Epistemic Experiments: Utilities, Beliefs, and Irrational Play

Paul J. Healy (OSU)

An Embarrassing Timeline

A pet project:

2010: Tenure. First attempt at 2×2 games.

2013: Redo experiment on pencil & paper

2014: Present at Pitt

2015: Add centipede games

2016: Add no-elicitation benchmark

2017: Add sequential-move 2×2 games

2019: "I'm never presenting this again."

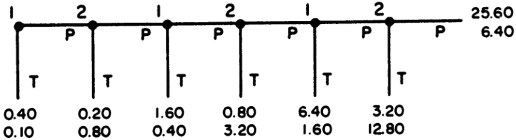
2020: COVID writing retreat, 1st draft

2023: Present at Pitt

The Standard Approach

Standard game theory experiment:

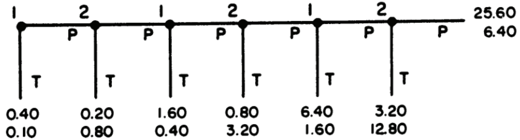
1. Interesting game form



The Standard Approach

Standard game theory experiment:

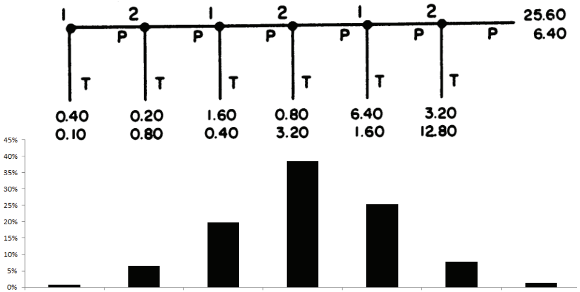
1. Interesting game form
2. Baseline theory/assumptions:
    - Selfish prefs, "Rational" behavior (eg, backwards induction)



The Standard Approach

Standard game theory experiment:

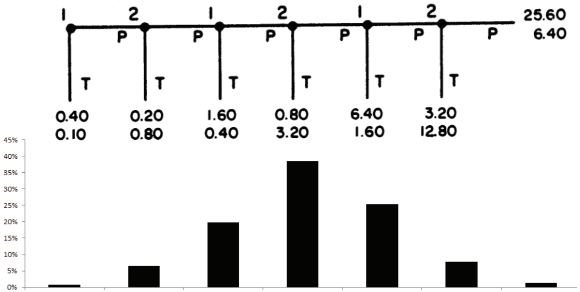
1. Interesting game form
2. Baseline theory/assumptions:
    - Selfish prefs, “Rational” behavior (eg, backwards induction)
3. Observe deviations



The Standard Approach

Standard game theory experiment:

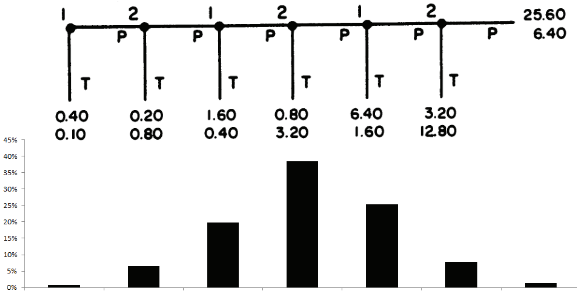
1. Interesting game form
2. Baseline theory/assumptions:
    - Selfish prefs, “Rational” behavior (eg, backwards induction)
3. Observe deviations
4. Posit alternative theory



The Standard Approach

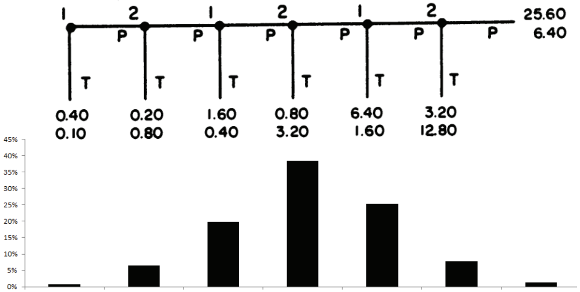
Standard game theory experiment:

1. Interesting game form
2. Baseline theory/assumptions:
    - Selfish prefs, “Rational” behavior (eg, backwards induction)
3. Observe deviations
4. Posit alternative theory
5. New treatments to test comparative statics

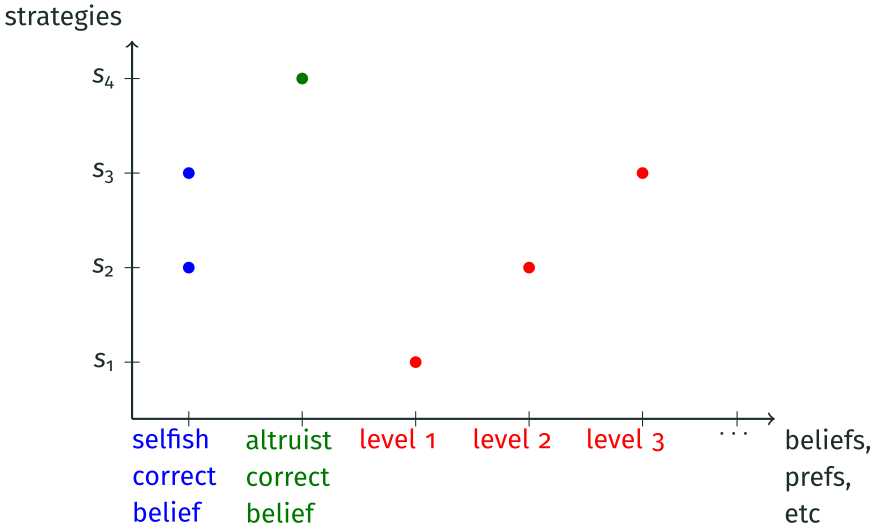


Alternative "Solution Concepts"

1. Nash with Altruism, Inequality Aversion
2. Reputation-building/Gang of Four
3. Level-*k* (wrong beliefs)
4. QRE (noisy equilibrium play)



A Risky Picture



Direct Measurement

Each solution concept makes specific assumptions about utilities, beliefs, rationality, etc.

