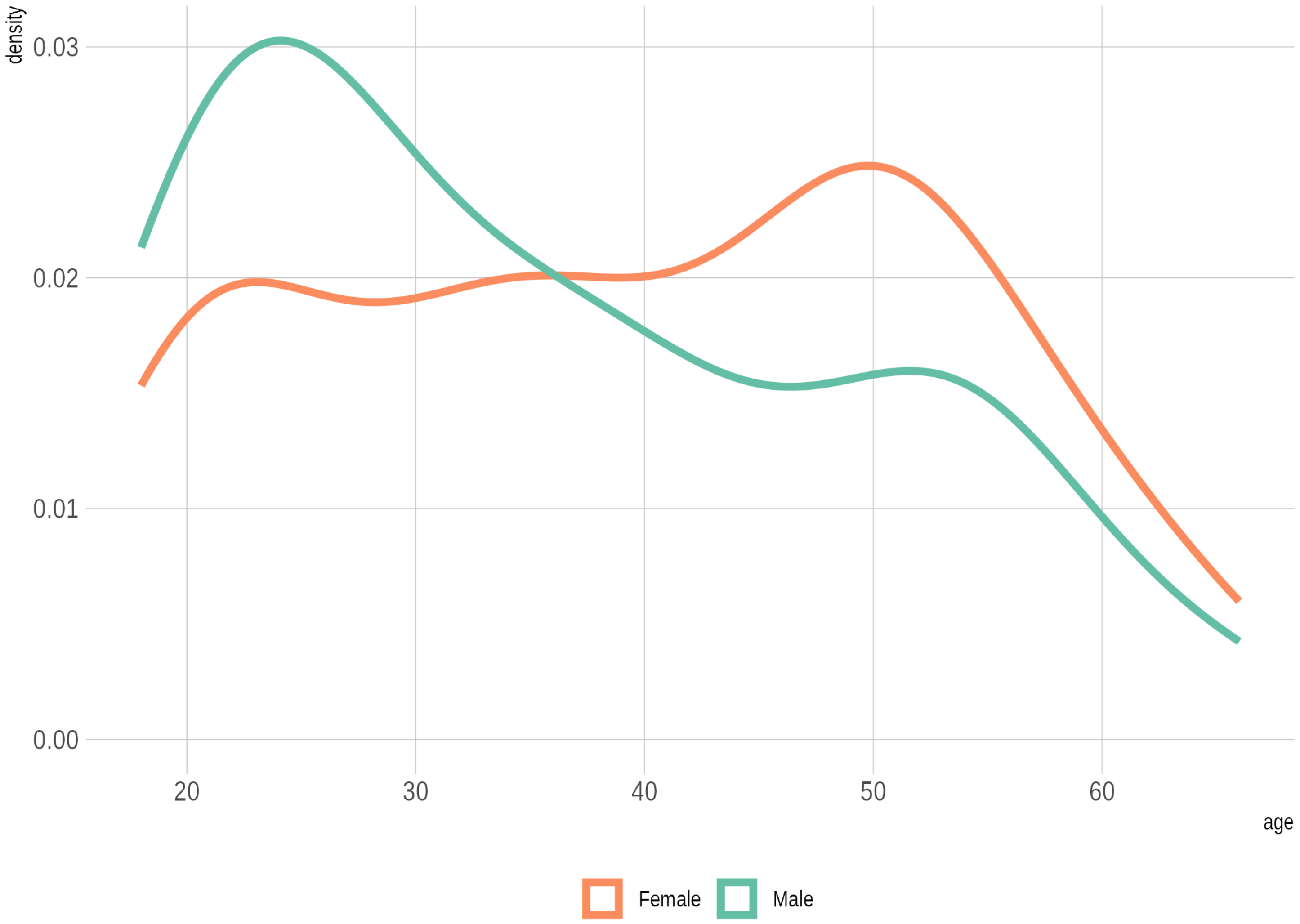
Sample

- **161** subjects over 5 sessions
- **General** population
- Very **low** attrition over 14 days: **2 dropout / 161**
- Average payment ~**77€** (for the 14 days)



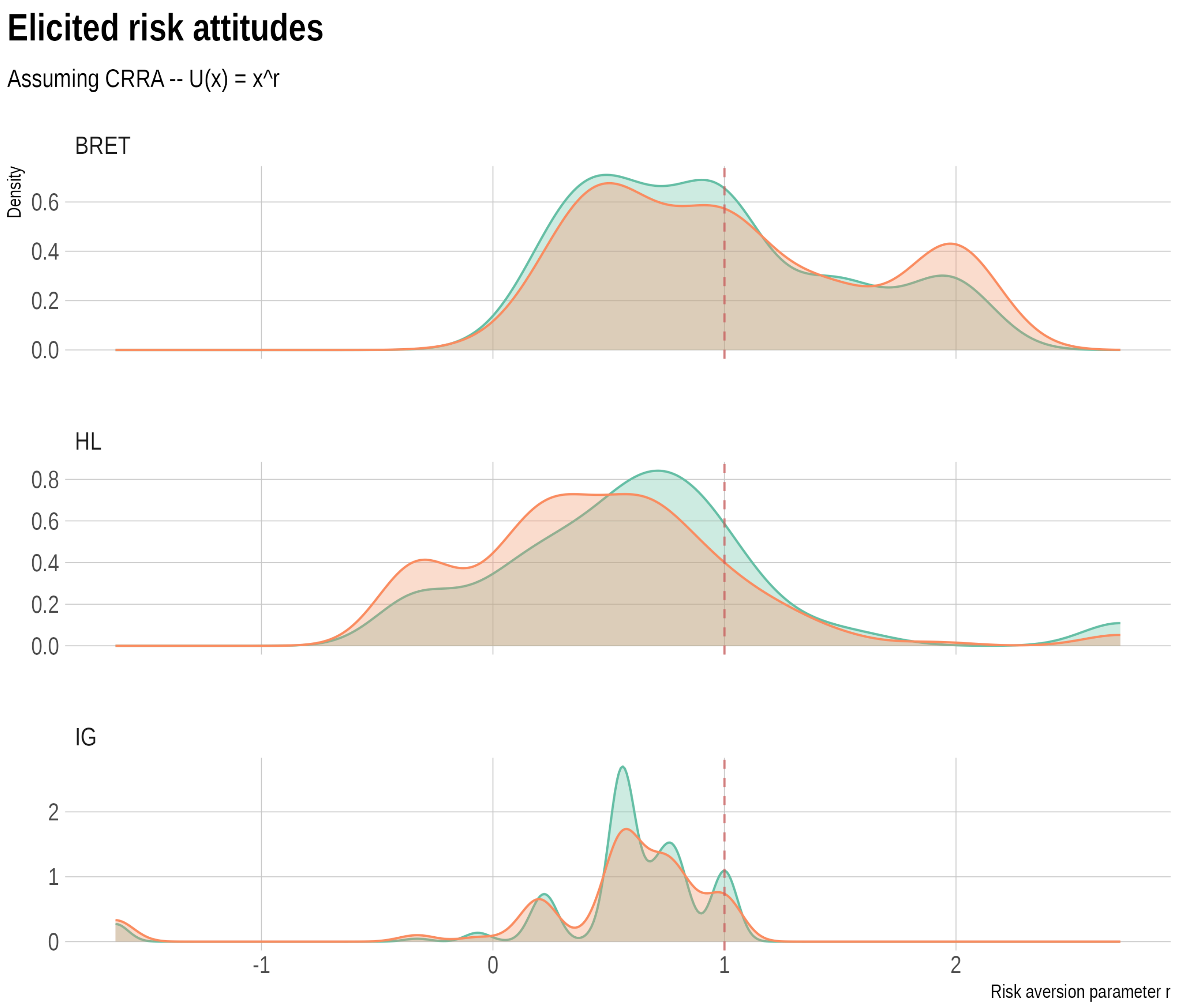
GENDER	N	AGE
f	100	39.8 (13.8)
m	59	35.4 (13.6)

