Testing Elicitation Mechanisms via Team Chat

Paul J. Healy (OSU)
John Kagel (OSU)

Introduction

Suppose you need to elicit people's beliefs.

There are many mechanisms

1. Which incentive mechanism is "best"? Is there consensus?
2. Theoretically, under what assumptions is that mechanism IC?
3. Empirically, do we know that it's actually IC?
    - Testing IC requires that we know their true beliefs!
4. How frequently do subjects actually misreport?
    - Danz, Vesterlund & Wilson (2022)

Introduction

Suppose you need to elicit people's beliefs.

There are many mechanisms

1. Which incentive mechanism is "best"? Is there consensus?
2. Theoretically, under what assumptions is that mechanism IC?
3. Empirically, do we know that it's actually IC?
    - Testing IC requires that we know their true beliefs!
4. How frequently do subjects actually misreport?
    - Danz, Vesterlund & Wilson (2022)

This paper: An attempt to answer these questions

There Are Many Mechanisms

Popular mechanisms:

- Unincentivized
  - More noise, more default answers like 50%
- Quadratic scoring rule (“QSR”; Brier 1950)
  - Other scoring rules: Logarithmic, spherical…
  - Can correct for risk aversion (Harrison et al. 2014)
- “Binarized” scoring rules (BSR; Savage 1971, Hossain&Okui 2013)
- BDM for probabilities (Marschak 1963, Grether 1981)
  - Ascending-clock BDM (Karni 2009)
  - Multiple Price List (“MPL”; Holt & Smith 2016)

How do these perform in the lab?

How Do These Perform in the Lab?

- Offerman & Sonnemans (2004): QSR$\sim$None
- Armantier & Treich (2013): QSR$\succ$None
- Huck & Weizsacker (2002): QSR$\succ$BDM
- Hollars et al. (2010): BDM$\succ$QSR
- Hao & Houser (2012): BDM-Clock$\succ$BDM
- Hossain & Okui (2013): BSR$\succ$QSR
- Harrison et al. (2014): BSR$\sim$QSR-Corr$\succ$QSR
- Wilson & Vespa (2017): BSR$\succ$PU-BSR
- Holt & Smith (2016); MPL$\succ$BDM

BSR and MPL have never "lost", but haven't been compared

Our Motivations

- Offerman & Sonnemans (2004): QSR$\sim$None
- Armantier & Treich (2013): QSR$\succ$None
- Huck & Weizsacker (2002): QSR$\succ$BDM
- Hollars et al. (2010): BDM$\succ$QSR
- Hao & Houser (2012): BDM-Clock$\succ$BDM
- Hossain & Okui (2013): BSR$\succ$QSR
- Harrison et al. (2014): BSR$\sim$QSR-Corr$\succ$QSR
- Wilson & Vespa (2017): BSR$\succ$PU-BSR
- Holt & Smith (2016); MPL$\succ$BDM

Motivation #1: Compare MPL to BSR, both in theory and in the lab

So, What Are the BSR and MPL?

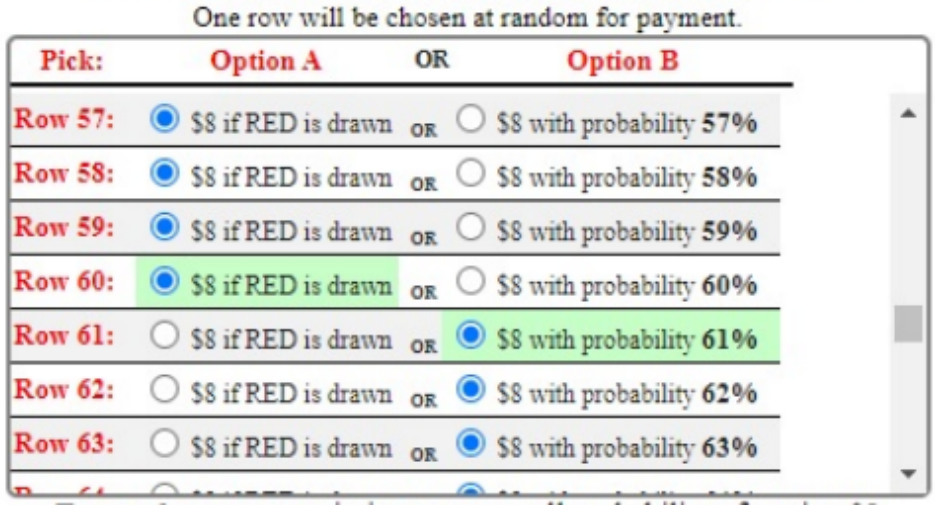
Suppose event is "Red Jar was chosen". Subject's true belief: p