Why not measure these things directly???

Direct Measurement

Each solution concept makes specific assumptions about utilities, beliefs, rationality, etc.

Why not measure these things directly???

(Yes, eliciting these things might change behavior. I'll get to that.)

Direct Measurement

Each solution concept makes specific assumptions about utilities, beliefs, rationality, etc.

Why not measure these things directly???

(Yes, eliciting these things might change behavior. I'll get to that.)

OK... but then what exactly should we measure?

Direct Measurement

Each solution concept makes specific assumptions about utilities, beliefs, rationality, etc.

Why not measure these things directly???

(Yes, eliciting these things might change behavior. I'll get to that.)

OK... but then what exactly should we measure?

We need a framework that encompasses *all* such theories

...a level playing field in which no theory is the null hypothesis

The Epistemic Game Theory Framework

The Observable Experiment: (I, S, X, π)

1. Players: $i \in I = \{1, 2\}$
2. Strategies: $s_i \in S_i$ *Ex*: when to Take
3. Outcomes: $(x_1, x_2) \in X$ *Ex*: $(\$6.40, \$1.60)$
4. Outcome function: $\pi(s_1, s_2) \in X$

i's Private Information: $\omega_i = (u_i, s_i, \vec{p}_i)$

1. Utility: $u_i(x_1, x_2)$
    - Non-selfish, but consequentialist
2. Chosen Strategy: $s_i \in S_i$
    - Mixing is only in beliefs (Aumann)
3. Beliefs
    - First-order: $p_i^1(u_{-i}, s_{-i})$
    - Second-order: $p_i^2(p_{-i}^1, u_{-i}, s_{-i})$
    - Hierarchy: $\vec{p}_i = (p_i^1, p_i^2, p_i^3, \ldots)$

Example: Nash Equilibrium

Players are in a (selfish) Nash equilibrium at $\omega = (\omega_1, \ldots, \omega_n)$ if:

1. Utility: $u_i(x_i, x_{-i}) = x_i$ (“selfish”)
2. Beliefs: correct beliefs about u_{-i}, s_{-i}.
3. Strategy: $s_i \in \arg\max_{s_i'} \left[\sum_{(u_{-i},s_{-i})} p_i^1(u_{-i}, s_{-i}) u_i(\pi(s_i', s_{-i}))\right]$

Example: Nash Equilibrium

Players are in a (selfish) Nash equilibrium at $\omega = (\omega_1, \ldots, \omega_n)$ if:

1. Utility: $u_i(x_i, x_{-i}) = x_i$ ("selfish")
2. Beliefs: correct beliefs about u_{-i}, s_{-i}.
3. Strategy: $s_i \in \arg\max_{s_i'} \left[\sum_{(u_{-i}, s_{-i})} p_i^1(u_{-i}, s_{-i}) u_i(\pi(s_i', s_{-i}))\right]$
    - Player i is **rational** at $\omega_i = (u_i, s_i, \vec{p}_i)$ if this is true
    - Let R_i be those (p_i^1, u_i, s_i) where i is rational
    - i's belief that $-i$ is rational is $p_i^2(R_{-i})$
    - Can define common knowledge of rationality, etc.

Example: Nash Equilibrium

Players are in a (selfish) Nash equilibrium at $\omega = (\omega_1, \ldots, \omega_n)$ if:

1. Utility: $u_i(x_i, x_{-i}) = x_i$ ("selfish")
2. Beliefs: correct beliefs about u_{-i}, s_{-i}.
3. Strategy: $s_i \in \arg\max_{s_i'} \left[\sum_{(u_{-i}, s_{-i})} p_i^1(u_{-i}, s_{-i}) u_i(\pi(s_i', s_{-i}))\right]$
    - Player i is **rational** at $\omega_i = (u_i, s_i, \vec{p}_i)$ if this is true
    - Let R_i be those (p_i^1, u_i, s_i) where i is rational
    - i's belief that $-i$ is rational is $p_i^2(R_{-i})$
    - Can define common knowledge of rationality, etc.
    - Aumann (1995): Nash equil. does **not** require c.k. of rationality
    - Rationality & c.k. of rationality $\Rightarrow$ IESDS

Example: Altruistic Nash Equilibrium

Players are in an Altruistic Nash equilibrium at $\omega = (\omega_1, \ldots, \omega_n)$ if:

1. Utility: $u_i(x_i, x_{-i}) = x_i + \alpha x_{-i}$
2. Beliefs: correct beliefs about u_{-i}, s_{-i}.
3. Strategy: $s_i \in \arg\max_{s_i'} \left[\sum_{(u_{-i}, s_{-i})} p_i^1(u_{-i}, s_{-i}) u_i(\pi(s_i', s_{-i}))\right]$
    - Player i is **rational** at $\omega_i = (u_i, s_i, \vec{p}_i)$ if this is true
    - Let R_i be those (p_i^1, u_i, s_i) where i is rational
    - i's belief that $-i$ is rational is $p_i^2(R_{-i})$
    - Can define common knowledge of rationality, etc.
    - Aumann (1995): Nash equil. does **not** require c.k. of rationality
    - Rationality & c.k. of rationality $\Rightarrow$ IESDS

Example: Level-k

Level-1:

1. Utility: selfish
2. Beliefs: u_{-i} selfish, s_{-i} uniformly distributed
3. Strategy: s_i is rational, given utility & beliefs

Level-$k > 1$:

1. Utility: selfish
2. Beliefs: u_{-i} selfish, s_{-i} is Level-$k-1$ strategy
3. Strategy: s_i is rational, given utility & beliefs

So, What Should We Measure?