Age distribution by gender

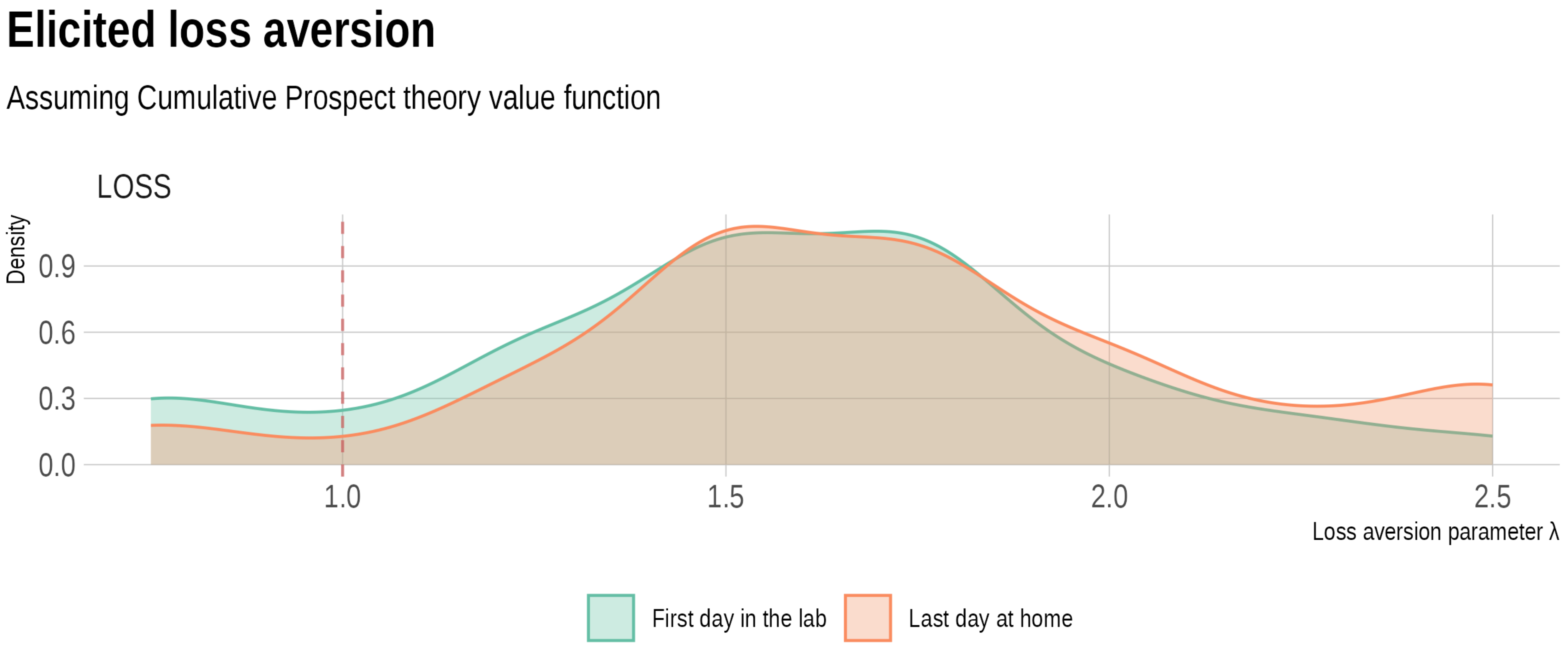


Test-retest

Elicited risk attitudes



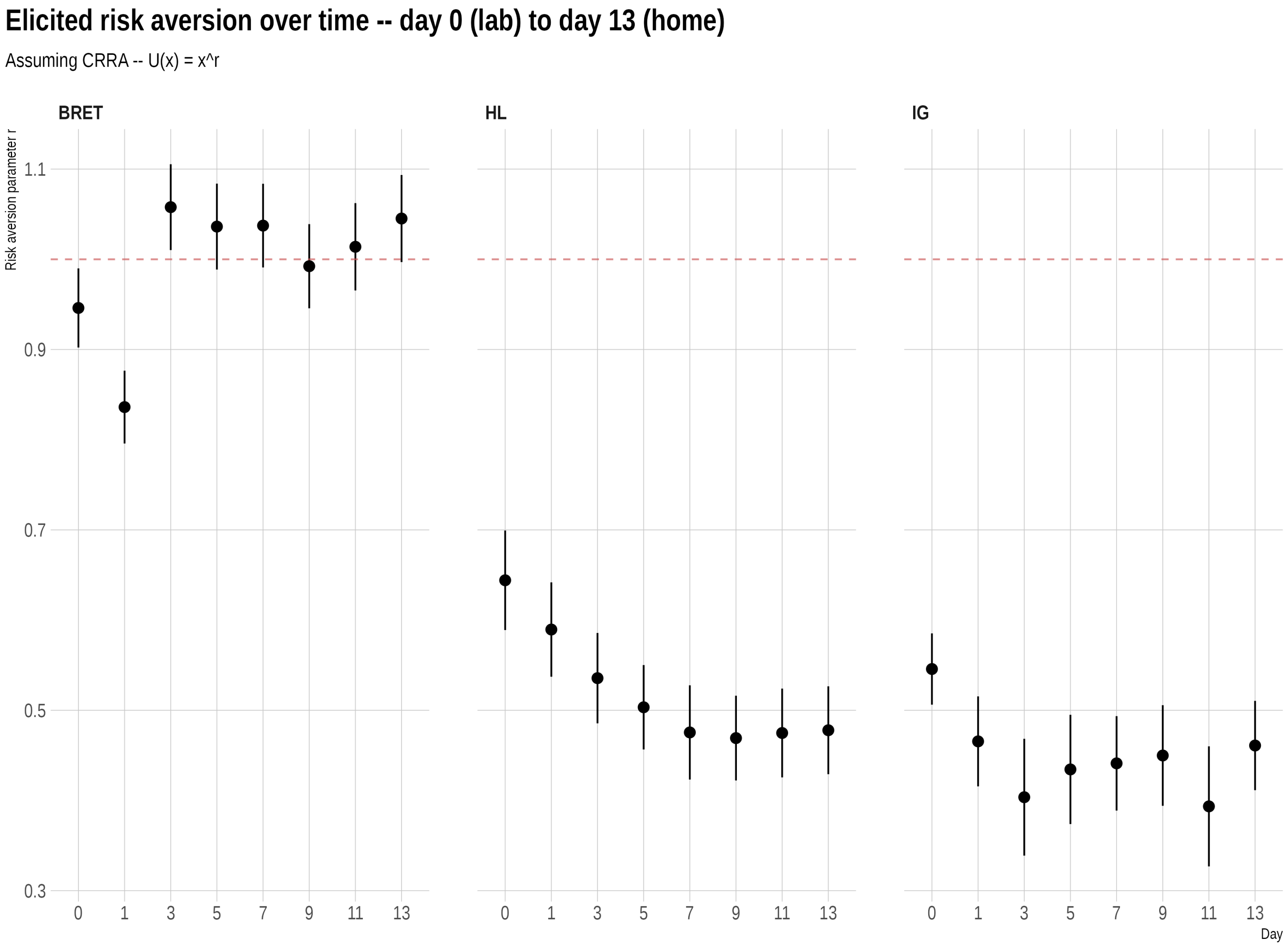
Elicited loss attitudes



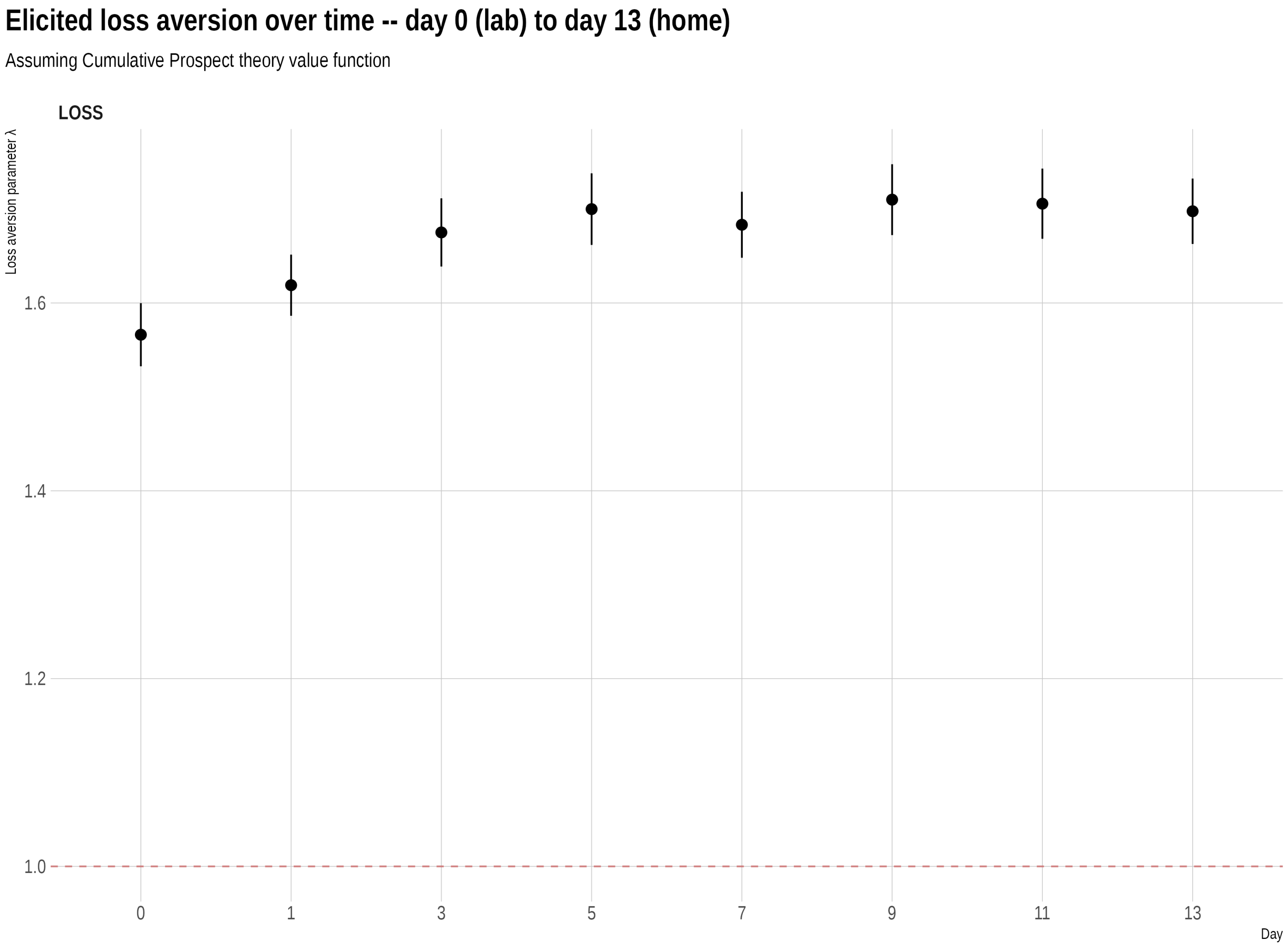
Elicited loss aversion

Assuming Cumulative Prospect theory value function

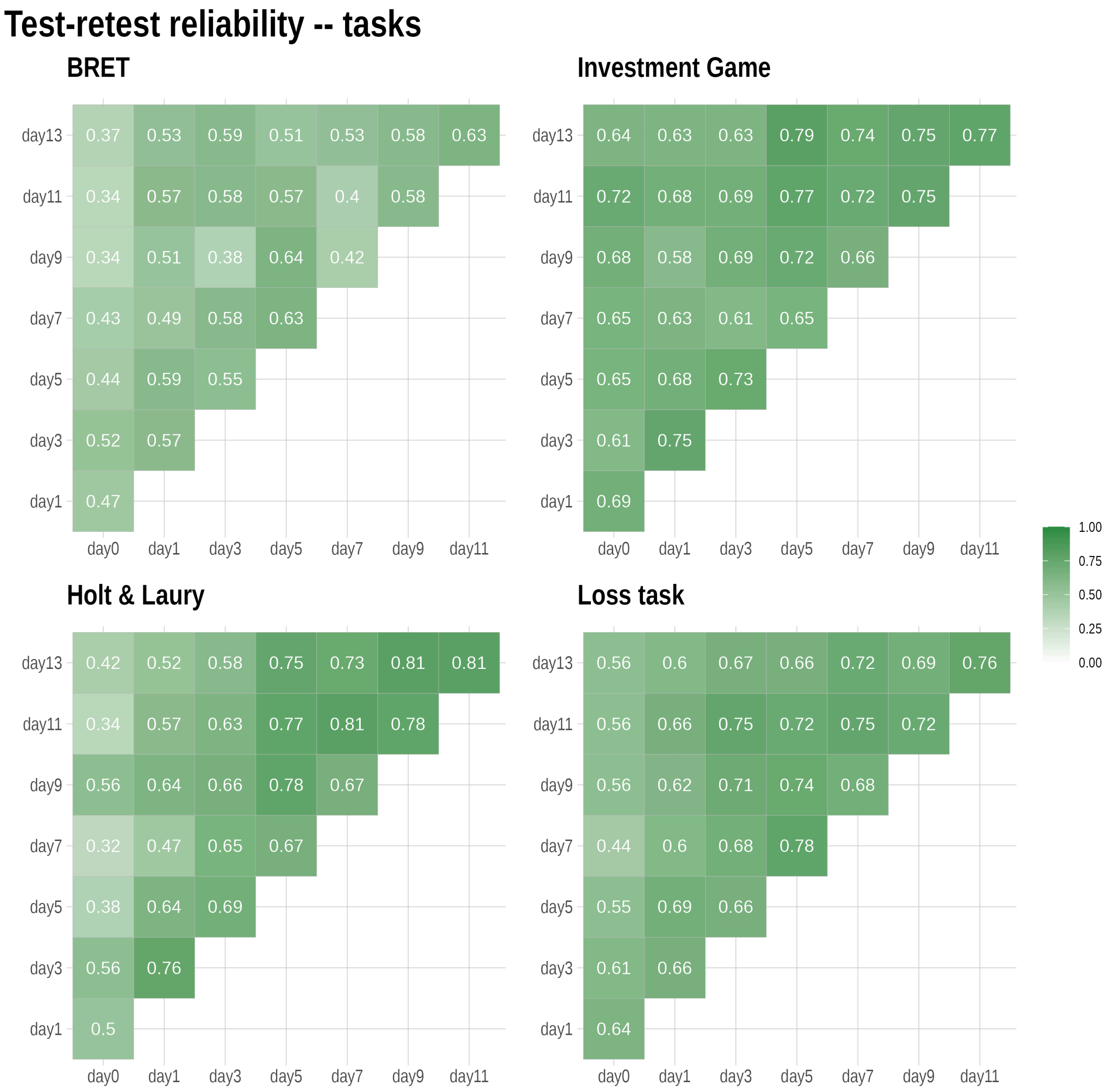
Evolution over 14 days: risk



Evolution over 14 days: loss



Test-retest reliability: tasks



Test-retest reliability -- tasks

BRET

	day0	day1	day3	day5	day7	day9	day11
day13	0.37	0.53	0.59	0.51	0.53	0.58	0.63
day11	0.34	0.57	0.58	0.57	0.4	0.58	
day9	0.34	0.51	0.38	0.64	0.42		
day7	0.43	0.49	0.58	0.63			
day5	0.44	0.59	0.55				
day3	0.52	0.57					
day1	0.47						