BQSR

- Report: q (may not be p)
- Payment if Red Jar:
  $1 - (1 - q)^2$
- Payment if Blue Jar:
  $1 - (0 - q)^2$
- Payment is not money, but % chance of \$8
- $Pr(\$8) = p \cdot [1 - (1 - q)^2] + (1 - p) \cdot [1 - (0 - q)^2]$

MPL

- Pick a switch point in this list:


One row will be chosen at random for payment.

Pick:	Option A	OR	Option B
Row 57:	◉ \$8 if RED is drawn	OR	○ \$8 with probability 57%
Row 58:	◉ \$8 if RED is drawn	OR	○ \$8 with probability 58%
Row 59:	◉ \$8 if RED is drawn	OR	○ \$8 with probability 59%
Row 60:	◉ \$8 if RED is drawn	OR	○ \$8 with probability 60%
Row 61:	○ \$8 if RED is drawn	OR	◉ \$8 with probability 61%
Row 62:	○ \$8 if RED is drawn	OR	◉ \$8 with probability 62%
Row 63:	○ \$8 if RED is drawn	OR	◉ \$8 with probability 63%

- Interpret row # as report q

How Can You Test if an Elicitation Mechanism Works??

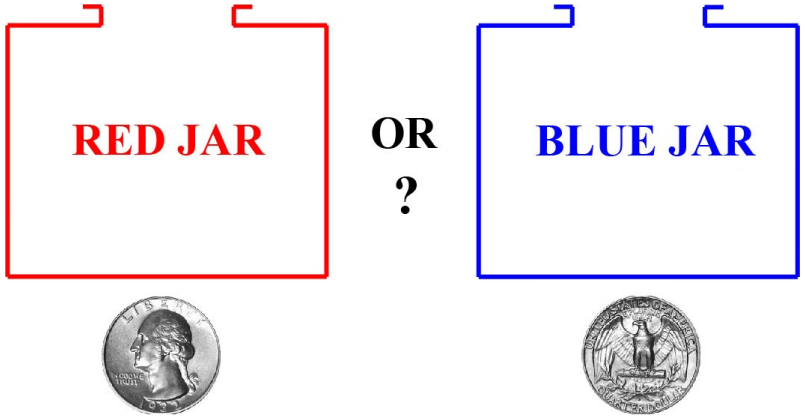
Motivation #2: Experiments testing elicitation are... tricky

- Need to know their belief to test whether they report truthfully

Example: Objective-Easy Questions

Holt & Smith (2016), Danz et al. (2022), etc.

Easy Task:



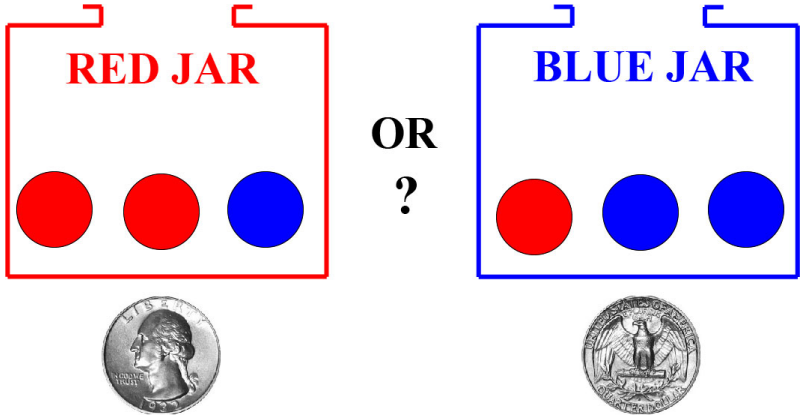
Pro: Almost certainly know their belief

Con: Too suspicious! “Deviation” might be distrust, confusion

Example: Objective-Hard Questions

Holt & Smith (2016), Danz et al. (2022), etc.