1. Utility: $u_i(\pi(s_1, s_2))$

A Game Form

	L	R
U	\$2,\$2	\$0,\$3
D	\$3,\$0	\$1,\$1

$\swarrow \qquad \downarrow \qquad \searrow$

A Game

	L	R
U	$2, 2$	$0, 3$
D	$3, 0$	$1, 1$

A Game

	L	R
U	$4, 4$	$0, 0$
D	$0, 0$	$2, 2$

A Game

	L	R
U	$0, 0$	$0, 0$
D	$0, 0$	$0, 0$

How To Measure Cardinal Utility

	L	R
U	\$2,\$2	\$0,\$3
D	\$3,\$0	\$1,\$1

A Game Form

- Elicit $u_i(x_1, x_2)$ for each cell
  - or, for each terminal node
- How?
  - Let $\overline{x} = (\$20, \$20)$, $\underline{x} = (\$0, \$0)$
  - "I'm indifferent between $(\$3, \$0)$ and getting $\overline{x}$ w/ prob. p^*"

$$
\begin{aligned}
u_i(\$3, \$0) &= p^* \underbrace{u_i(\overline{x})}_{=1} + (1 - p^*) \underbrace{u_i(\underline{x})}_{=0} \\
&= p^*
\end{aligned}
$$

Multiple Price List Elicitation

Row#	**Option A**	OR	**Option B**
1	(\$3,\$0)	or	(\$20,\$20) w/ prob 1%
2	(\$3,\$0)	or	(\$20,\$20) w/ prob 2%
$\vdots$	$\vdots$	$\vdots$	$\vdots$
q	(\$3,\$0)	or	(\$20,\$20) w/ prob q%
$q+1$	(\$3,\$0)	or	(\$20,\$20) w/ prob $q+1$%
$q+2$	(\$3,\$0)	or	(\$20,\$20) w/ prob $q+2$%
$q+3$	(\$3,\$0)	or	(\$20,\$20) w/ prob $q+3$%
$\vdots$	$\vdots$	$\vdots$	$\vdots$
99	(\$3,\$0)	or	(\$20,\$20) w/ prob 99%
100	(\$3,\$0)	or	(\$20,\$20) w/ prob 100%

Choose Option A or Option B (single switch point q)

One row randomly selected for payment

Multiple Price List Elicitation

Row#	Option A	OR	Option B
1	(\$3,\$0)	or	(\$20,\$20) w/ prob 1%
2	(\$3,\$0)	or	(\$20,\$20) w/ prob 2%
$\vdots$	$\vdots$	$\vdots$	$\vdots$
q	(\$3,\$0)	or	(\$20,\$20) w/ prob q%
$q+1$	(\$3,\$0)	or	(\$20,\$20) w/ prob $q+1$%
$q+2$	(\$3,\$0)	or	(\$20,\$20) w/ prob $q+2$%
$q+3$	(\$3,\$0)	or	(\$20,\$20) w/ prob $q+3$%
$\vdots$	$\vdots$	$\vdots$	$\vdots$
99	(\$3,\$0)	or	(\$20,\$20) w/ prob 99%
100	(\$3,\$0)	or	(\$20,\$20) w/ prob 100%

If you lie, you get the less-preferred option on some rows

I.C. as long as subject respects **statewise dominance** in rows

Issue 1: Consequentialism

- Elicit $u_i(\$15, \$5)$, e.g.
- u_i captures non-selfish preferences.
- u_i captures risk aversion.

Problem: Game theory: utility over *strategies*: $U_i(s_i, s_j)$

We elicit: utility over *outcomes*: $u_i(\pi(s_i, s_j))$

Solution: Assume consequentialism:

$$U_i(s_i, s_j) = u_i(\pi(s_i, s_j))$$

Is consequentialism reasonable?? Is it even testable??

Issue 1: Consequentialism

Example violating consequentialism:

	Nice	Mean
Foolish	\$5, \$5	\$5, \$5
Wise	\$100, \$5	\$5, \$5

$\pi(\text{Foolish}, \text{Nice}) = \pi(\text{Wise}, \text{Mean})$, but, intuitively $U_1(\text{Foolish}, \text{Nice}) \neq U_1(\text{Wise}, \text{Mean})$.

But how could you possibly observe that??

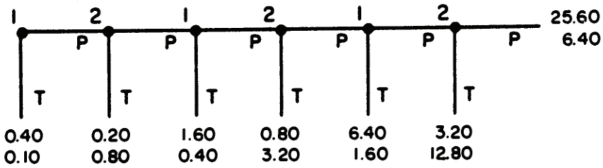
Claim: Cannot elicit $U_i(s_i, s_j)$. *Must* assume consequentialism.

Messy Solution: Elicit $u_i(\pi(s_i, s_j))$ in the *context* of the game.

So, What Should We Measure?

2. Strategies: s_i

- Easy. Just play the game.
- Complete contingent plan
  - "When will you Take?"
- Can re-elicit this at each node
  - Even when not active



So, What Should We Measure?