Investment Game

	day0	day1	day3	day5	day7	day9	day11
day13	0.64	0.63	0.63	0.79	0.74	0.75	0.77
day11	0.72	0.68	0.69	0.77	0.72	0.75	
day9	0.68	0.58	0.69	0.72	0.66		
day7	0.65	0.63	0.61	0.65			
day5	0.65	0.68	0.73				
day3	0.61	0.75					
day1	0.69						

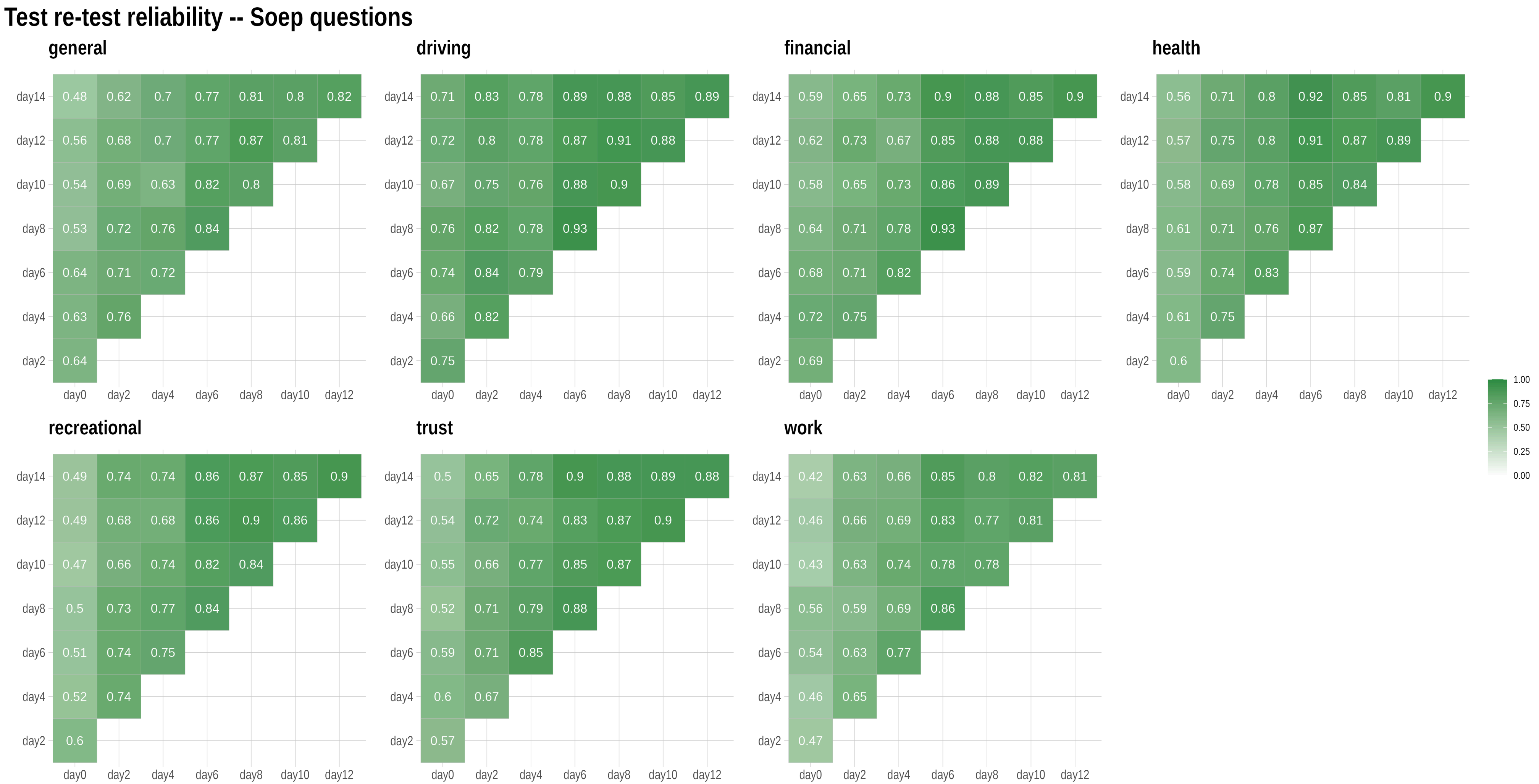
Holt & Laury

	day0	day1	day3	day5	day7	day9	day11
day13	0.42	0.52	0.58	0.75	0.73	0.81	0.81
day11	0.34	0.57	0.63	0.77	0.81	0.78	
day9	0.56	0.64	0.66	0.78	0.67		
day7	0.32	0.47	0.65	0.67			
day5	0.38	0.64	0.69				
day3	0.56	0.76					
day1	0.5						

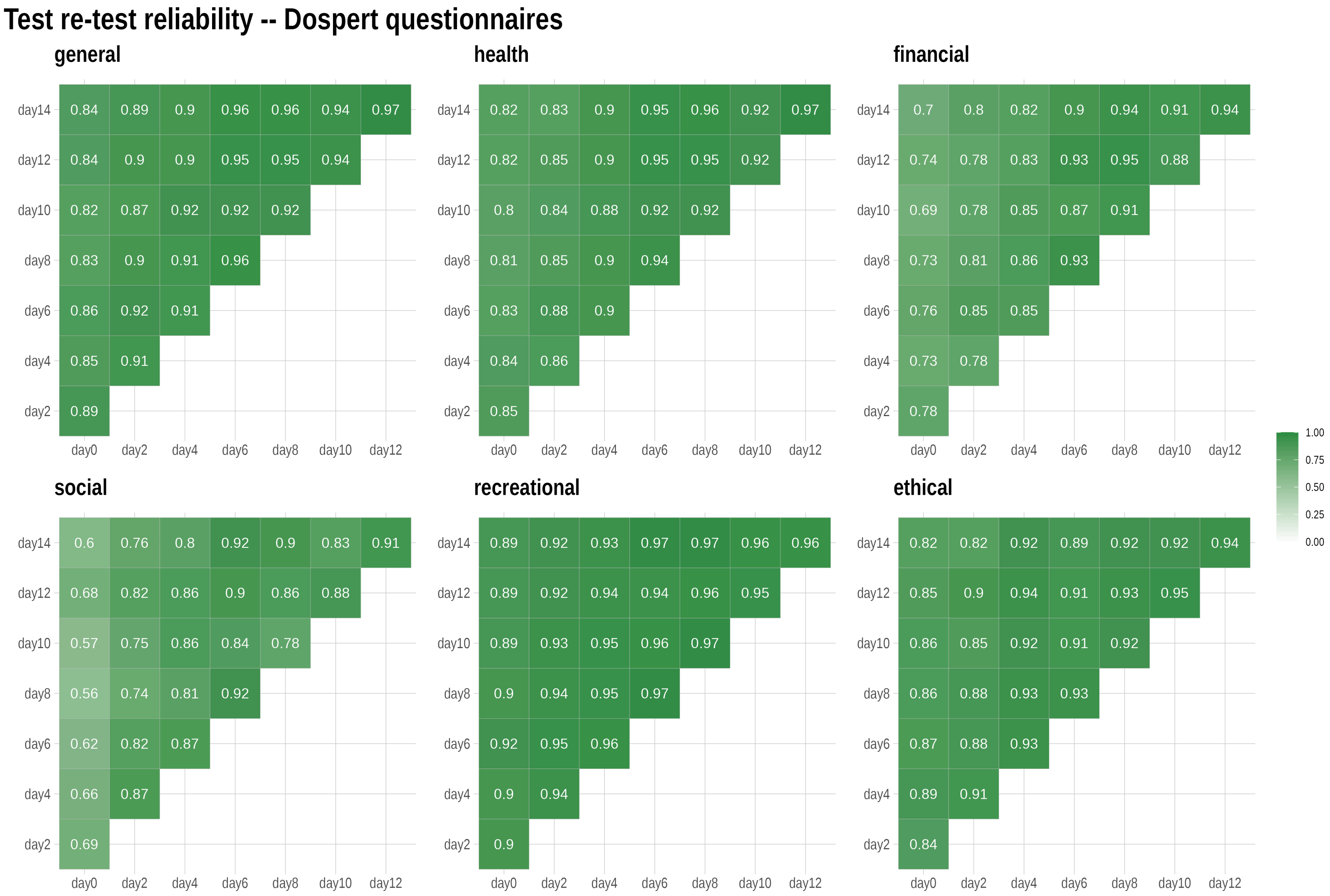
Loss task

	day0	day1	day3	day5	day7	day9	day11
day13	0.56	0.6	0.67	0.66	0.72	0.69	0.76
day11	0.56	0.66	0.75	0.72	0.75	0.72	
day9	0.56	0.62	0.71	0.74	0.68		
day7	0.44	0.6	0.68	0.78			
day5	0.55	0.69	0.66				
day3	0.61	0.66					
day1	0.64						

Test-retest reliability: SOEP



Test-retest reliability: DOSPERT



Test re-test reliability -- Dospert questionnaires

general

	day0	day2	day4	day6	day8	day10	day12
day14	0.84	0.89	0.9	0.96	0.96	0.94	0.97
day12	0.84	0.9	0.9	0.95	0.95	0.94	
day10	0.82	0.87	0.92	0.92	0.92		
day8	0.83	0.9	0.91	0.96			
day6	0.86	0.92	0.91				
day4	0.85	0.91					
day2	0.89						

health

	day0	day2	day4	day6	day8	day10	day12
day14	0.82	0.83	0.9	0.95	0.96	0.92	0.97
day12	0.82	0.85	0.9	0.95	0.95	0.92	
day10	0.8	0.84	0.88	0.92	0.92		
day8	0.81	0.85	0.9	0.94			
day6	0.83	0.88	0.9				
day4	0.84	0.86					
day2	0.85						

financial

	day0	day2	day4	day6	day8	day10	day12
day14	0.7	0.8	0.82	0.9	0.94	0.91	0.94
day12	0.74	0.78	0.83	0.93	0.95	0.88	
day10	0.69	0.78	0.85	0.87	0.91		
day8	0.73	0.81	0.86	0.93			
day6	0.76	0.85	0.85				
day4	0.73	0.78					
day2	0.78						

social

	day0	day2	day4	day6	day8	day10	day12
day14	0.6	0.76	0.8	0.92	0.9	0.83	0.91
day12	0.68	0.82	0.86	0.9	0.86	0.88	
day10	0.57	0.75	0.86	0.84	0.78		
day8	0.56	0.74	0.81	0.92			
day6	0.62	0.82	0.87				
day4	0.66	0.87					
day2	0.69						