Hard Task:



Signal: Two BLUE marbles were drawn w/ replacement

Pro: Less suspicious

Con: Too hard! "Deviation" might be confusion, errors

Wouldn't it be nice to have a machine that lets us see true beliefs???

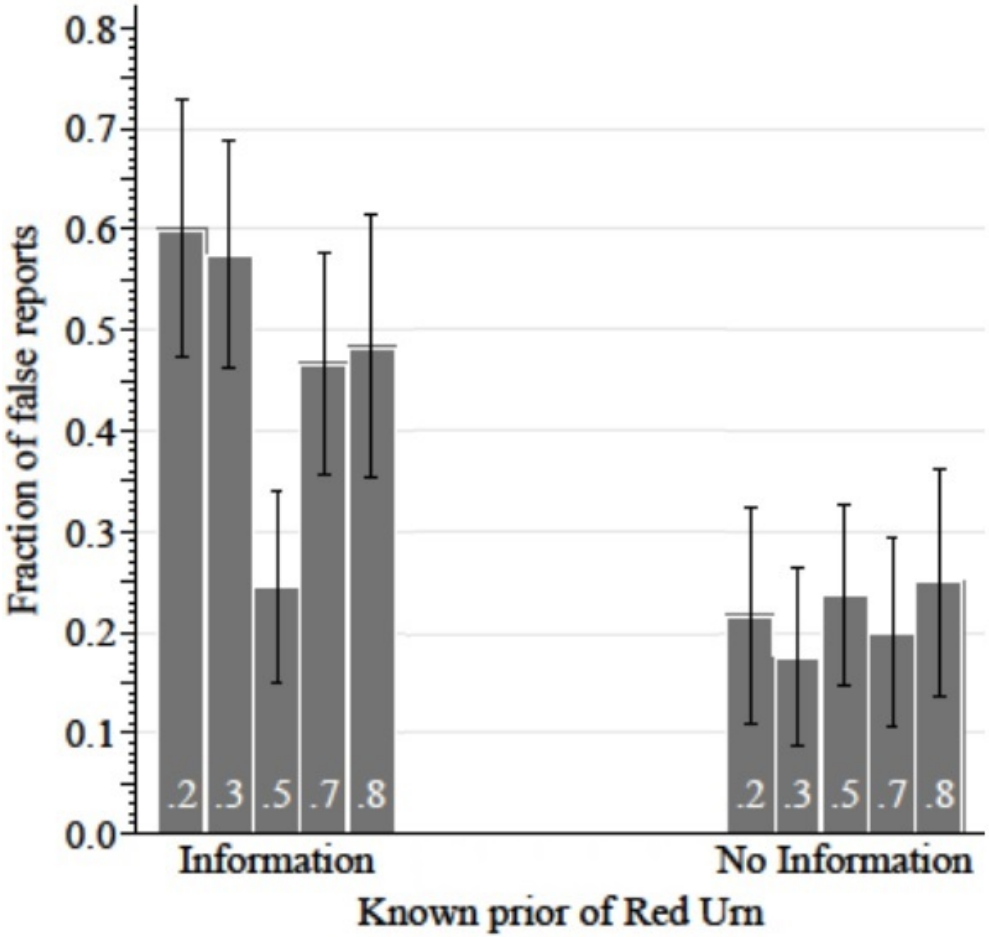
Our Project

John Kagel has such a machine!! The **team chat** protocol

- Have subjects in teams of two, working together via chat
  - Cooper & Kagel (2005,2009,2020)
- Scan chat transcripts for (1) true beliefs, (2) manipulation
- Can do this with *any type* of question
  - Easy, hard, subjective
  - Probabilities, means, medians
- Compare **BSR** to **MPL**
- For fun, we also test a **Non-IC** mechanism

Danz, Vesterlund & Wilson (2022)

Motivation #3: Is it really better to hide incentives?



