Limited Depth of Reasoning in Games

Duarte Gonçalves

University College London

Experimental Economics

Overview

1. Level-k and Cognitive Hierarchy Models

2. Identifying Higher-Order Rationality via Ring Games

3. Unstable Levels

4. Implications and Design Lessons

Overview

1. Level-k and Cognitive Hierarchy Models

2. Identifying Higher-Order Rationality via Ring Games

3. Unstable Levels

4. Implications and Design Lessons

Beauty Contest

Beauty Contest (Nagel, 1995 AER)

Participants choose $s_i \in [0, 100]$.

Winner is person whose chosen number is closest to p times the average; ties broken uniformly at random.

k-Rationalisable strategies $R^k := \{s_i \leq 100 p^k\}$.

Idea from Keynes's metaphor to describe stock market: what's important is to guess what everyone else is thinking.

$p \in (0, 1)$: dominance solvable; 0 is unique NE.

$p > 1$: all strategies rationalisable; 0 and 100 are unique NE.

Beauty Contest

Beauty Contest (Nagel, 1995 AER)

Participants choose $s_i \in [0, 100]$.

Winner is person whose chosen number is closest to p times the average; ties broken uniformly at random.

k-Rationalisable strategies $R^k := \{s_i \leq 100 p^k\}$.

Idea from Keynes's metaphor to describe stock market: what's important is to guess what everyone else is thinking.

$p \in (0, 1)$: dominance solvable; 0 is unique NE.

$p > 1$: all strategies rationalisable; 0 and 100 are unique NE.

Level-k:

Anchor/level 0 given, s^0; Level-k s^k=BR(s^{k-1}).

Beauty Contest

Beauty Contest (Nagel, 1995 AER)

Logistics

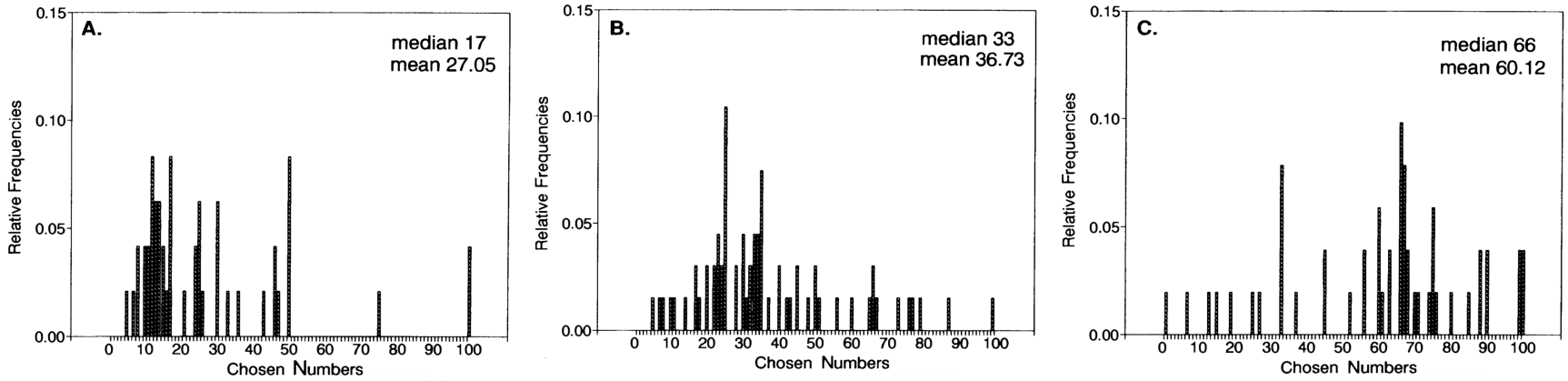
4 rounds with feedback; 15-18 participants per session, fixed matching, students; participate only once.

Treatments: p = 1/2 (sessions 1-3), 2/3 (4-7), 4/3 (8-10).

Instructions read aloud (used to be standard, now not so much).

Bonus per round 20 DM (≈ USD 13); Participation fee 5 DM (≈ USD 3). Duration: about 45min.

Beauty Contest



A: $p = 1/2$, B: $p = 2/3$, C: $p = 4/3$.

Initial Choices

Choices not 1-rationalisable: $p = 1/2$ 6%; $p = 2/3$ 10%.

100*p* Choices: $p = 1/2$ 8%; $p = 2/3$ 6%.

Suggestive of peaks around:

50, 25, and 12.5 for $p = 1/2$;

66, 33, and 22 for $p = 2/3$;

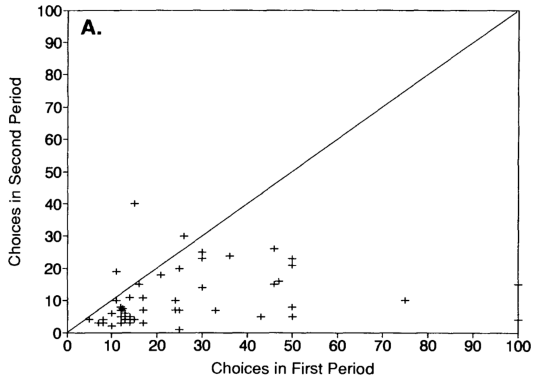
33, 66, 75, 100 for $p = 4/3$.

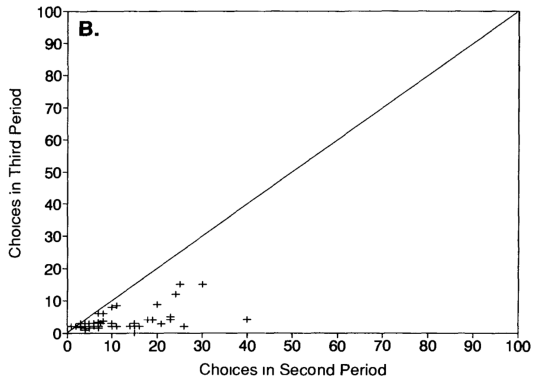
Anchor of 50 not very credible; choices too dispersed.