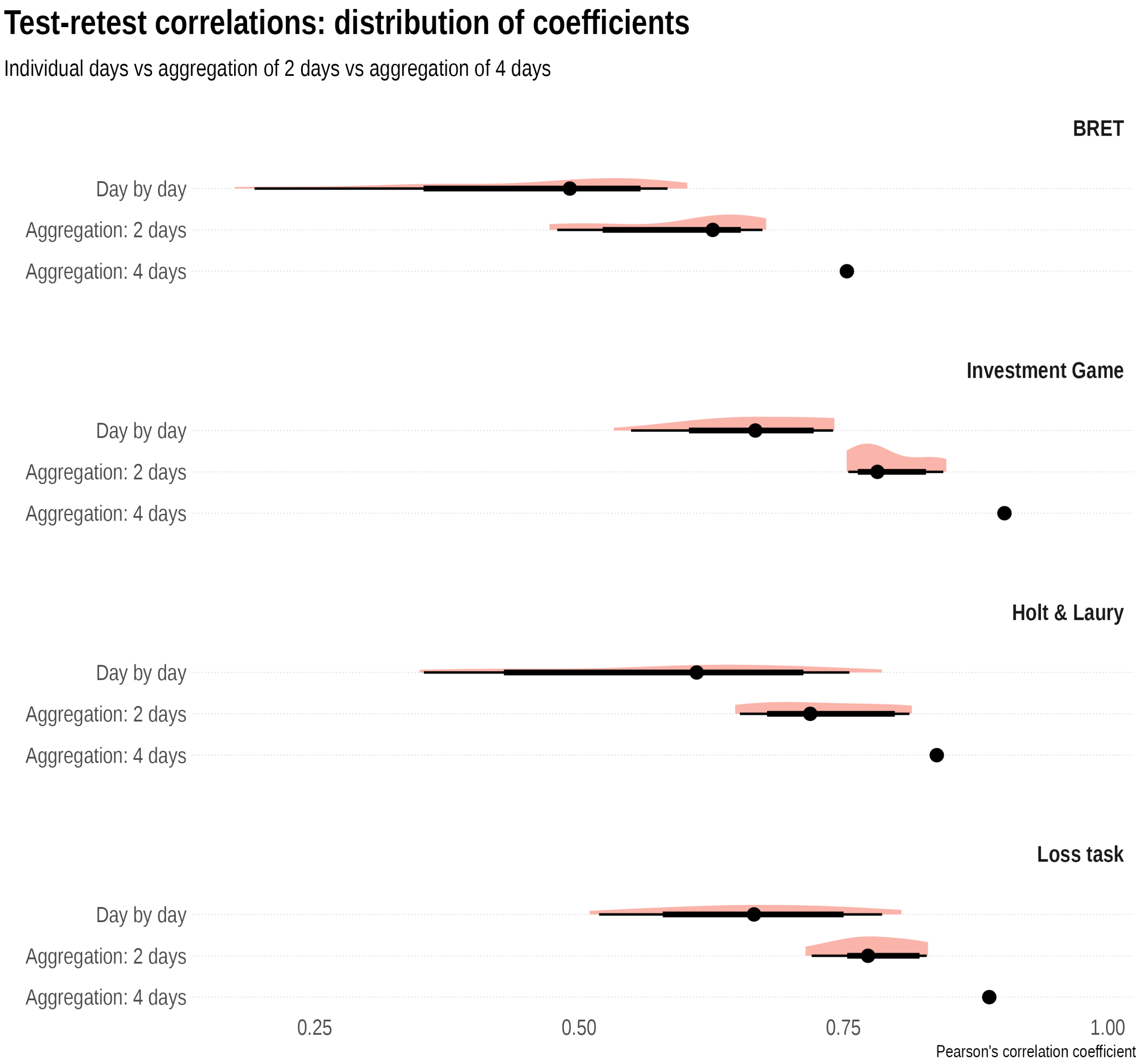
recreational

	day0	day2	day4	day6	day8	day10	day12
day14	0.89	0.92	0.93	0.97	0.97	0.96	0.96
day12	0.89	0.92	0.94	0.94	0.96	0.95	
day10	0.89	0.93	0.95	0.96	0.97		
day8	0.9	0.94	0.95	0.97			
day6	0.92	0.95	0.96				
day4	0.9	0.94					
day2	0.9						

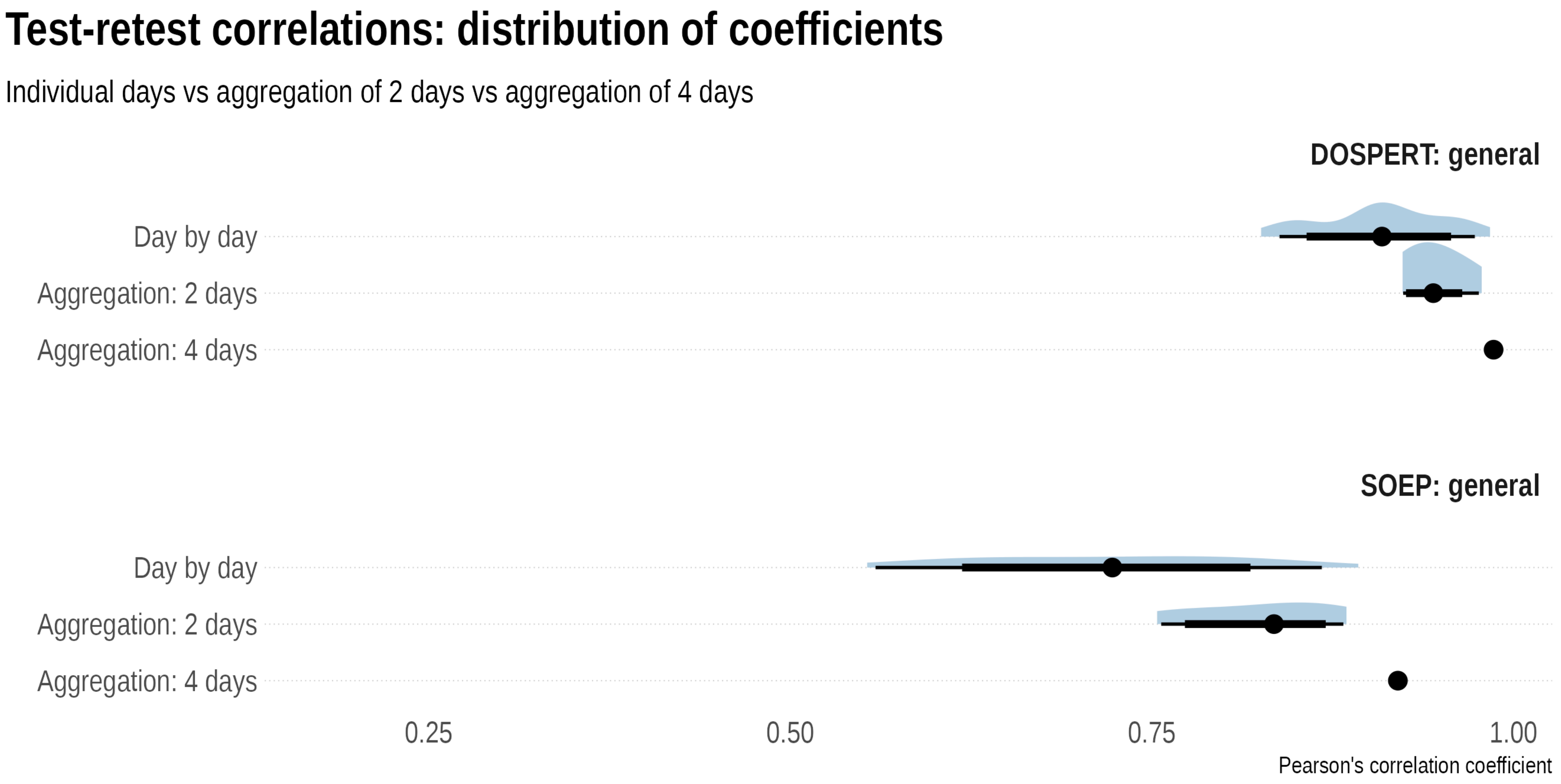
ethical

	day0	day2	day4	day6	day8	day10	day12
day14	0.82	0.82	0.92	0.89	0.92	0.92	0.94
day12	0.85	0.9	0.94	0.91	0.93	0.95	
day10	0.86	0.85	0.92	0.91	0.92		
day8	0.86	0.88	0.93	0.93			
day6	0.87	0.88	0.93				
day4	0.89	0.91					
day2	0.84						