3. Beliefs: $p_i^1(u_{-i}, s_{-i})$, $p_i^2(p^1_{-i}, u_{-i}, s_{-i})$, ...

Measure before the game:

1. Best guess of $u_{-i}(x_1, x_2)$ at each terminal node

Measure at every node:

1. Probability of each s_{-i} (call that $p_i^1(s_{-i})$)
2. Best guess of $p^1_{-i}(s_i)$
3. Probability $-i$ is *rational*

My Wish List:

1. Entire distribution over u_{-i}
2. Correlation between u_{-i} and s_{-i}
3. Correlation between p^1_{-i} and s_{-i}

Issue 2: Contamination

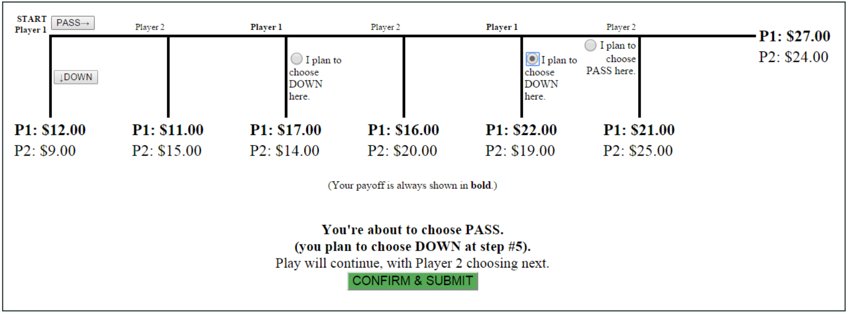
Does elicitation contaminate game play? PROBABLY!
Does game play contaminate elicitation?? PROBABLY!

- I embrace it! This is a *fully contaminated* experiment!
  - Necessary evil for the methodology
  - Intuitively: provides an upper bound on rationality

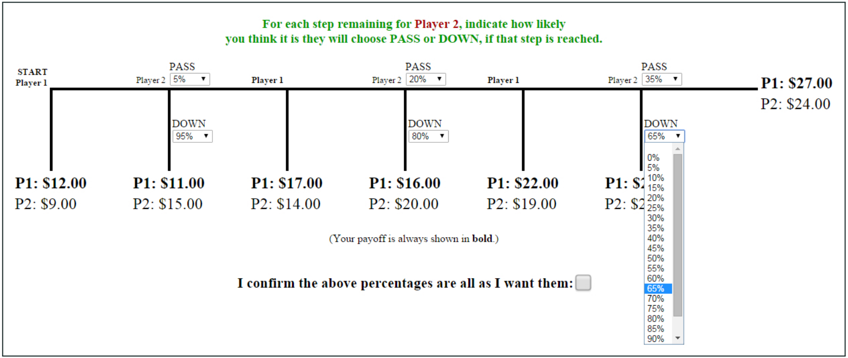
Empirically, I think it probably doesn't matter:

- In five 2×2 games, play w/out elicitation was the same in 4 of 5
- Behavior pretty similar to previous papers