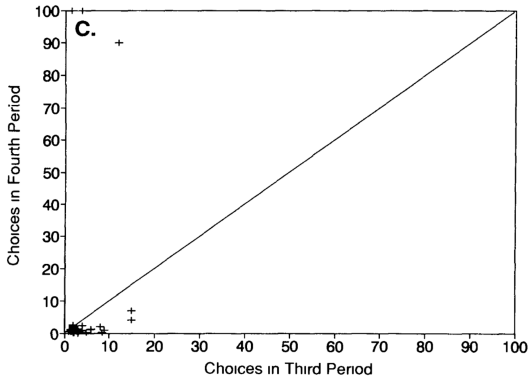
Nonetheless, stepwise reasoning still apparent.

Beauty Contest

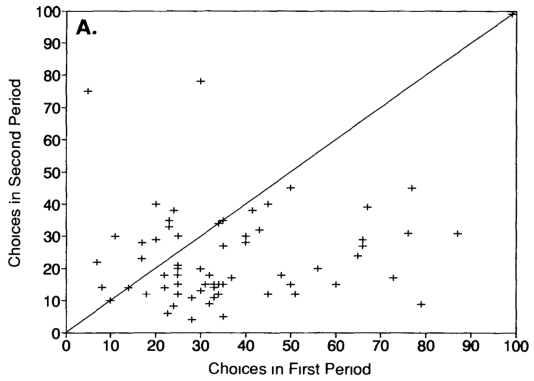
p = 1/2

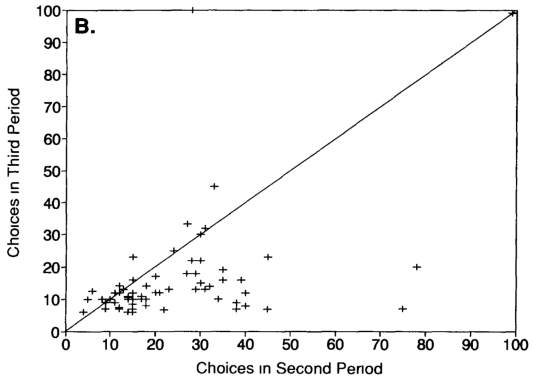


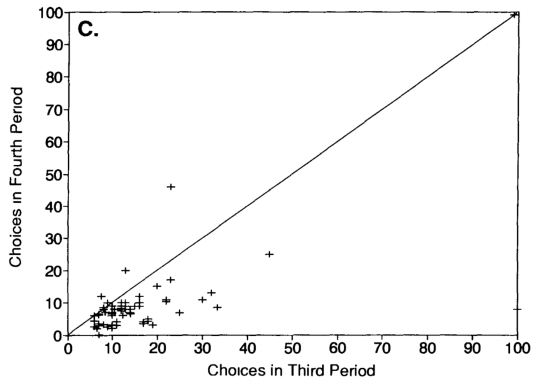




p = 2/3







Learning with Feedback: Significant learning over just 4 rounds.

For p = 1/2, 2/3, choices converge toward zero; faster for p = 1/2.

For p = 4/3, choices converge toward one.

Ho, Camerer, & Weigelt (1998 AER) replicate the results (10 rounds);

Finite # IESDS rounds with strategy space lower bound > 0, e.g., [100, 200].

Beauty Contest

Bosch-Domènech, Montalvo, Nagel, & Satorra (2002 AER) report on a multitude of replications.

Beauty Contest

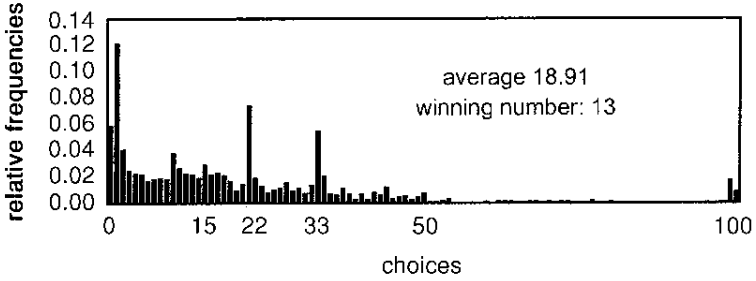
TABLE 3—DESIGN AND STRUCTURE OF 17 EXPERIMENTS, CLASSIFIED INTO SIX GROUPS

Experiment (Month/year)	Data from	Subject pool	Number of players per session (total)	Payoffs	Time to submit the number	Submission by type	Comments
1. Lab # 1–5 (8/1991, 3/1994)	Nagel (1995, 1998)	Undergraduates from various departments at Bonn and Caltech (#5)	15–18 (86)	20 DM to winners, 5 DM show-up fee, $20 and $5 show-up fee, split if tie	5 min.	Immediately	Optional
2. Classroom # 6, 7 (10/1997)	Collected by Teachers at UPF: Charness, Hurkens, Lopez, Nagel	2nd-year economic undergrads UPF, in Economic Theory class. Limited knowledge in game theory	30–50 (138)	3,000 Pesetas ($24), split if tie	5 min.	Immediately	Optional
3. Take-home # 8, 9 (10/1997)			30–50 (119)		1 week	Hand in personally	Optional
4. Theorists #10 (12/1997)	Collected by Rockenbach	3rd–4th-year undergraduates in Game Theory class, Bonn	54	30 DM ($18), split if tie	3 weeks	Hand in personally	Optional
# 11, 12 (6,10/1997)	Collected by Nagel	Game theorists/ Economists in Conference	20–40 (92)	$20 split if tie	5 min.	Immediately or e-mail	Optional
# 13 (11/1995) by e-mail		Profs/doctorates of Department of Business/Economics in UPF		*Handbook of Experimental Economics*. Random draw if tie	1 week		
5. Internet newsgroup # 14 (10/1997)	Collected by Participant in S. See Selten and Nagel (1998)	Newsgroup in Internet (responses via e-mail)	150	30 DM ($18) or book	1 week	e-mail	Optional
6. Newspaper # 15 (5/1997)	Thaler (1997) in *Financial Times*	Readers of FT	1476	2 tickets London–NY or London–Chicago	2 weeks	Letters	Required to become a winner
# 16 (5/1997)	Bosch, Nagel (1997) in *Expansión*	Readers of E	3696	100.000 Pesetas ($800)	1 week	Letter, e-mail, fax	Optional
# 17 (10/1997)	Selten, Nagel (1998) in *Spektrum der Wissenschaft*	Readers of S	2728	1,000 DM ($600), random draw if tie	2 weeks	Letter, e-mail	Optional