Test-retest: aggregating tasks

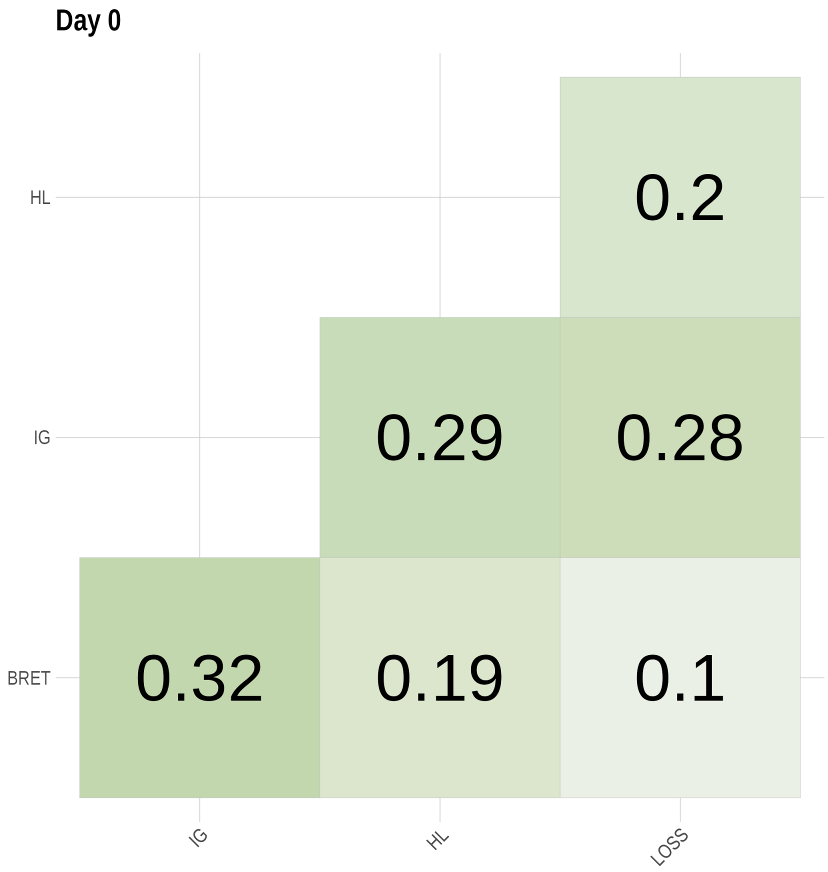


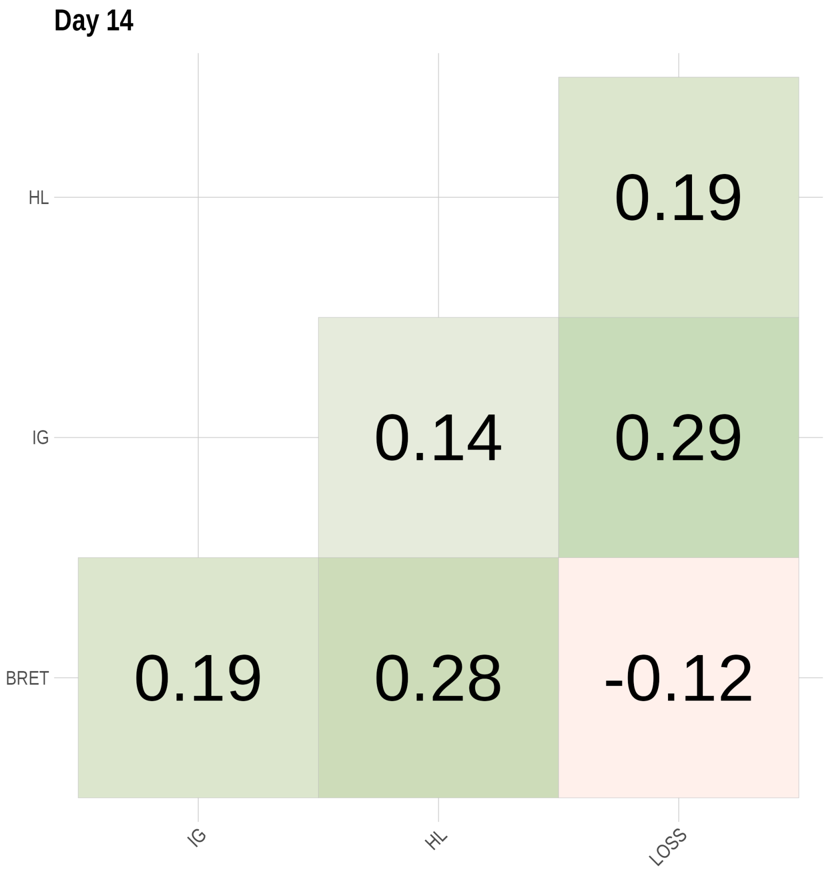
Test-retest: aggregating questionnaires



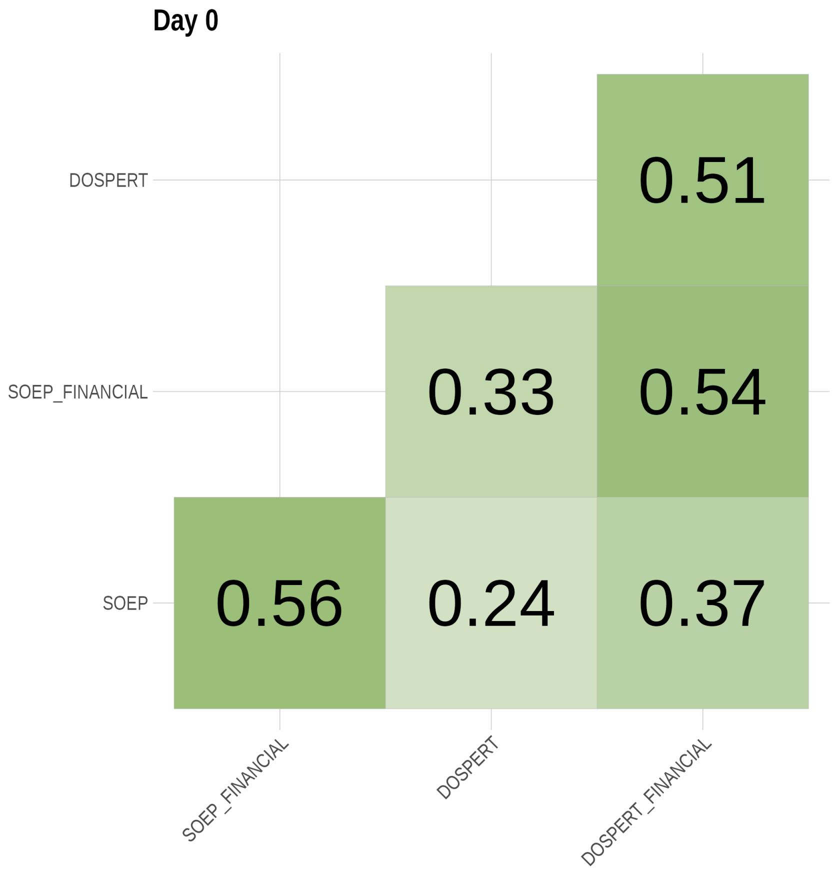
Convergent validity

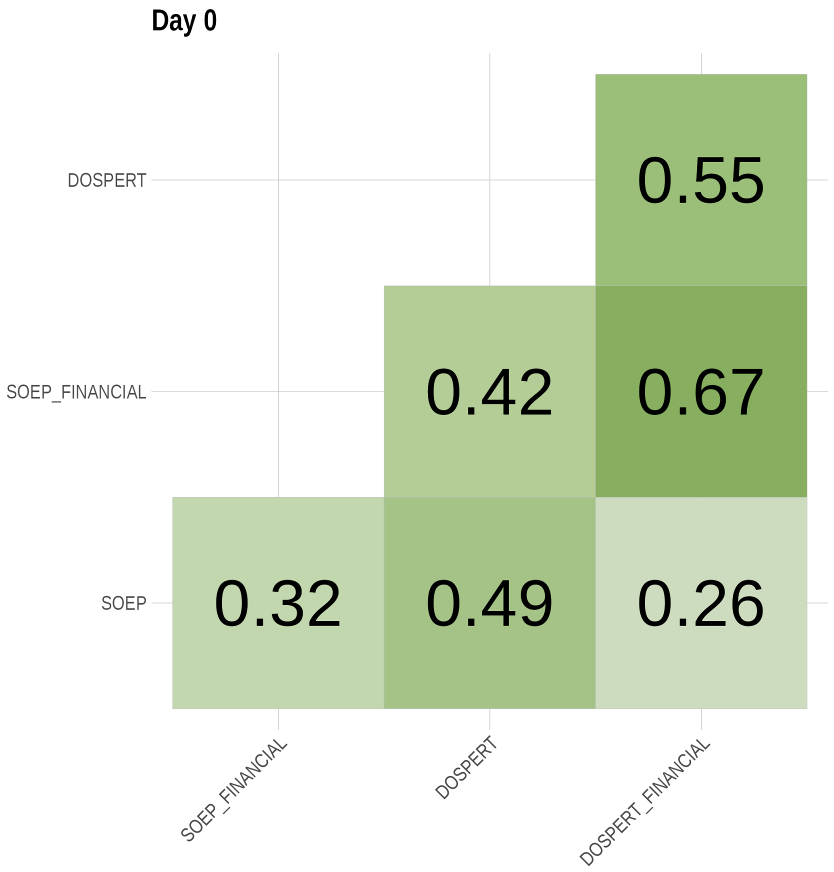
Convergence: tasks


Day 0

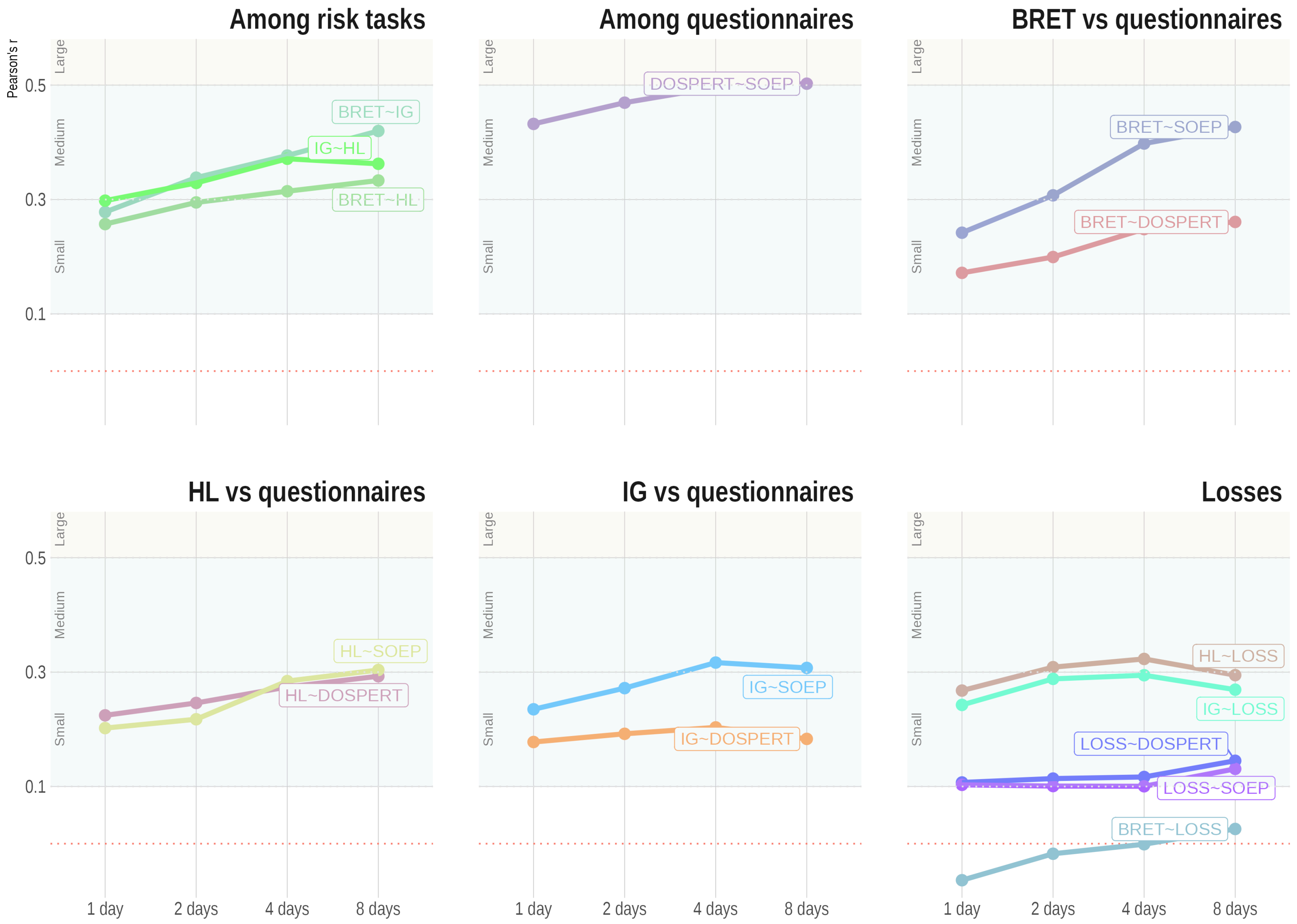

Day 14

Convergence: questionnaires





Convergence: aggregating



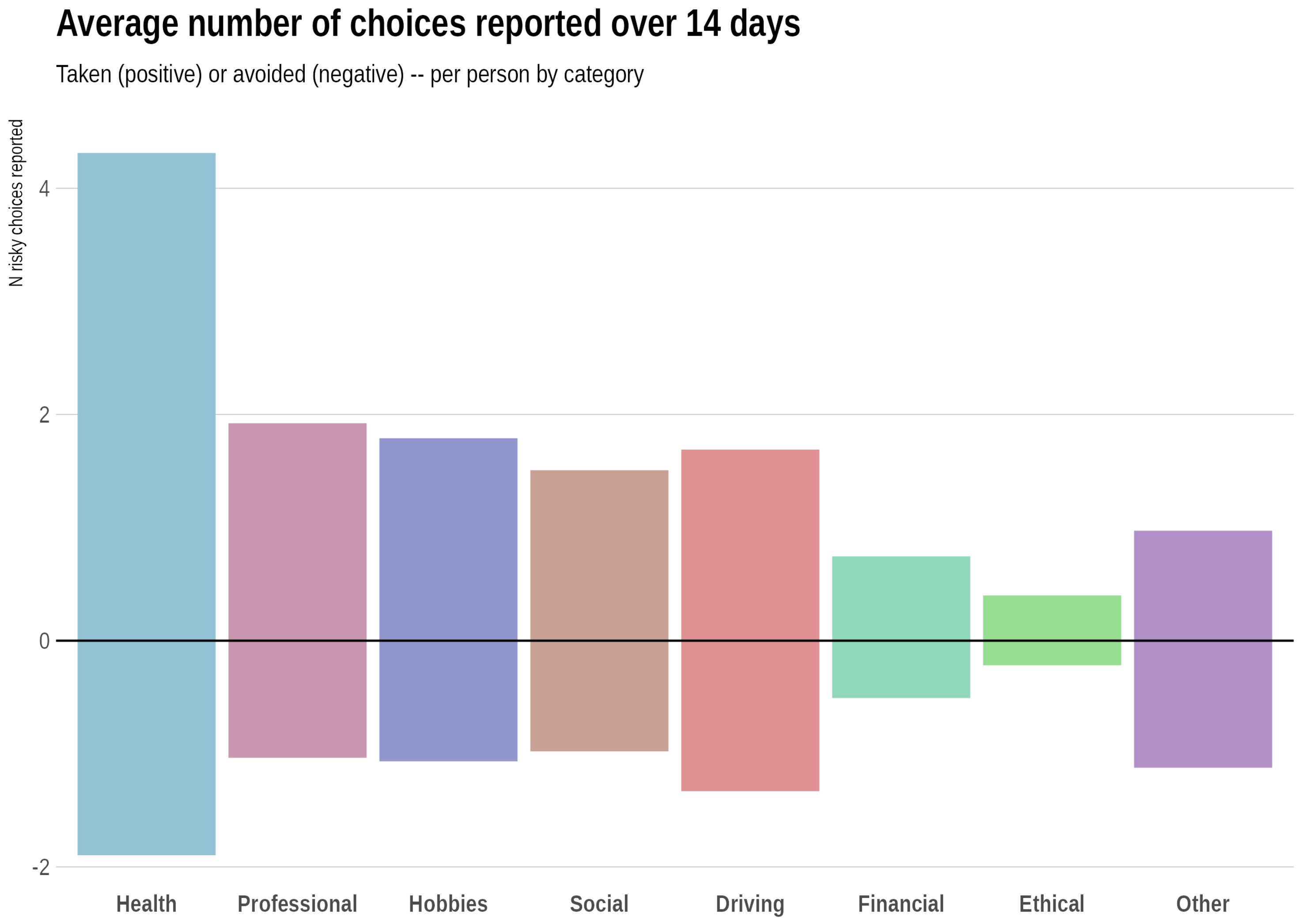
External validity: DRM

Descriptives

> 159 subjects carried out for 14 days, **3392** activities

- Subjects reported 1.52 choices involving risk per day
- 0.95 per day, they took risks – 0.57 per day, they avoided risks
- Mostly in **Health**, followed by **Driving, Professional** and **Hobbies**