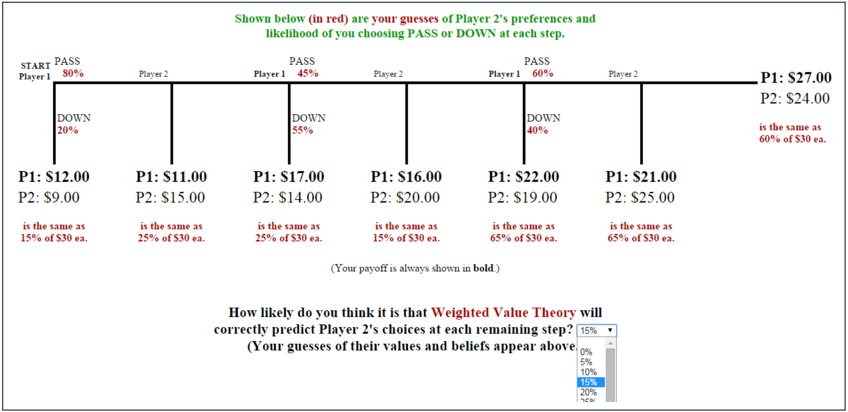
Screenshot: Eliciting Strategies



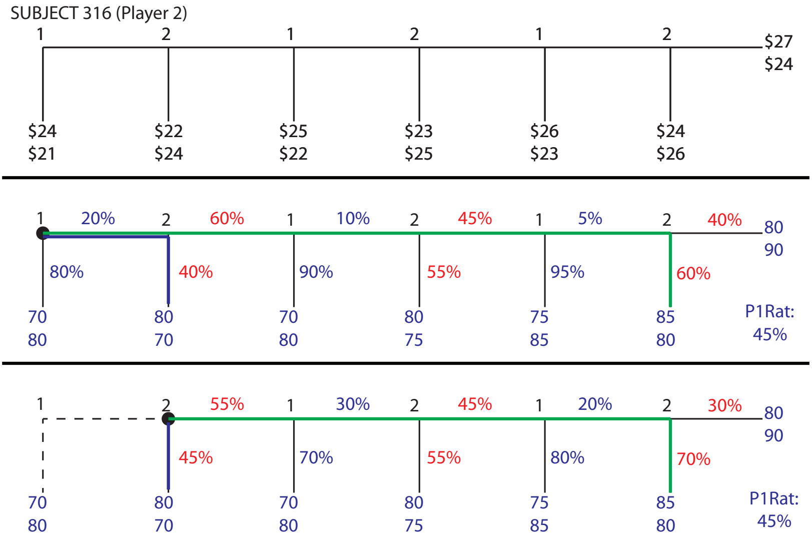
Screenshot: 1st-Order Beliefs



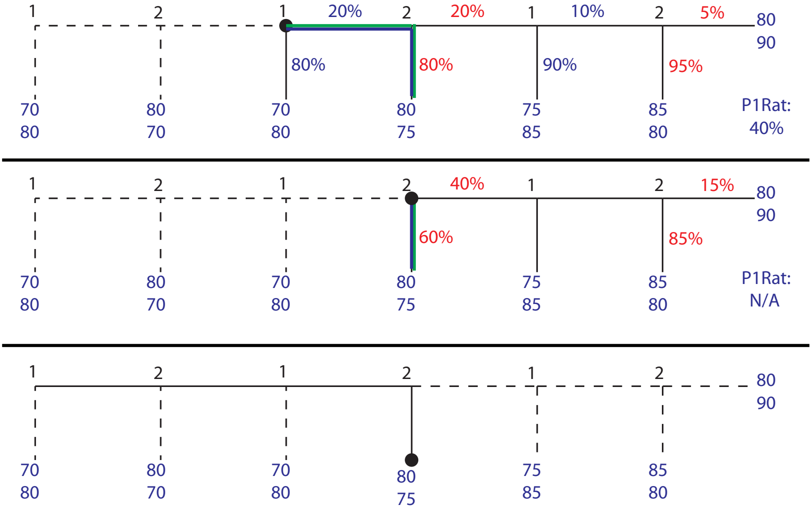
Screenshot: Belief of Rationality



Example Observation

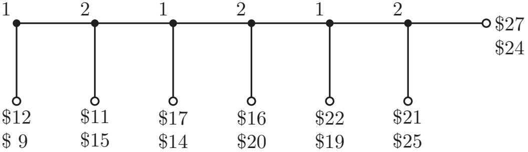


Example Observation

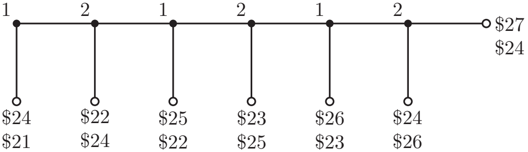


Centipede Treatments

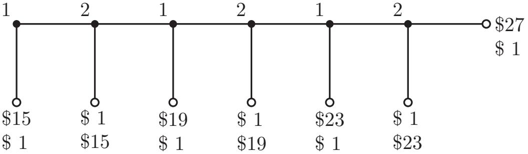
CENT-LO:
“Risk $1 to gain $5”



CENT-HI:
“Risk $2 to gain $1”



CENT-ALL:
“Risk ALL to gain $4”



Design Details

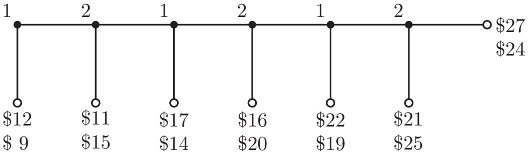
- OSU subject pool
- Custom software, ORSEE recruiting
- Between-subjects treatment (LO vs HI vs ALL)
- Play 4 periods. Elicitation only in last 2
  - Random rematching with feedback
- Only one game *or* elicitation is paid
- $19 average
- # subjects:

CENT-LO	CENT-HI	CENT-ALL
54	36	62

Results

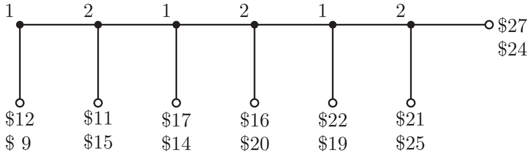
The Story

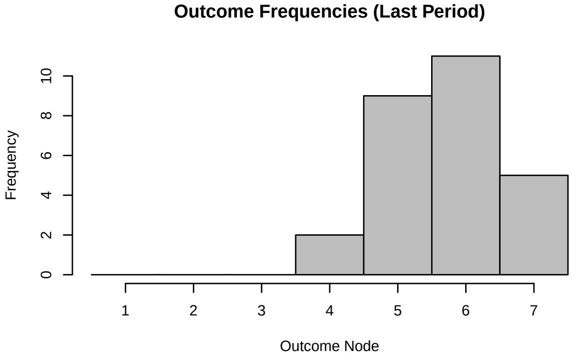
CENT-LO:
"Risk \$1 to gain \$5"



The Story

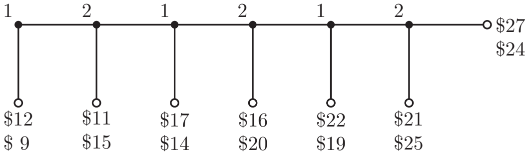
CENT-LO:
"Risk $1 to gain $5"

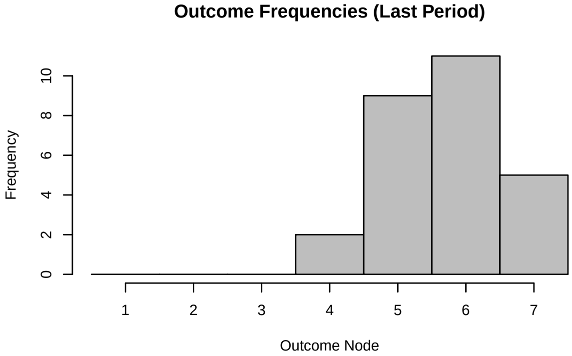




The Story

CENT-LO:
"Risk \$1 to gain \$5"

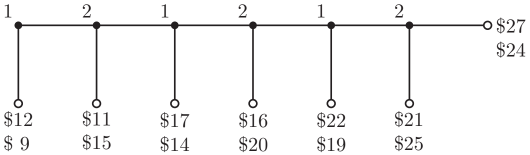




1. Node 7 freq
2. Story???

The Story

CENT-LO:
"Risk $1 to gain $5"