Beauty Contest

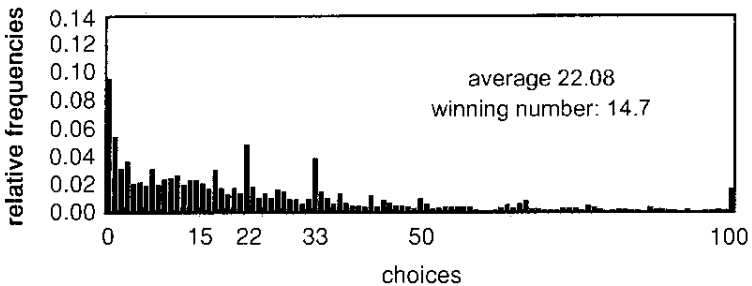
(a)

Financial Times **experiment (1,468 subjects)**



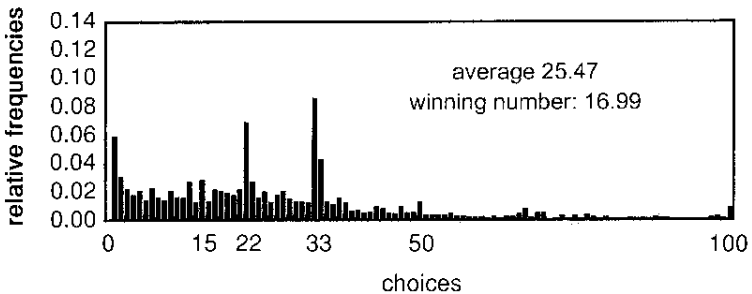
(b)

Spektrum **experiment (2,729 subjects)**



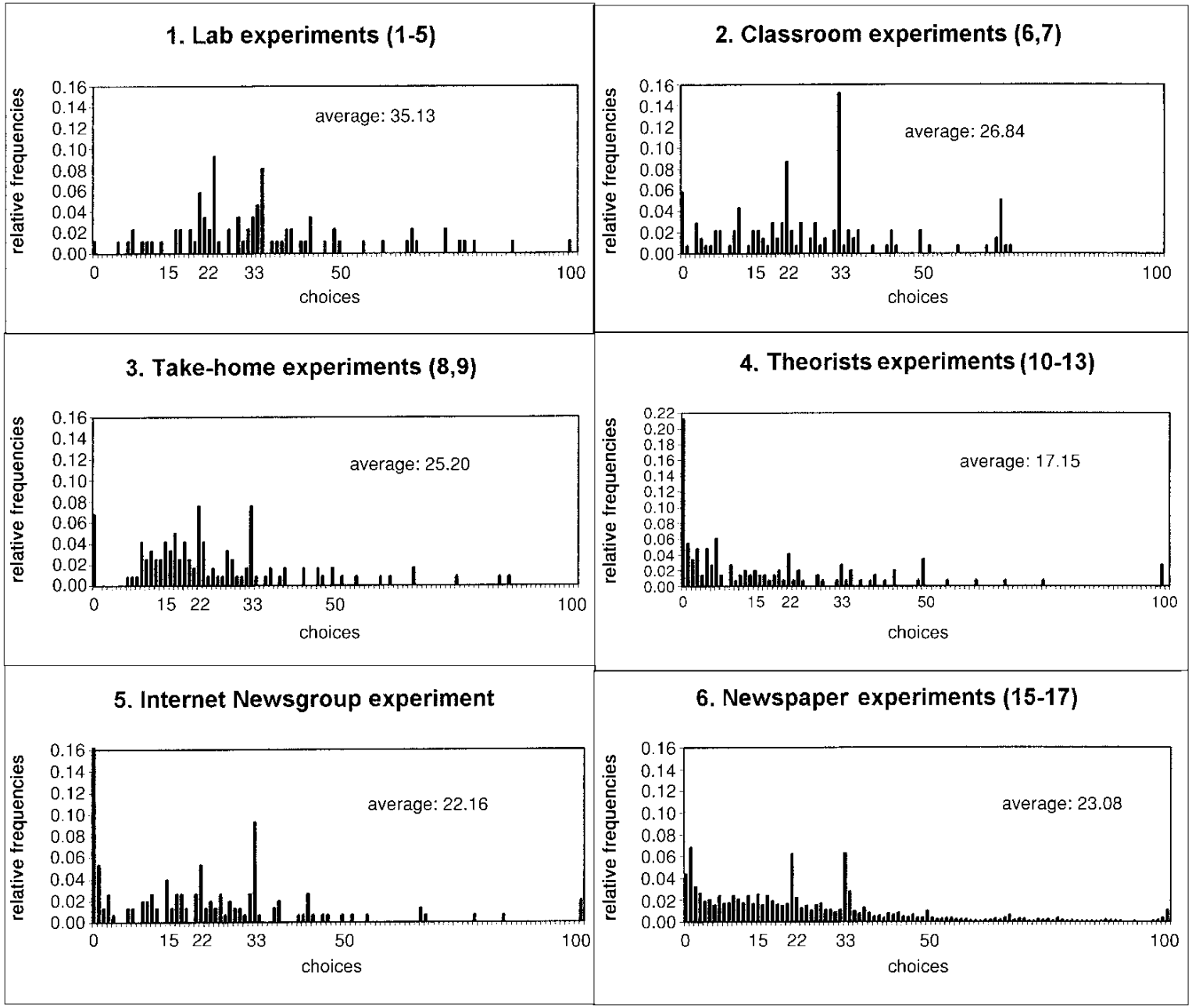
(c)

Expansión **experiment (3,696 subjects)**



Clear Stepwise Reasoning: Now clear peaks around 33 = BR(50), 22 = BR(33), and the dominance solution 0.