- **Financial** and **Ethical** risk less frequent (<1/person over 2 weeks)

Descriptives



Measures

- **Absolute** frequency measure:

  $\sum_{d=1}^{14} N\_take_d$

- **Relative** frequency measure:

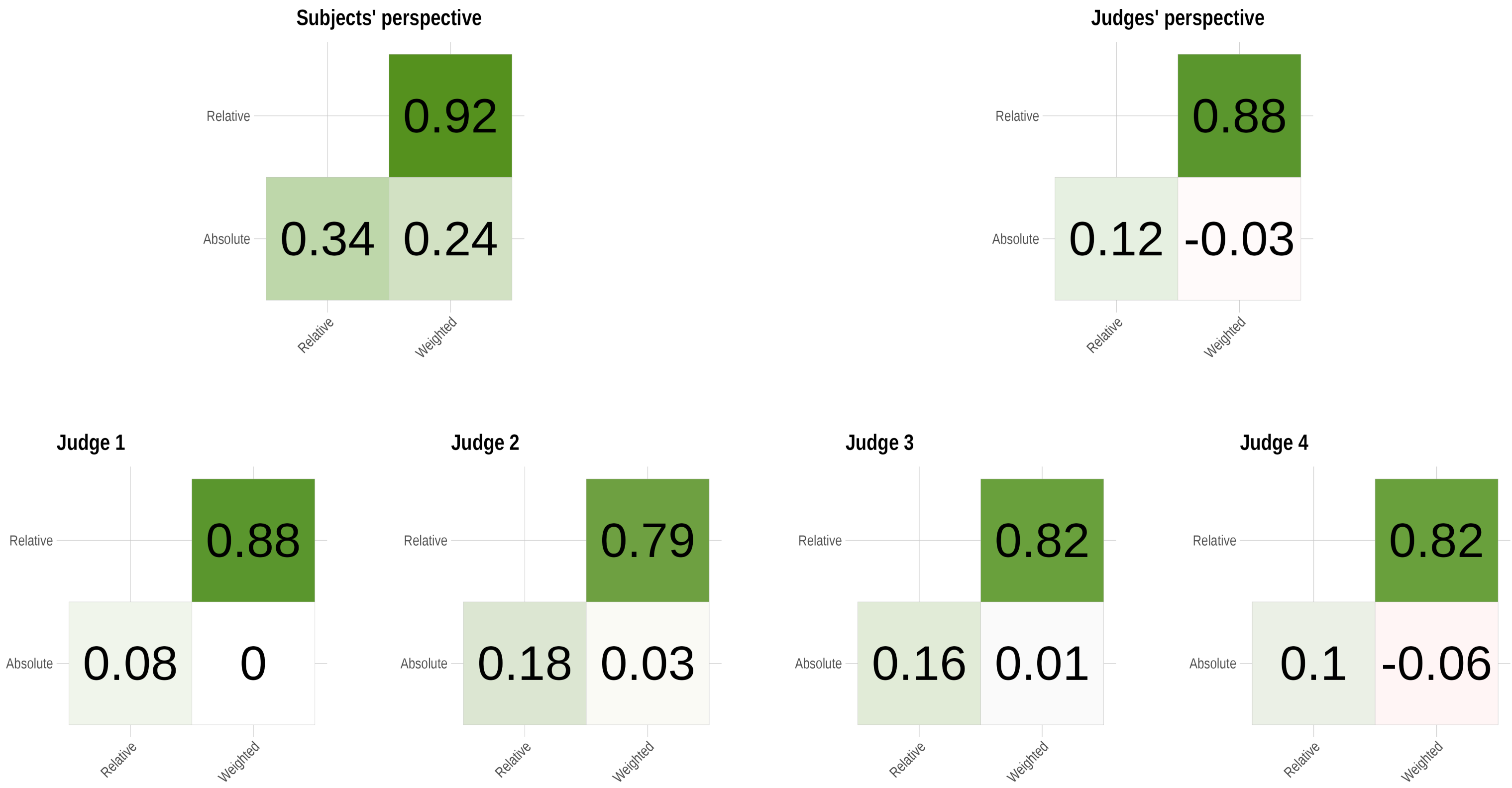
  $\sum_{d=1}^{14} \frac{N\_take_d}{N\_take_d + N\_avoid_d}$

- **Weighted** measure: average risk rating

  $\frac{1}{14} \sum_{d=1}^{14} W_d,\ w \in [-10, 10]$

Measures’ validity

Convergent validity of DRM measures

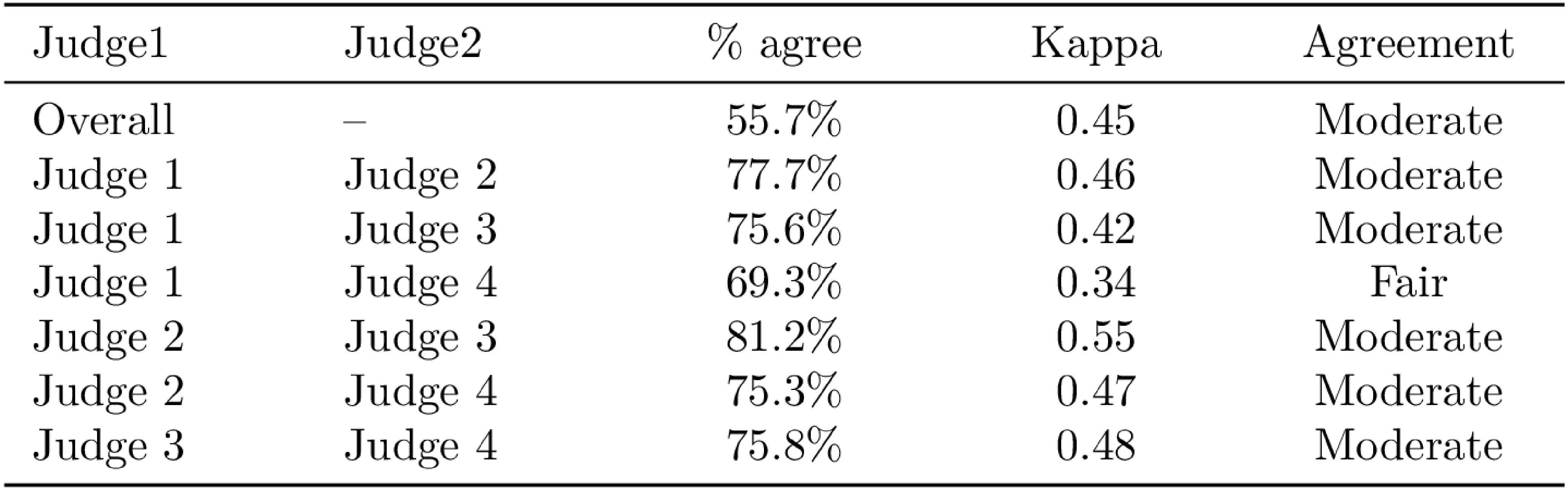


Judges have a higher bar for risk

> Subjects report **3392** choices in presence of risk…
>
> …but jdges think that many of them ***are not*** risky

- **Judge 1:** 2376/3992 (59%)
- **Judge 2:** 2411/3992 (60%)
- **Judge 3:** 2369/3992 (59%)
- **Judge 4:** 1992/3992 (49%)