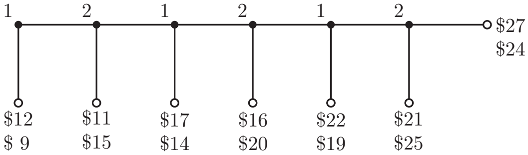


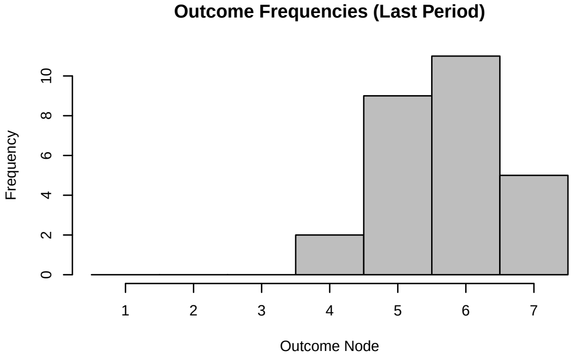
What do we learn from elicitation?

1. There are altruists who prefer Pass even if opponent will Take
    - Many people will give up $1 to give $6
2. Selfish people know that altruists are common
3. *Early nodes:* Selfish people Pass, knowing altruists Pass back
4. *Later nodes:* Selfish people Take. Altruists keep Passing

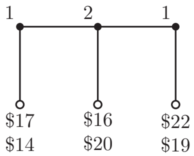
The Story

CENT-LO:
"Risk $1 to gain $5"

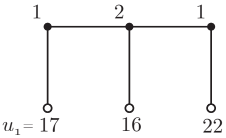




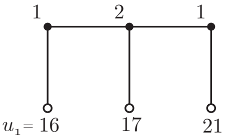
The Unit of Analysis: 3-Node Segments



A 3-Node Segment



Selfish u_1



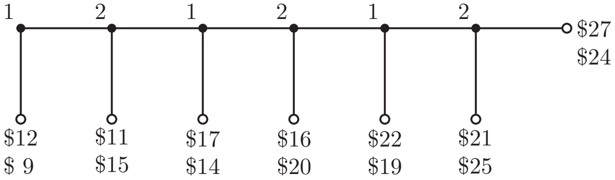
Altruist u_1

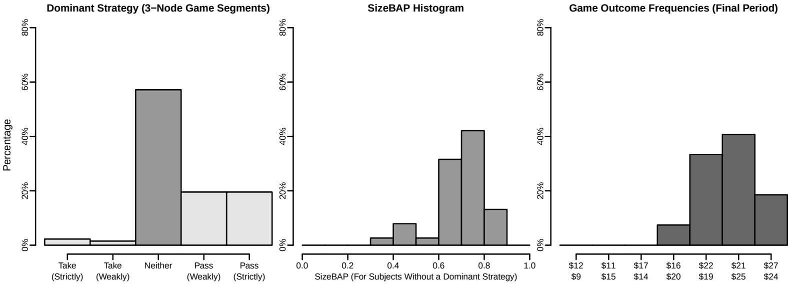
SizeBAP: A measure of the temptation to Pass

- Let $p =$ subjective prob. next mover will Pass
- **Selfish:** Pass is BR if $p \in [1/6, 1]$
  - SizeBAP for this u_1 is 5/6. Very likely to Pass.
- **Altruist:** Pass is BR if $p \in [0, 1]$ (strict Dom.Strat.)
  - SizeBAP for this u_1 is 1. Guaranteed to Pass.
- SizeBAP is a statistic for u_i (and nothing else)

Pooling All 3-Node Segments

CENT-LO Treatment:

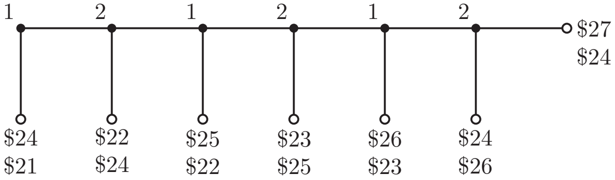


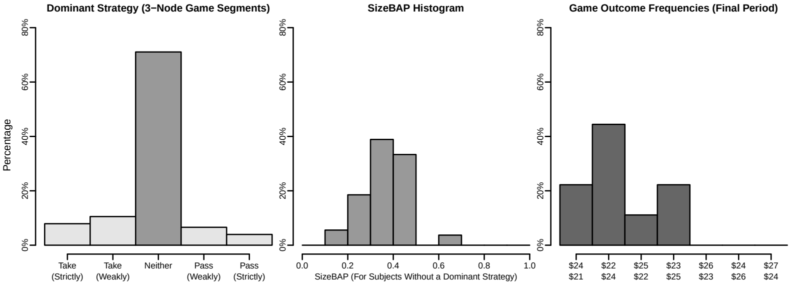


Selfish SizeBAP $\approx$ 0.833

CENT-HI

CENT-HI Treatment:

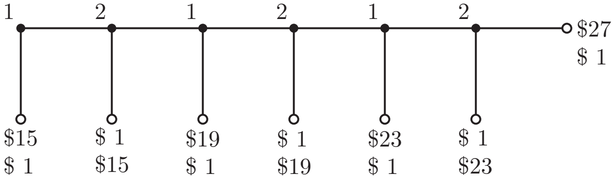


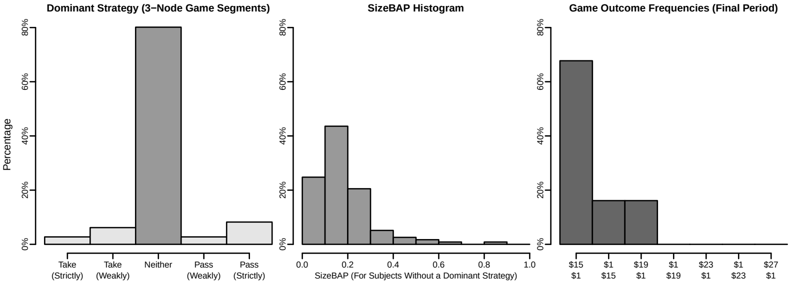


Selfish SizeBAP ≈ 0.333

CENT-ALL

CENT-ALL Treatment:





Selfish SizeBAP $\approx 0.22, 0.18, 0.15$

Verifying the Story: CENT-LO

CENT-LO:

1. Altruists exist
   - Pass is DomStrat in 43.7% of segments
2. Altruists pass
   - 89.7% of the time
   - $43.7\% \times 89.7\% = 39.2\%$ overall chance of Pass from altruists
3. Non-altruists believe Pass is reasonably likely
   - 54.8% have $Pr(Pass) > 39.2\%$ (median = 40%)
   - Self-similarity hides direct belief in altruism
4. Non-altruists BR to that belief
   - 83.8% play BR, given p_i^1 and u_i

Verifying the Story: CENT-HI

CENT-HI:

1. Altruists don't exist
   - Pass is DomStrat in 8.9% of segments
2. Altruists pass but they're very rare
   - Small sample: 6 out of 9
   - $8.9\% \times 66.6\% = 5.9\%$ overall chance of Pass
3. Non-altruists believe Pass is reasonably unlikely
   - Median = 20%
   - Self-similarity hides direct belief in altruism
4. Non-altruists BR to that belief
   - 58.5% play BR, given p_i^1 and u_i
   - Beliefs only elicited for those that Pass, which is a small sample

Verifying the Story: CENT-ALL

CENT-ALL:

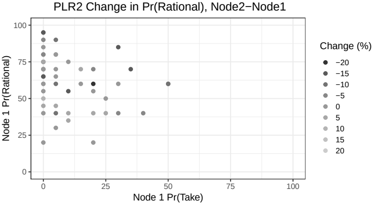
1. Altruists don't exist
    - Pass is DomStrat in 8.7% of segments
2. Altruists pass but they're very rare
    - Small sample: 2 out of 12
    - $8.7\% \times 16.67\% = 1.45\%$ overall chance of Pass
3. Non-altruists believe Pass is reasonably unlikely
    - Median = 17.5%
    - Self-similarity hides direct belief in altruism
4. Non-altruists BR to that belief
    - 38.3% play BR, given p_i^1 and u_i
    - Beliefs only elicited for those that Pass, which is a small sample

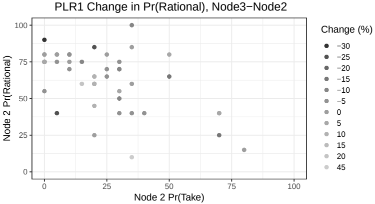
Belief in Rationality & Backward Induction

- Does common belief in rationality $\Rightarrow$ backwards induction?
- Depends how people react to surprises (Reny 1993)
  - RCSBR: continue to believe in rationality after surprises
  - (Surprises $\Rightarrow$ belief in irrationality) $\Rightarrow$ Surprises!
- Surprise: Pr(Take)=100%, Pr(Rational)=100%, but then Pass

Belief in Rationality & Backward Induction

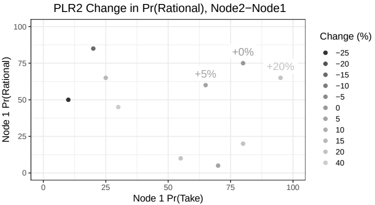
- Does common belief in rationality $\Rightarrow$ backwards induction?
- Depends how people react to surprises (Reny 1993)
  - RCSBR: continue to believe in rationality after surprises
  - (Surprises $\Rightarrow$ belief in irrationality) $\Rightarrow$ Surprises!
- Surprise: Pr(Take)=100%, Pr(Rational)=100%, but then Pass
- CENT-LO: Pr(Take) never near 100%
  - It's *not* a game of complete information!

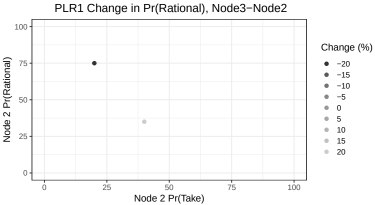




Belief in Rationality & Backward Induction

- Does common belief in rationality $\Rightarrow$ backwards induction?
- Depends how people react to surprises (Reny 1993)
  - RCSBR: continue to believe in rationality after surprises
  - (Surprises $\Rightarrow$ belief in irrationality) $\Rightarrow$ Surprises!
- Surprise: Pr(Take)=100%, Pr(Rational)=100%, but then Pass
- CENT-ALL: Very few surprises since everyone Takes!
  - Unsurprisingly, surprises are rare





Prisoners Dilemma

The Prisoners' Dilemma

	C	D
C	\$10, \$10	\$1, \$15
D	\$15, \$1	\$5, \$5

The Prisoners' Dilemma Game Form

- New treatment: SIM
- Five 2×2 games without feedback, random matching
- Elicitation in every game
- Pencil & paper
- $n = 150$

The Prisoners' Dilemma

	C	D
C	\$10, \$10	\$1, \$15
D	\$15, \$1	\$5, \$5

The Prisoners' Dilemma Game Form

30.4% play C.

Why???