Beauty Contest



Beauty Contest

TABLE 2—RELATIVE FREQUENCIES OF THE DIFFERENT TYPES OF REASONING FROM THE COMMENTS OF E AND S EXPERIMENTS

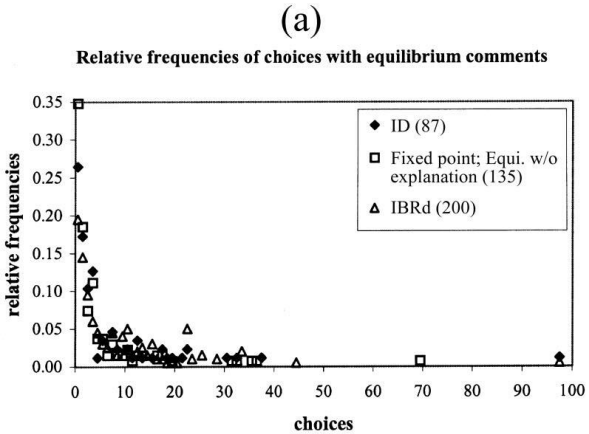
Types of reasoning processes	Relative frequencies
Fixed point	2.56 percent
Equilibrium, without further explanation	14.61 percent
Iterated dominance (ID)	13.77 percent of which 11.10 percentage points are Level-∞
Iterated best-reply degenerate (IBRd)	54.71 percent of which 25.45 percentage points are Level-∞ 12.47 percentage points are Level 0
Iterated best-reply nondegenerate (IBRnd)	9.28 percent
Experimenters	5.09 percent

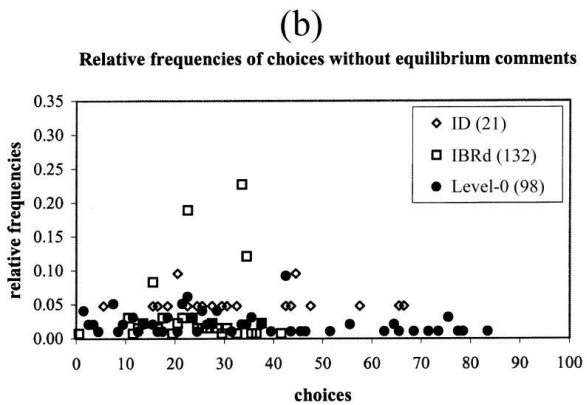
(Self-reported) thought process:

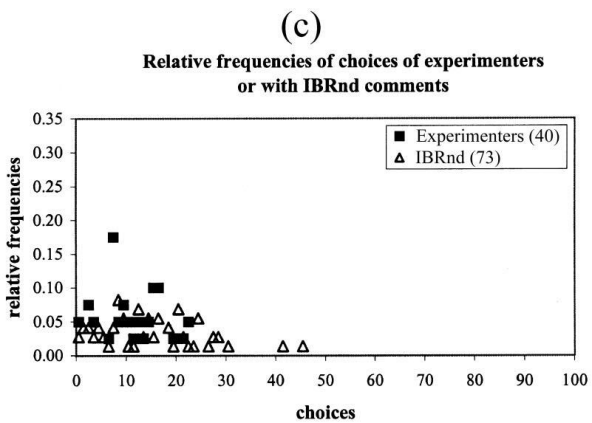
Interesting! Use debrief comments!

(Note 'experimenters': "I decided to run an experiment with a group of friends. Since I believed that the sample was representative of the participants in the general experiment, I assumed the result of the experiment would be a good indicator of the solution.")

Beauty Contest







(Self-reported) thought process:

Revealing of choices.

Extending Level-k

Cognitive Hierarchy: Camerer, Ho, & Chong (2004 QJE) provide portable level-k extension.

Model: level 0 is anchor (uniform);

level k BR to truncated distribution of levels 0,..., $k-1$.

Distribution of levels assumed to be Poisson $f(k) := e^{-k}\tau^k/k!$.

Mostly, data-fitting device.

Extending Level-k

Cognitive Hierarchy: Camerer, Ho, & Chong (2004 QJE) provide portable level-k extension.

Model: level 0 is anchor (uniform);
level k BR to truncated distribution of levels 0,..., $k-1$.

Distribution of levels assumed to be Poisson $f(k) := e^{-k}\tau^k/k!$.

Mostly, data-fitting device.

How Stable are Predictions?

"Across games, the Poisson-CH model fits only a little less accurately than when estimates are common within games"

However, CHC also say: "A sensible intuition is that when stakes are higher, subjects will use more steps of reasoning (and may think others will think harder too). (...) When stakes are higher, τ is estimated to be twice as large (5.01 versus 2.51), which is a clue that some sort of cost-benefit analysis may underlie steps of reasoning. (...) In future work, variation in estimates of τ could be useful in sharpening a theory of how steps of thinking are chosen endogenously."

Cognitive Hierarchy

Pioneering references for level-k-type models: Stahl (1993 GEB) and Stahl & Wilson (1994 JEBO, 1995 GEB)

Papers also allowed for (naive-)NE, logit BRs, and cognitive hierarchy.

Cognitive Hierarchy

Pioneering references for level-k-type models: Stahl (1993 GEB) and Stahl & Wilson (1994 JEBO, 1995 GEB)

Papers also allowed for (naive-)NE, logit BRs, and cognitive hierarchy.

Logistics

10 and 12 symmetric 3x3 games, dominance solvable, weak dominance solvable, unique PSNE, unique MSNE, multiple NE.

Small samples (1990s!) involving US undergrads.

No feedback, one round per game, random matching ex-post.

Average game payoffs across rounds determine prob. bonus.

Cognitive Hierarchy

Pioneering references for level-k-type models: Stahl (1993 GEB) and Stahl & Wilson (1994 JEBO, 1995 GEB)

Papers also allowed for (naive-)NE, logit BRs, and cognitive hierarchy.

Logistics

10 and 12 symmetric 3x3 games, dominance solvable, weak dominance solvable, unique PSNE, unique MSNE, multiple NE.

Small samples (1990s!) involving US undergrads.

No feedback, one round per game, random matching ex-post.

Average game payoffs across rounds determine prob. bonus.

Fit structural model. Find some type 0 behaviour (choice of dominated actions), and nontrivial frequencies of types 1 and 2 as well as of NE.