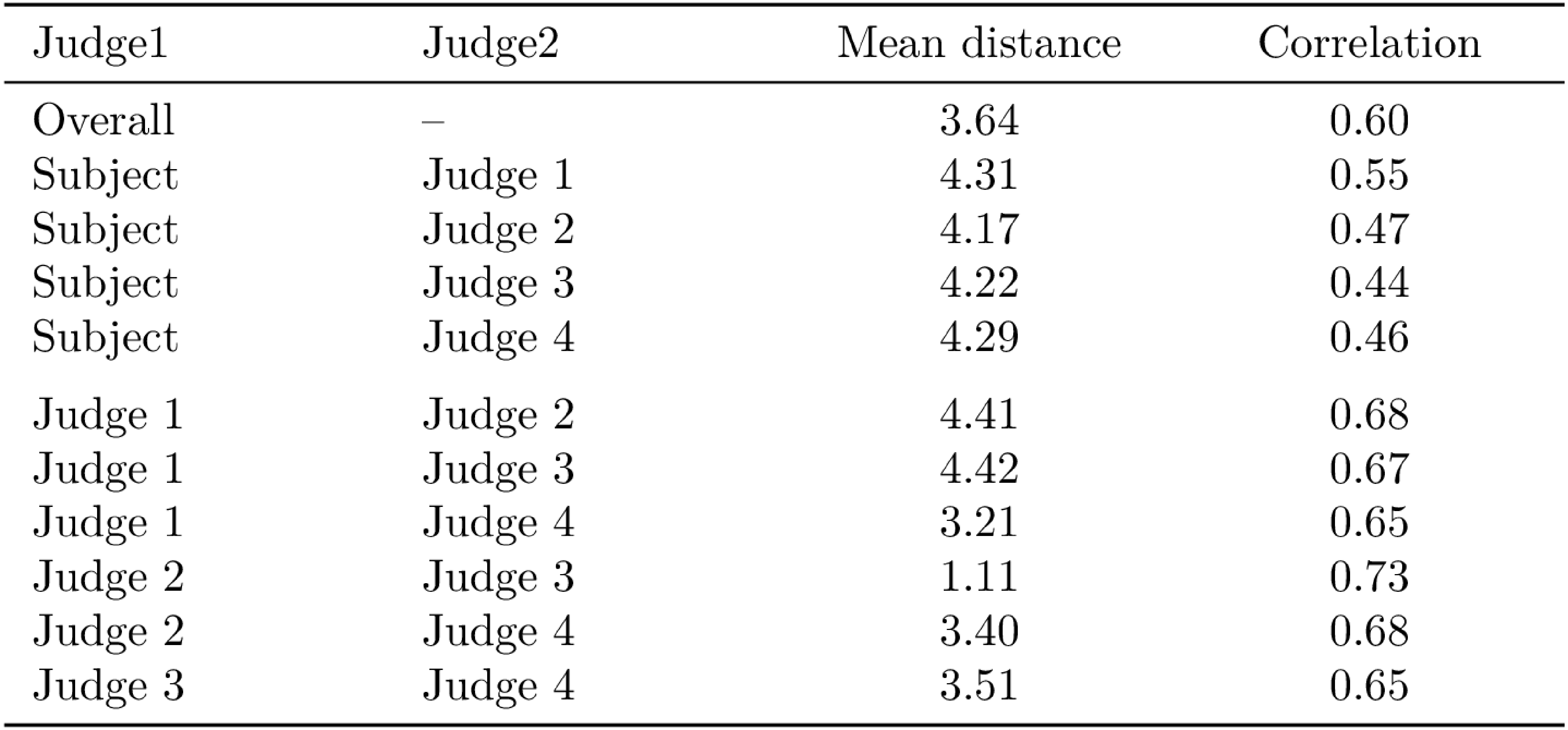
Judges’ reliability



Judge1	Judge2	% agree	Kappa	Agreement
Overall	–	55.7%	0.45	Moderate
Judge 1	Judge 2	77.7%	0.46	Moderate
Judge 1	Judge 3	75.6%	0.42	Moderate
Judge 1	Judge 4	69.3%	0.34	Fair
Judge 2	Judge 3	81.2%	0.55	Moderate
Judge 2	Judge 4	75.3%	0.47	Moderate
Judge 3	Judge 4	75.8%	0.48	Moderate

Judges’ agreement rate: is the activity risky?

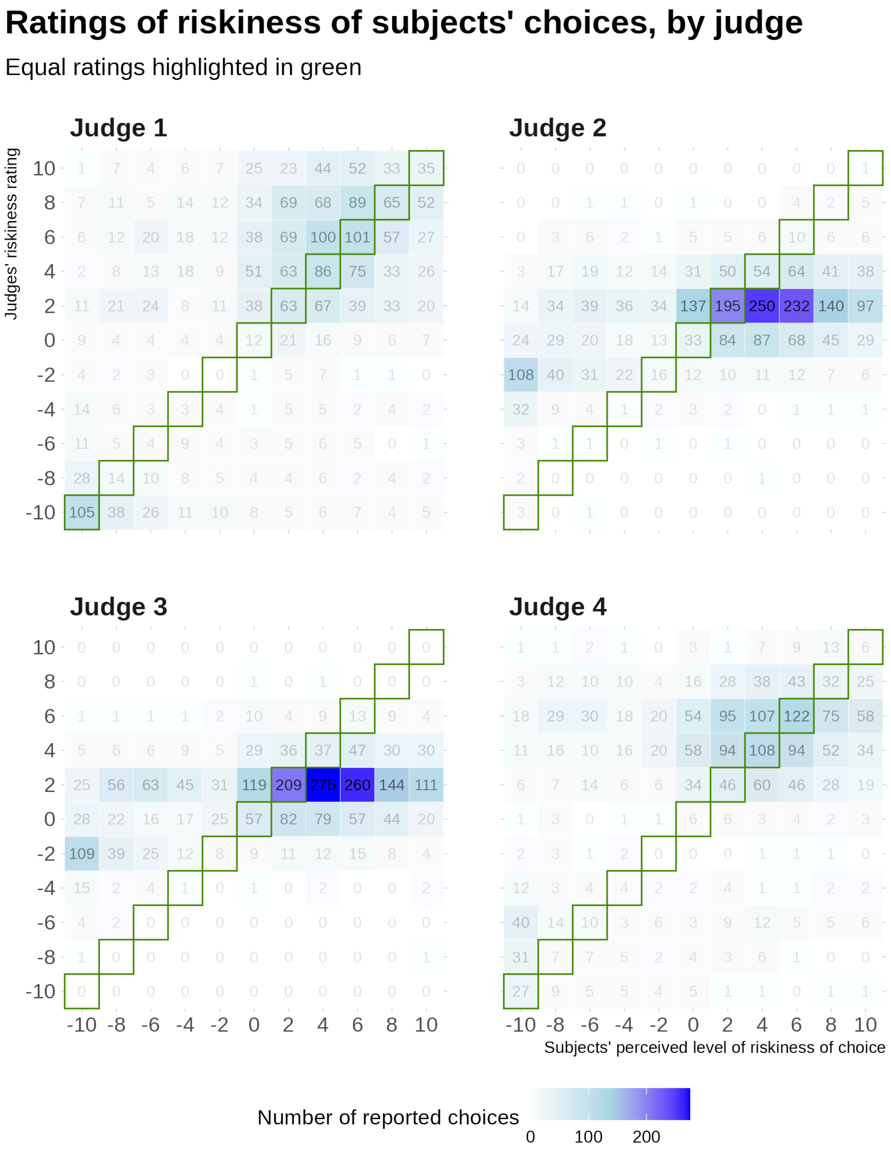
Judges’ reliability



Judge1	Judge2	Mean distance	Correlation
Overall	–	3.64	0.60
Subject	Judge 1	4.31	0.55
Subject	Judge 2	4.17	0.47
Subject	Judge 3	4.22	0.44
Subject	Judge 4	4.29	0.46
Judge 1	Judge 2	4.41	0.68
Judge 1	Judge 3	4.42	0.67
Judge 1	Judge 4	3.21	0.65
Judge 2	Judge 3	1.11	0.73
Judge 2	Judge 4	3.40	0.68
Judge 3	Judge 4	3.51	0.65

Judge’s agreement rate: how much risk was taken?

Judge’s and subject’s risk perception



Judge 1

	-10	-8	-6	-4	-2	0	2	4	6	8	10
10	1	7	4	6	7	25	23	44	52	33	35
8	7	11	5	14	12	34	69	68	89	65	52
6	6	12	20	18	12	38	69	100	101	57	27
4	2	8	13	18	9	51	63	86	75	33	26
2	11	21	24	8	11	38	63	67	39	33	20
0	9	4	4	4	4	12	21	16	9	6	7
-2	4	2	3	0	0	1	5	7	1	1	0
-4	14	6	3	3	4	1	5	5	2	4	2
-6	11	5	4	9	4	3	5	6	5	0	1
-8	28	14	10	8	5	4	4	6	2	4	2
-10	105	38	26	11	10	8	5	6	7	4	5

Judge 2

	-10	-8	-6	-4	-2	0	2	4	6	8	10
10	0	0	0	0	0	0	0	0	0	0	1
8	0	0	1	1	0	1	0	0	4	2	5
6	0	3	6	2	1	5	5	6	10	6	6
4	3	17	19	12	14	31	50	54	64	41	38
2	14	34	39	36	34	137	195	250	232	140	97
0	24	29	20	18	13	33	84	87	68	45	29
-2	108	40	31	22	16	12	10	11	12	7	6
-4	32	9	4	1	2	3	2	0	1	1	1
-6	3	1	1	0	1	0	1	0	0	0	0
-8	2	0	0	0	0	0	0	1	0	0	0
-10	3	0	1	0	0	0	0	0	0	0	0

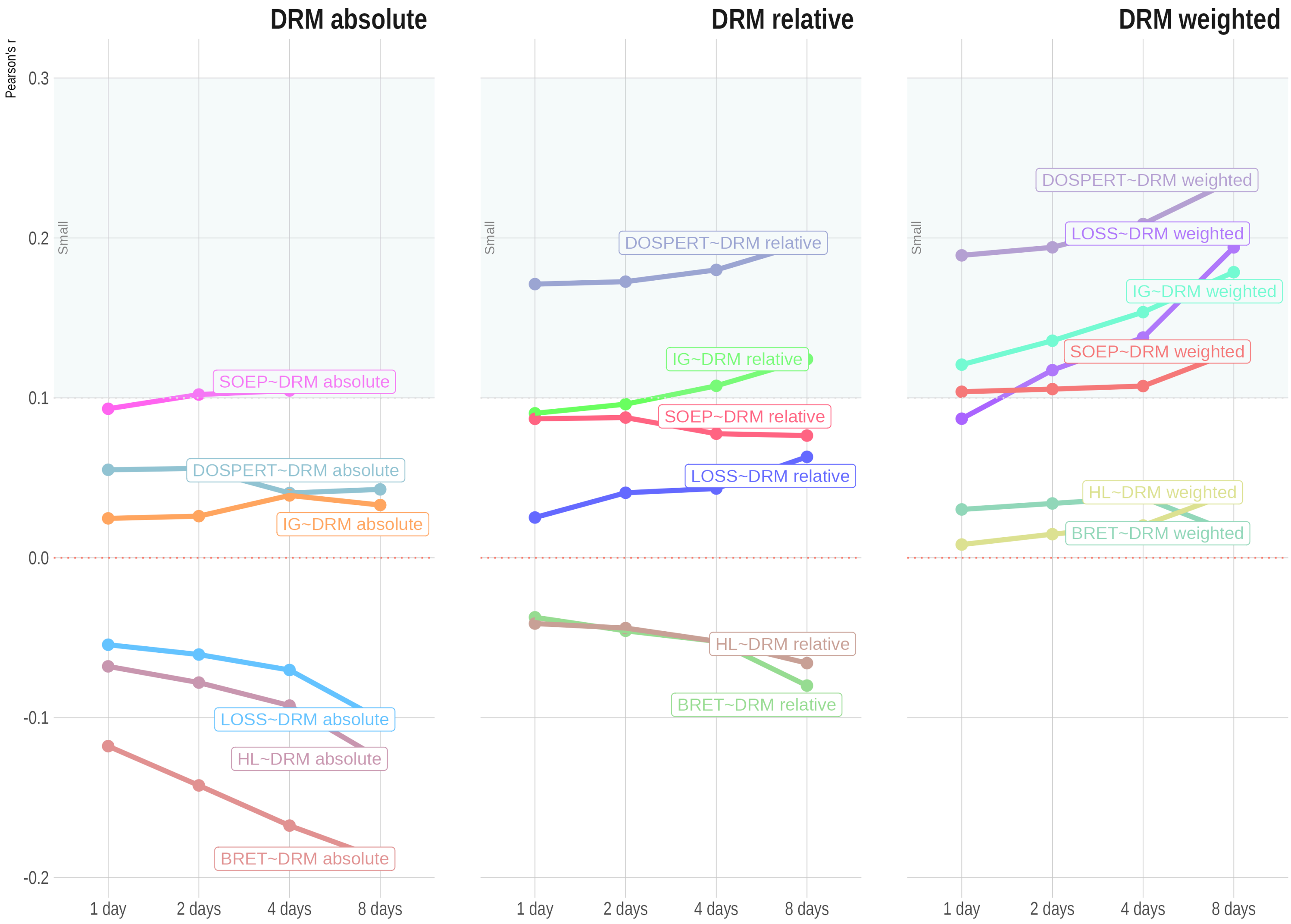
Judge 3

	-10	-8	-6	-4	-2	0	2	4	6	8	10
10	0	0	0	0	0	0	0	0	0	0	0
8	0	0	0	0	0	1	0	1	0	0	0
6	1	1	1	1	2	10	4	9	13	9	4
4	5	6	6	9	5	29	36	37	47	30	30
2	25	56	63	45	31	119	209	275	260	144	111
0	28	22	16	17	25	57	82	79	57	44	20
-2	109	39	25	12	8	9	11	12	15	8	4
-4	15	2	4	1	0	1	0	2	0	0	2
-6	4	2	0	0	0	0	0	0	0	0	0
-8	1	0	0	0	0	0	0	0	0	0	1
-10	0	0	0	0	0	0	0	0	0	0	0