The Prisoners' Dilemma

	C	D
C	\$10, \$10	\$1, \$15
D	\$15, \$1	\$5, \$5

The Prisoners' Dilemma Game Form

				$BR_i(p_i^{1s}\|u_i) = C$		$BR_i(p_i^{1s}\|u_i) = D$	
Pref. Type	$BR_i(C)$	$BR_i(D)$	% Subj.	$s_i = C$	$s_i = D$	$s_i = C$	$s_i = D$
Selfish	D	D	68.0%	–	–	18	**79**
Cond. Coop.	C	D	19.7%	15	5	3	6
Reverse	D	C	2.7%	1	2	0	1
Uncond. Coop.	C	C	9.5%	8	6	–	–

The Prisoners' Dilemma

	C	D
C	\$10, \$10	\$1, \$15
D	\$15, \$1	\$5, \$5

The Prisoners' Dilemma Game Form

				$BR_i(p_i^{1s}\|u_i) = C$		$BR_i(p_i^{1s}\|u_i) = D$	
Pref. Type	$BR_i(C)$	$BR_i(D)$	% Subj.	$s_i = C$	$s_i = D$	$s_i = C$	$s_i = D$
Selfish	D	D	68.0%	–	–	18	**79**
Cond. Coop.	C	D	19.7%	15	5	3	6
Reverse	D	C	2.7%	1	2	0	1
Uncond. Coop.	C	C	9.5%	8	6	–	–

Only 53% of cooperation (C) is rational

Failure of consequentialism or dominance

Iterated Dominance

A Dominance-Solvable Game Form

	L	R
U	\$10, \$ 5	\$15, \$15
D	\$5, \$10	\$1, \$1

A Dominance Solvable Game Form

- Row players: 100% play U
  - 71 of 75: U is a dominant strategy
  - 4 of 75: U is a best response
- Column players: 25% play L
  - Why???

A Dominance-Solvable Game Form

	L	R
U	\$10, \$ 5	\$15, \$15
D	\$5, \$10	\$1, \$1

A Dominance Solvable Game Form

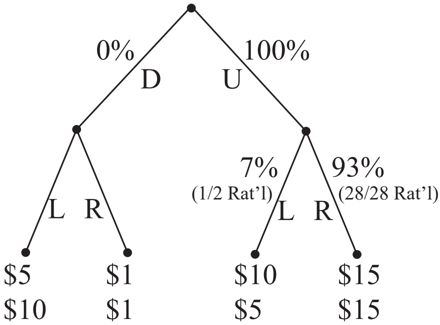
				$BR_i(p_i^{1s}\|u_i) = L$		$BR_i(p_i^{1s}\|u_i) = R$	
Pref. Type	$BR_i(U)$	$BR_i(D)$	% Subj.	$s_i = L$	$s_i = R$	$s_i = L$	$s_i = R$
Selfish	R	L	91.9%	0	0	14	53
DomStrat L	L	L	5.4%	3	1	–	–
DomStrat R	R	R	2.7%	–	–	1	1
Reversed	L	R	0%	0	0	0	0

Violation of consequentialism and/or EU

Conjecture: avoiding (\$1, \$1), despite stated preferences. Strategic uncertainty.

Sequential-Move DomSolv

SEQ treatment: $n = 60$



Irrationality *disappears* when strategic uncertainty is removed

Coordination

Asymmetric Coordination

	L	R
U	\$15, \$ 5	\$2, \$1
D	\$1, \$2	\$5, \$10

An Asymmetric Coordination Game Form

- Row: 93% play U
- Col: 49.3% play L
    - Why??? Beliefs?

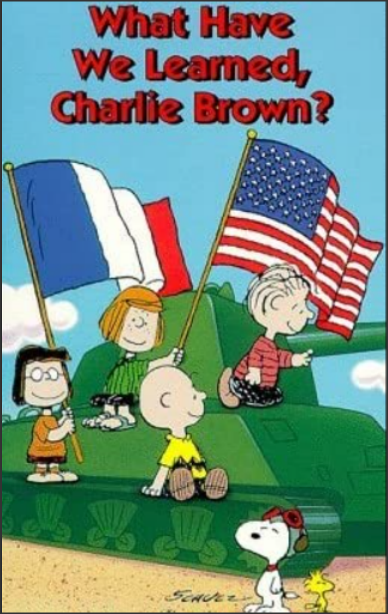
Asymmetric Coordination

	L	R
U	\$15, \$ 5	\$2, \$1
D	\$1, \$2	\$5, \$10

An Asymmetric Coordination Game Form

				$BR_1(p_1^{1S}\|u_1) = U$		$BR_1(p_1^{1S}\|u_1) = D$	
Row's Type	$BR_1(L)$	$BR_1(R)$	% Subj.	$s_1 = U$	$s_1 = D$	$s_1 = U$	$s_1 = D$
Selfish	U	D	95.8%	43	2	20	3
DomStrat U	U	U	4.2%	3	0	–	–

				$BR_2(p_2^{1S}\|u_2) = L$		$BR_2(p_2^{1S}\|u_2) = R$	
Col's Type	$BR_2(U)$	$BR_2(D)$	% Subj.	$s_2 = L$	$s_2 = R$	$s_2 = L$	$s_2 = R$
Selfish	L	R	93.0%	26	27	5	8
DomStrat L	L	L	7.0%	5	0	–	–


What Have
We Learned,
Charlie Brown?

Lessons

1. Most experiments are Bayesian games, not complete info
2. The story changes from one game to the next
3. Centipede game forms:
    - Altruists pass $\Rightarrow$ selfish pass
    - Backwards induction seems to work fine here
4. Prisoners' dilemma:
    - Non-consequential preference for cooperating
5. Beliefs are generally pretty accurate
6. Don't write a solo-authored paper post-tenure