Analysis lacks clear hypotheses, model lacks a point other than fitting data, also lacks portability.

Overview

1. Level-k and Cognitive Hierarchy Models

2. Identifying Higher-Order Rationality via Ring Games

3. Unstable Levels

4. Implications and Design Lessons

Identifying Levels

Ring games (Kneeland, 2015 Ecta)

n-player ring game is game where payoffs of player $i \in \{0, 1, ..., n-1\}$ only depend on player i and $(i+1)$'s actions (modulo n) and exactly one player has a dominant strategy.

Identifying Levels

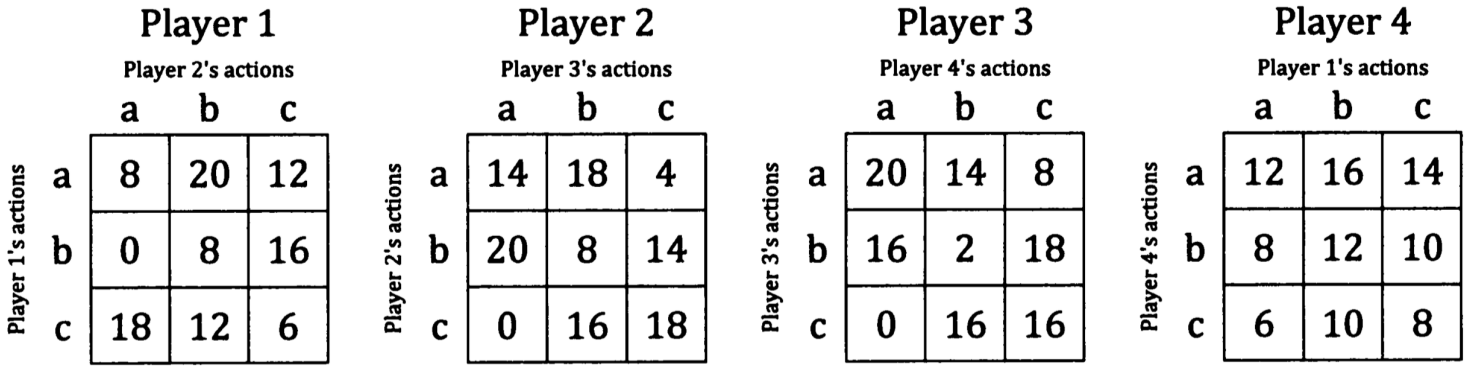


FIGURE 5.—G1.

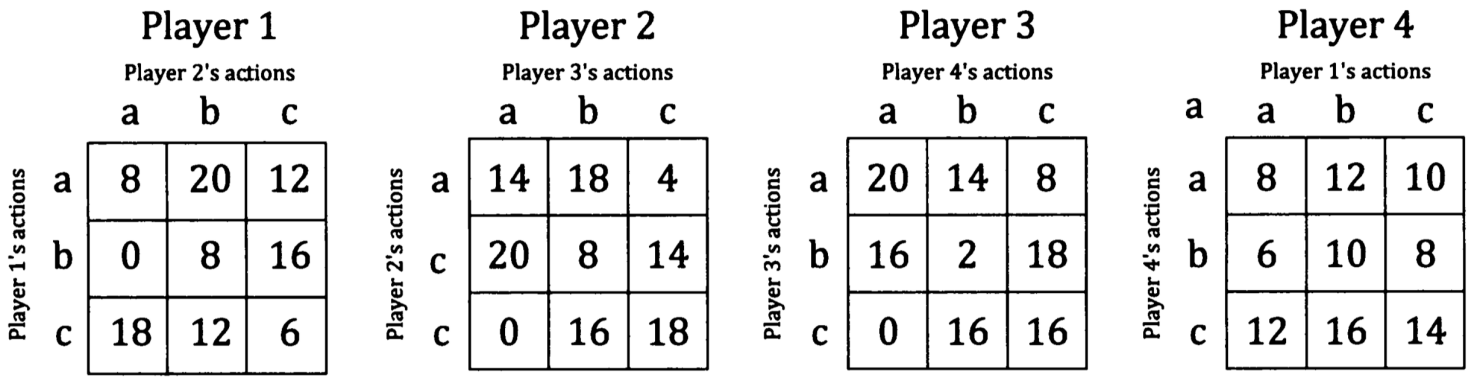


FIGURE 6.—G2.

Identification:

If L1/2/3, then Player 1's choices are the same in G1 and G2, but not if L4.

If L1/2, then Player 2's choices are the same in G1 and G2, but not if L3/4; etc.

Use action variation across games to identify level.

Identifying Levels

TABLE I

PREDICTED ACTIONS UNDER RATIONALITY AND ASSUMPTIONS ER IN THE EIGHT GAMES

	Games							
	P1		P2		P3		P4	
Type	G1	G2	G1	G2	G1	G2	G1	G2
R1	$(a,a)(b,b)(c,c)$		$(a,a)(b,b)(c,c)$		$(a,a)(b,b)(c,c)$		(a,c)	
R2	$(a,a)(b,b)(c,c)$		$(a,a)(b,b)(c,c)$		(a,b)		(a,c)	
R3	$(a,a)(b,b)(c,c)$		(b,a)		(a,b)		(a,c)	
R4	(a,c)		(b,a)		(a,b)		(a,c)	

Identifying Levels

Games:

8 4-player ring games.

Allow for one error in at most one of 8 games.

Only 5% chance that uniformly random player would get assined to R1-R4 (>6k combinations).

One round randomly selected for payment.

Avg payment CAD 17 incl. CAD 5 show-up fee; 45min.

Min 90 seconds on each game; no feedback.

Anonymous rematching within session.

116 students at UBC. Participant's matrix leftmost (80); also, robustness treatment with random order (36).