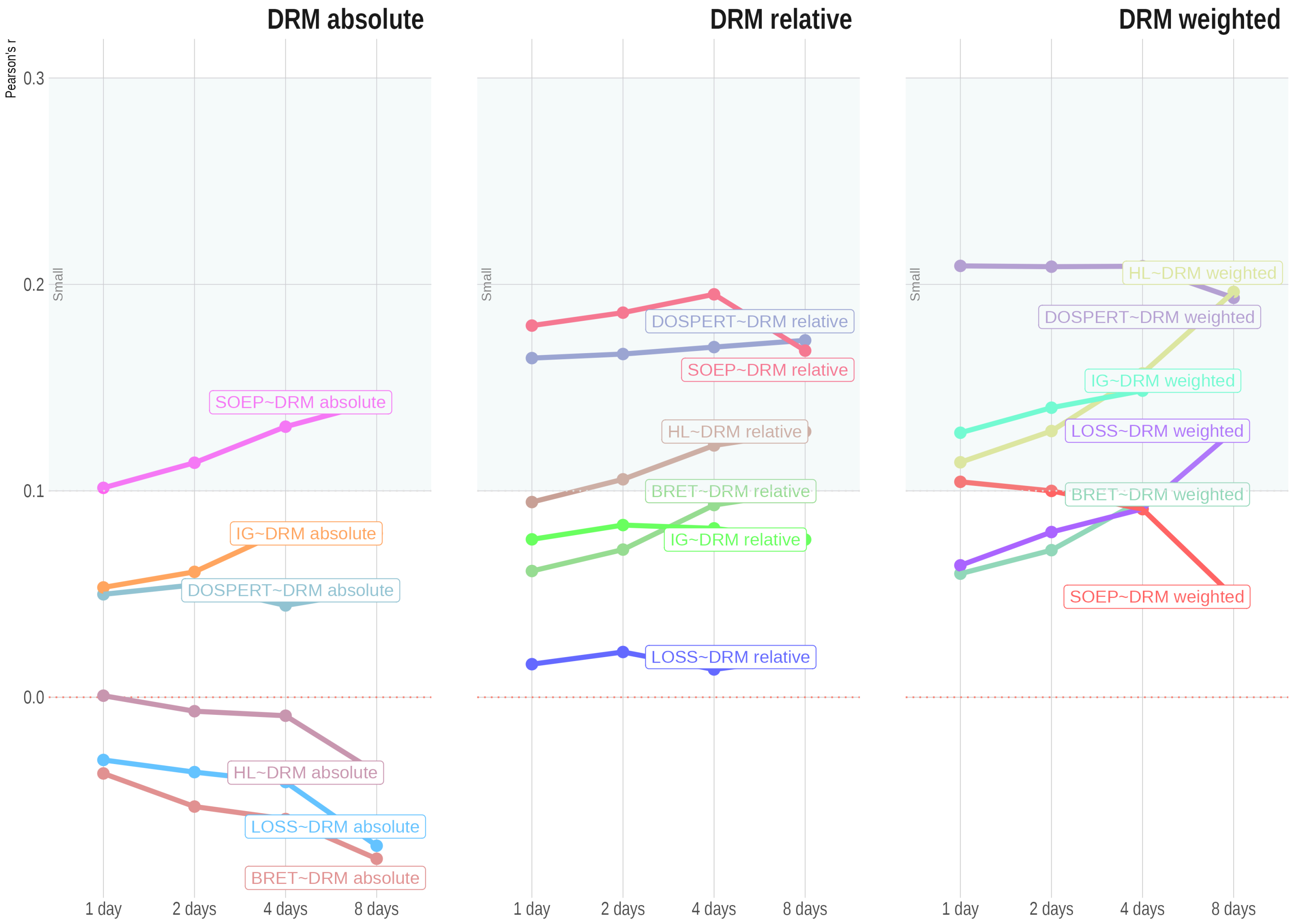
Judge 4

	-10	-8	-6	-4	-2	0	2	4	6	8	10
10	1	1	2	1	0	3	1	7	9	13	6
8	3	12	10	10	4	16	28	38	43	32	25
6	18	29	30	18	20	54	95	107	122	75	58
4	11	16	10	16	20	58	94	108	94	52	34
2	6	7	14	6	6	34	46	60	46	28	19
0	1	3	0	1	1	6	6	3	4	2	3
-2	2	3	1	2	0	0	0	1	1	1	0
-4	12	3	4	4	2	2	4	1	1	2	2
-6	40	14	10	3	6	3	9	12	5	5	6
-8	31	7	7	5	2	4	3	6	1	0	0
-10	27	9	5	5	4	5	1	1	0	1	1

External validity: aggregating (subj)



External validity: aggregating (judges)



Discussion

Not only measurement error

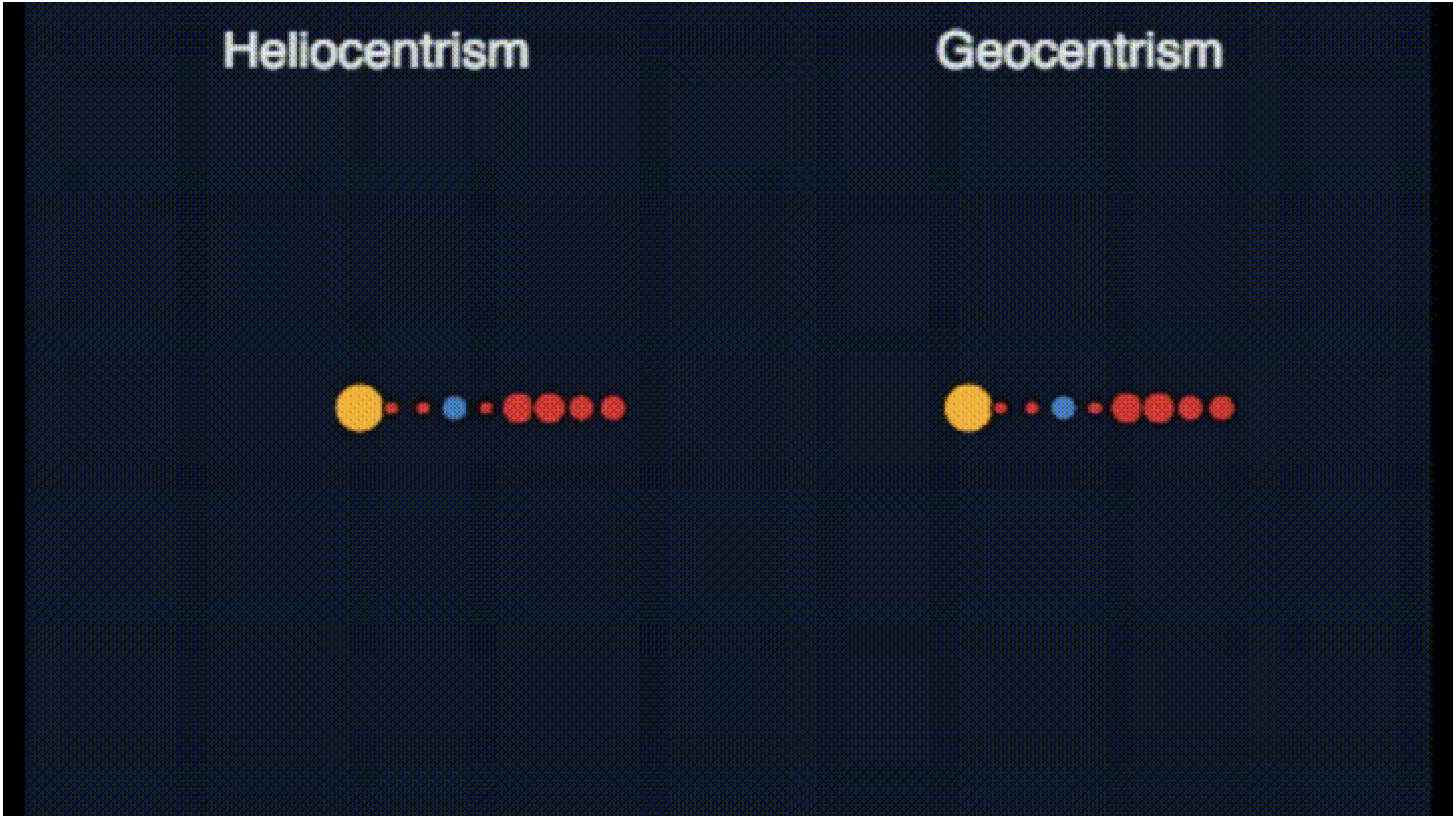
> Measurement error **plays a role**

- test-retest **dramatically up**
- convergent validity **slightly up**
- external validity **slightly up**

> But it **does not** close the risk validity gap

- **convergent** validity: tasks worse than questionnaires
- **external** validity still abysmally **low**

What if…



Have we got the right theory?

> We have so far **assumed EUT** (or **PT**) + noise. But it’s no more the only game in town!

- role of **risk perception** (Holzmeister et al. ManSci 2019)
  - laymen & traders perceive risk as probability of loss
  - perception seems *not* driven by variance but skewness

Have we got the right theory?

> We have so far **assumed EUT** (or **PT**) + noise. But it’s no more the only game in town!

- role of **risk modeling** (Crosetto & Flippin 202?)
  - risk, ambiguity, deep uncertainty
  - a nested structure, a nested experiment
    - what maps better to subjects’ behavior?

Have we got the right theory?

> We have so far **assumed EUT** (or **PT**) + noise. But it's no more the only game in town!

- role of **cognitive noise**
  - models of *noisy coding*: what we see as *risk aversion* could be just risk neutraility + the *way we see the world*
    - logarihtmic number perception (Khaw et al. 2021)
    - Bayesian Inference Model (Vieider 2024)
    - Role of complexity & Cognition (Oprea 2024)

TO BE
CONTINUED....

Thank you!