Identifying Levels

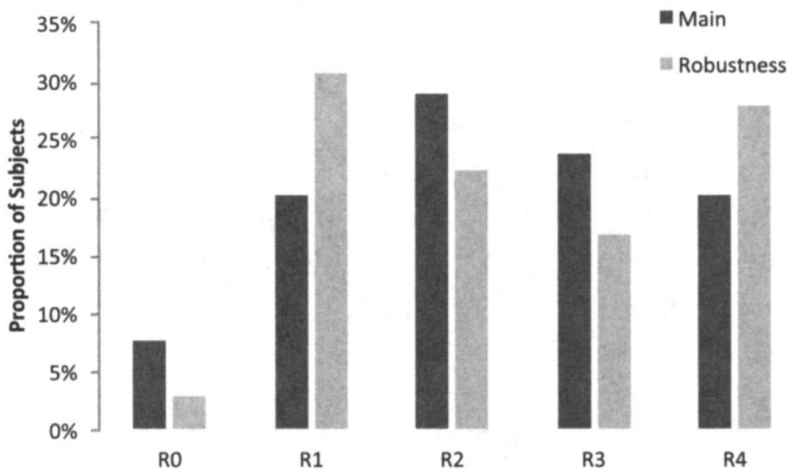


FIGURE 7.—Subjects classified by order of rationality, by treatment.

Only 6% are not identified (R0); over 68% are *exactly* identified.

23% participants failed quiz; of R0, this goes to 71%; of R4, only 8%.

Takeaways

Higher-order rationality is scarce even when design isolates beliefs, reinforcing need for models with shallow reasoning depth.

Supports use of level-k/CH but emphasises heterogeneity; not all subjects reach L3.

Provides methodology for testing epistemic concepts beyond best response (common knowledge hierarchy).

Overview

1. Level-k and Cognitive Hierarchy Models

2. Identifying Higher-Order Rationality via Ring Games

3. Unstable Levels

4. Implications and Design Lessons

Stable Levels?

Stable Levels: Is Level-*k* meaningful individual characteristic capturing 'strategic sophistication'?

If not stable across games, then

(1) stepwise reasoning is not good descriptor of thought process, or

(2) level is not capturing stable characteristic.

Stable Levels?

Stable Levels: Is Level-*k* meaningful individual characteristic capturing 'strategic sophistication'?

If not stable across games, then

(1) stepwise reasoning is not good descriptor of thought process, or

(2) level is not capturing stable characteristic.

Georganas, Healy, & Weber (2015 JET) study individual-level sophistication stability within and across classes of games: undercutting games and guessing games.

Undercutting Games

	1	2	3	4	5	6	7
1	1 1	10 -10	0 0	0 0	0 0	0 0	-11 0
2	-10 10	0 0	10 -10	0 0	0 0	0 0	0 0
3	0 0	-10 10	0 0	10 -10	0 0	0 0	0 0
4	0 0	0 0	-10 10	0 0	10 -10	10 -10	10 -10
5	0 0	0 0	0 0	-10 10	0 0	0 0	0 0
6	0 0	0 0	0 0	-10 10	0 0	0 0	0 0
7	0 -11	0 0	0 0	-10 10	0 0	0 0	-11 -11

Fig. 1. Undercutting game 1 (UG1).

Undercutting games: symmetric 2-player game, $s_i \in \{1, \ldots, n\}$.

Payoffs: (in USD)

$u_i(s) = 1 = u_j(s)$ if $s_i = s_j = 1$;

$u_i(s) = 10 = -u_j(s)$ if $s_i = m < s_j$ or $s_i = s_j - 1 \leq m$;

$u_i(s) = -11$ if $s_i = n$ and $s_j \in \{1, n\}$;

otherwise $u_i(s) = 0$.

Undercutting Games

	1	2	3	4	5	6	7
1	1 1	10 -10	0 0	0 0	0 0	0 0	-11 0
2	-10 10	0 0	10 -10	0 0	0 0	0 0	0 0
3	0 0	-10 10	0 0	10 -10	0 0	0 0	0 0
4	0 0	0 0	-10 10	0 0	10 -10	10 -10	10 -10
5	0 0	0 0	0 0	-10 10	0 0	0 0	0 0
6	0 0	0 0	0 0	-10 10	0 0	0 0	0 0
7	0 -11	0 0	0 0	-10 10	0 0	0 0	-11 -11

Fig. 1. Undercutting game 1 (UG1).

Level-k:

$s_i > m$ never BR, 2-player game, hence strictly dominated (requires elimination by mixed strat.); iterate.

m= # IESDS steps with maximal deletion.

If level 0 unif randomises, level k plays $s_i^k = m + 1 - k$.

Games chosen: $(n, m) = (7, 4), (9, 4), (9, 6)$ and a variant of $(7, 4)$ increasing payoffs $u_i(m, m+1) = 30$.

Guessing Games

Guessing games: (Costa-Gomes and Crawford, 2006 AER) asymmetric 2-player game; $S_i = [a_i, b_i]$.

Payoffs: (in USD)

Decreasing in $e_i := |s_i - p_i s_j|$.

$u_i(s_i, s_j) := 15 - (11/200)e_i$ if $e_i \leq 200$;

$u_i(s_i, s_j) := 5 - (1/200)e_i$ if $e_i \in (200, 1000]$;

$u_i(s_i, s_j) := 0$ if otherwise.

Level-k:

If level 0 unif randomises over $[a_j, b_j]$ or plays midpoint, level 1 plays $s_i^1 = p_i(a_j + b_j)/2$ (or nearest point in S_i).

Iterating, converges to NE with player choosing on boundary of S_i and other best-responding.

Three asymmetric games.

Procedures

116 OSU students.

Quizzes: Mensa IQ items, eye-gaze test (theory of mind), digit span, cognitive reflection, takeover game.

Games in fixed order: undercutting games, guessing games in one role, then in other role.

Played all 10 games/roles, 3 choices in each:

- Against a random opponent.
- Against opponent scoring highest in quizzes.
- Against opponent scoring lowest in quizzes.

No feedback, could go back and revise choices up to 4 times per game (most participants don't).

Randomly selected 2 undercutting games and 2 guessing games for payment.

Participants "earning less than $6 (the standard show-up fee) were paid $6 for their time." (rather unconventional; induces weird incentives.)

Aggregate Level Distribution

Table 2
Frequency of levels in each game, and when pooling each family of games.

Game	L0	L1	L2	L3	Nash
UG1	7.76%	32.76%	19.83%	10.34%	29.31%
UG2	7.76%	32.76%	22.41%	7.76%	29.31%
UG3	5.17%	27.59%	18.10%	5.17%	43.97%
UG4	6.03%	31.03%	29.31%	5.17%	28.45%
UGs pooled	4.31%	28.45%	26.72%	5.17%	35.34%
GG5	6.03%	70.69%	9.48%	12.07%	1.72%
GG6	0.86%	65.52%	17.24%	11.21%	5.17%
GG7	43.10%	37.07%	13.79%	1.72%	4.31%
GG8	6.90%	39.66%	24.14%	21.55%	7.76%
GG9	5.17%	42.24%	23.28%	4.31%	25.00%
GG10	9.48%	38.79%	24.14%	19.83%	7.76%
GGs pooled	1.72%	50.00%	10.34%	10.34%	27.59%

MLE estimates considering only levels 0–3, and NE and allowing for logit-type errors.

Aggregate distribution stable in undercutting games, not so much in guessing games.

Results: Type Persistence

Table 4
Markov transition between single-game levels within the four undercutting games.

From ↓ to →	L0	L1	L2	L3	Nash
L0	**43.0%**	22.6%	7.5%	9.7%	17.2%
L1	4.9%	**59.7%**	14.6%	4.4%	16.4%
L2	2.2%	20.2%	**57.1%**	9.0%	11.5%
L3	9.1%	19.2%	**28.3%**	18.2%	25.3%
Nash	3.5%	15.6%	7.9%	5.5%	**67.5%**
Overall	6.7%	31.0%	22.4%	7.1%	32.8%

Table 5
Markov transition between single-game levels within the six guessing games.

From ↓ to →	L0	L1	L2	L3	Nash
L0	8.7%	**48.2%**	18.1%	12.3%	12.8%
L1	11.7%	**53.1%**	16.8%	11.2%	7.1%
L2	11.5%	**44.2%**	27.4%	10.0%	6.9%
L3	12.4%	**46.6%**	15.9%	13.2%	12.0%
Nash	17.7%	**40.3%**	15.0%	16.3%	10.7%
Overall	11.9%	**49.0%**	18.7%	11.8%	8.6%

Stability $\Longrightarrow$ diagonal close to 1.

Comment: It would've been nice to have a measure of variability of assigned level and steady state distributions...

		L0	L1	L2	L3	NE
Steady state distributions:	UG	6.68%	31.05%	22.45%	7.13%	32.69%
	GG	11.91%	49.00%	18.68%	11.79%	8.63%

Results: Rank Persistence

Table 6
Observed frequency with which two players' levels strictly switch their ordering, compared to the expected frequency under independent, randomly-drawn levels.

	Data	Null hyp.
Pooled UGs vs. pooled GGs		
Switch frequency:	25.0%	24.9%
Non-switch frequency:	22.7%	24.9%
Switch ratio:	1.10	1.00
Undercutting games		
Switch frequency:	13.2%	27.1%
Non-switch frequency:	45.3%	27.1%
Switch ratio:	0.29	1.00
Guessing games		
Switch frequency:	19.9%	23.8%
Non-switch frequency:	22.2%	23.8%
Switch ratio:	0.89	1.00

Switch across pair of games: $k_i > k_j$ in one game and $k_i < k_j$ in another.

Non-switch across pair of games: $k_i > k_j$ in both games.

UGs: rank has high persistence. GGs: rank as low persistence.

Across game types: rank more likely to switch than not!

Robustness check: reversing order of games (GGs 1st, UGs 2nd) does not change conclusions.

Other Results

Correlation with Cognitive Quizzes:

Quizzes fail to substantively predict types. Only CRT significantly predicts earnings.

Comment: analysis opaque (box plots, MnL regressions, etc).

Would've preferred a simple Wilcoxon rank-sum test for each cognitive quiz vs pooled individual level.

Other Results

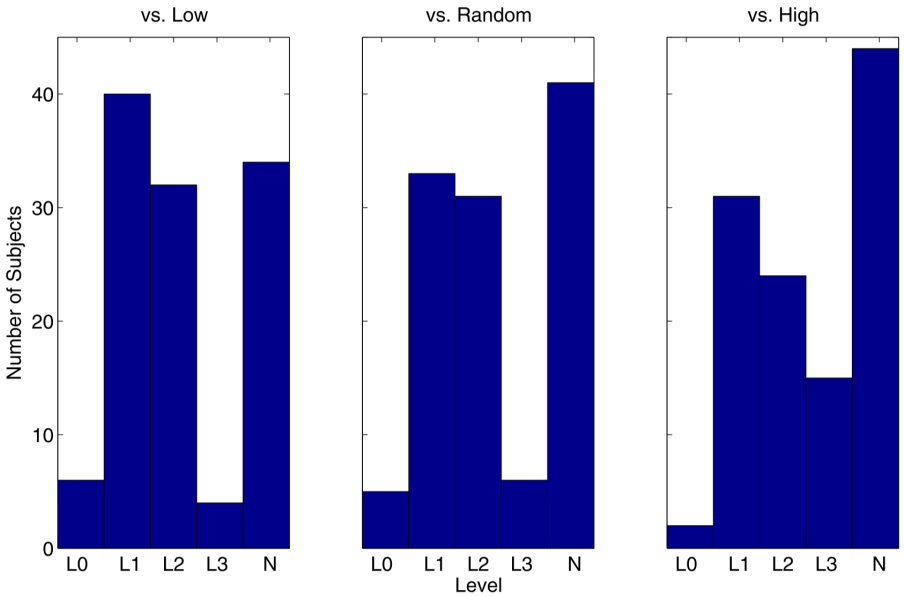


Fig. 8. Level distributions by opponent in the pooled undercutting games.

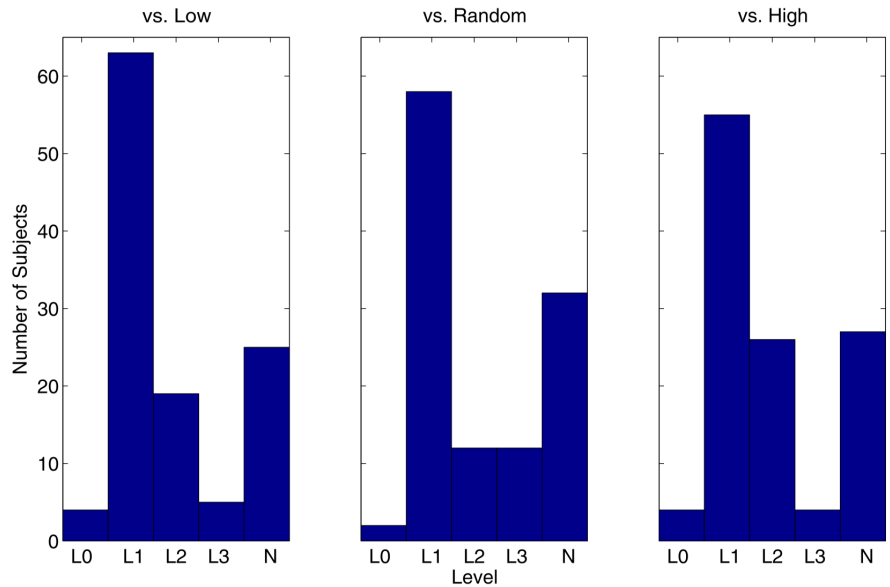


Fig. 10. Level distributions by opponent in the pooled guessing games.

Adjustment Depending on Opponent:

Some FOSD shift low quiz score $\rightarrow$ high quiz score, but minor.

Follow-Up on Inconsistent Depth

Cooper et al. (2024 AEJMicro): Five classes of games (imperfect price competition, minimum effort, traveller's dilemma, 11-20, all-pay auctions), four games per class.

- Min coordination (Goeree & Holt 2005 GEB): $u_i(s_i, s_j) := \min\{s_i, s_j\} - \alpha_i s_i$, $\alpha_i \in (0, 1)$; PSNE $s_i = s_j$.
- Traveller's Dilema (Capra et al. 1999 AER): $u_i(s_i, s_j) = s_i + 1\{s_i < s_j\}\,\alpha_i - 1\{s_i > s_j\}\,\alpha_i$, $\alpha_i > \max S_i - \min S_i$; ! NE $s_i = \min S_i > 0$.
- 11-20 (Arad & Rubinstein 2012 AER): $u_i(s_i, s_j) = s_i + 1\{s_i = s_j - 10\}\,\alpha$, $\alpha > 10$; no PSNE.

Level-k react to shifts in own (L1) or in opponent payoffs (L2). Paper does not consider L3 and above (strange for 11-20 game).

Follow-Up on Inconsistent Depth

Logistics

224 students, pencil-and-paper sessions; no feedback across 20 games.

For each class, subjects simultaneously chose actions for all four payoff permutations before observing any outcomes; no feedback across 20 games.

Monetary stakes: avg EUR 18–20 including EUR 5 show-up fee.

Findings

Approx. 20% of reaction patterns consistent with fixed level within class of games; virtually none across classes.

Structural estimates with mixture types: 89% of population classified as mixing levels either every game or every class.

Lessons on Heterogeneity

Distribution of depths stable.

Individual depth of reasoning is context-dependent: framing, payoff stakes, and environment trigger different anchors.

Mixture models (pure-mixing, semimixing) outperform fixed-type models even after accounting for noisy best responses.

Overview

1. Level-k and Cognitive Hierarchy Models

2. Identifying Higher-Order Rationality via Ring Games

3. Unstable Levels

4. Implications and Design Lessons

Key Takeaways

1. Dominance-solvable predictions systematically fail; choices denoting low depth of reasoning are pervasive.
2. Level-k and CH models capture aggregate behaviour, but individual depths are context-dependent (game-specific, payoff specific, opponent specific).

Limited depth of reasoning models used to explore deviations from rationalisability in more applied contexts, e.g.,:

IO: Hortaçsu et al. (2019 AER).

Macro: Farhi & Werning (2019 AER).

Mechanism design: Kneeland (2022 JET).

Revisiting Questions

What determines the depth of reasoning?

- It seems to depend on both own and others' limitations (Georganas et al. 2015 JET; also, Alaoui, Janezic, & Penta 2021 JET).
  When facing more sophisticated players, action distribution typically consistent with higher levels.
- It seems to increase with higher payoffs (cf. Camerer et al. 2004 QJE).
  Can we say more? (more later.)
- It varies across and within games. Why? How?
- What kinds of patterns should we expect to be robust? More rounds IESDS $\implies$ farther from rationalisable? Higher stakes $\implies$ higher k-rationalisable?
  Unclear if much can be said in this respect.

Revisiting Questions

What determines the depth of reasoning?

Is reasoning in strategic settings stepwise?

- Basic premise of level-*k*-style models, but most of the evidence comes from relying on games prone to stepwise reasoning or fitting somewhat ad-hoc structural models.
  Do we see evidence of it in other environments?
- Instability of individuals' levels within and across game classes poses significant challenge to the model.
  (My own) intuition suggests something noisier and less well-defined happens in general. (We'll say more about this later.)

Revisiting Questions

What determines the depth of reasoning?

Is reasoning in strategic settings stepwise?

- Basic premise of level-*k*-style models, but most of the evidence comes from relying on games prone to stepwise reasoning or fitting somewhat ad-hoc structural models.
  Do we see evidence of it in other environments?
- Instability of individuals' levels within and across game classes poses significant challenge to the model.
  (My own) intuition suggests something noisier and less well-defined happens in general. (We'll say more about this later.)

Anything else comes to mind?

Design Principles for Probing Rationalisability

Vary specific primitives to test hypotheses while holding primitives constant (e.g., change n, m in undercutting games).

Use **paired games** to isolate higher-order beliefs (ring games).

Introduce **opponent signals** (e.g., quiz-based matching) to see if behaviour adjusts endogenously.

Pre-test instructions with comprehension quizzes; ensure subjects common knowledge of understanding.

Collect **process data** (search, timing, revisions) to classify reasoning modes beyond outcomes.