- NoInfo ↓ misreports!
- We add a **NoInfo** treatment to check this

Our Findings

Theory Result (roughly):

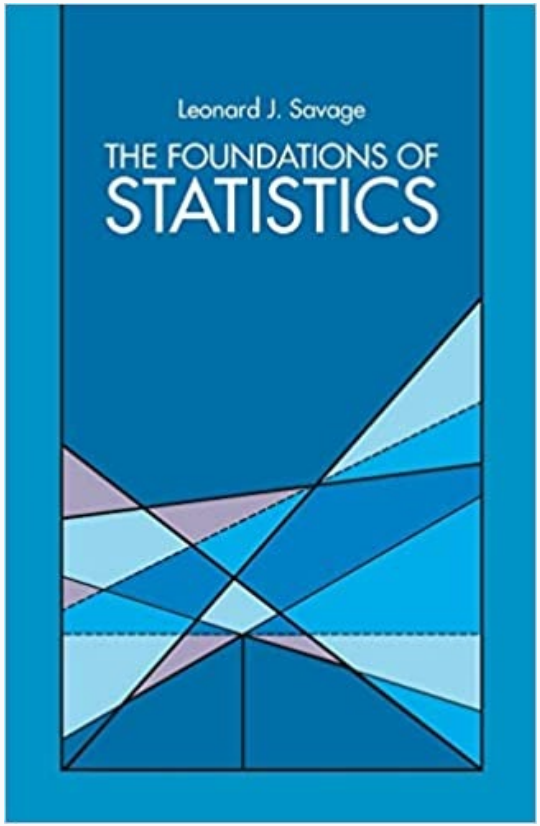
1. BSR is IC $\Rightarrow$ MPL is IC

Experimental Results:

1. Easy questions: BSR & MPL misreporting rates are very low
    - Very different from Danz et al. (2022)
2. NoInfo also does great, but so do MPL and BSR...
3. *Very* little evidence of manipulation in the chat
    - Sliiiightly more in BSR than MPL
4. Hard questions: misreporting is due to confusion and mistakes
5. Non-IC mechanism: Most people still tell the truth!
6. Danz et al. replication: Results sensitive to interface

Theory

Theory: Savage (1971)



(1954)

© Journal of the American Statistical Association
December 1971, Volume 66, Number 336
Theory and Methods Section

Elicitation of Personal Probabilities and Expectations

LEONARD J. SAVAGE*

Proper scoring rules, i.e., devices of a certain class for eliciting a person's probabilities and other expectations, are studied, mainly theoretically but with some speculations about application. The relation of proper scoring rules to other economic devices and to the foundations of the personalistic theory of probability is brought out. The implications of various restrictions, especially symmetry restrictions, on scoring rules is explored, usually with a minimum of regularity hypothesis.

1. INTRODUCTION

1.1 Preface

This article is about a class of devices by means of which an idealized *homo economicus*—and therefore, with some approximation, a real person—can be induced to reveal his opinions as expressed by the probabilities that he associates with events or, more generally, his personal expectations of random quantities. My emphasis here is theoretical, though some experimental considerations will be mentioned. The empirical importance of such studies in many areas is now recognized. It was emphasized for the area of economics in an address by Trygve Haavelmo [28, p. 357]:

pertaining to it has grown up, some of which will be cited in context and most of which can be found through the references cited, especially the recent and extensive [52] and others that I call "key references."

Bruno de Finetti and I began to write the present article in the spring of 1960, not yet aware of our predecessors and contemporaries. The impetus was de Finetti's, for he had brought us to rediscover McCarthy's [37] insight about convex functions. We expected to make short work of our "little note," but it grew rapidly in many directions and became inordinately delayed. Now we find that the material in the present article is largely mine and that de Finetti has published on diverse aspects of the same subject elsewhere [12, 13, 14, 17]. De Finetti has therefore withdrawn himself from our joint authorship and encouraged me to publish this article alone, though it owes so much to him at every stage, including the final draft.

The article is written for a diverse audience. Consequently, some will find parts of it mathematically too

(1971)

The Binarized Quadratic Scoring Rule (BSR)

Event: $X \in \{0, 1\}$
Subjective belief: $p = \Pr(X = 1)$

Binarized Quadratic Scoring Rule (Savage 1971; Hossai & Okui 2013):

- Fixed prize ($8)
- Announce subjective belief q
- Paid objective lotteries:
  - \$8 w/ prob $s_1(q) = 1 - (1 - q)^2$ if $X = 1$
  - \$8 w/ prob $s_0(q) = 1 - (0 - q)^2$ if $X = 0$

Conditions for Incentive Compatibility

Proof of Incentive Compatibility requires reduction:

$$U(q|p) = \underbrace{p \cdot s_1(q) + (1-p) \cdot s_0(q)}_{\text{Overall } Pr(\$8) \text{ when announcing } q}$$

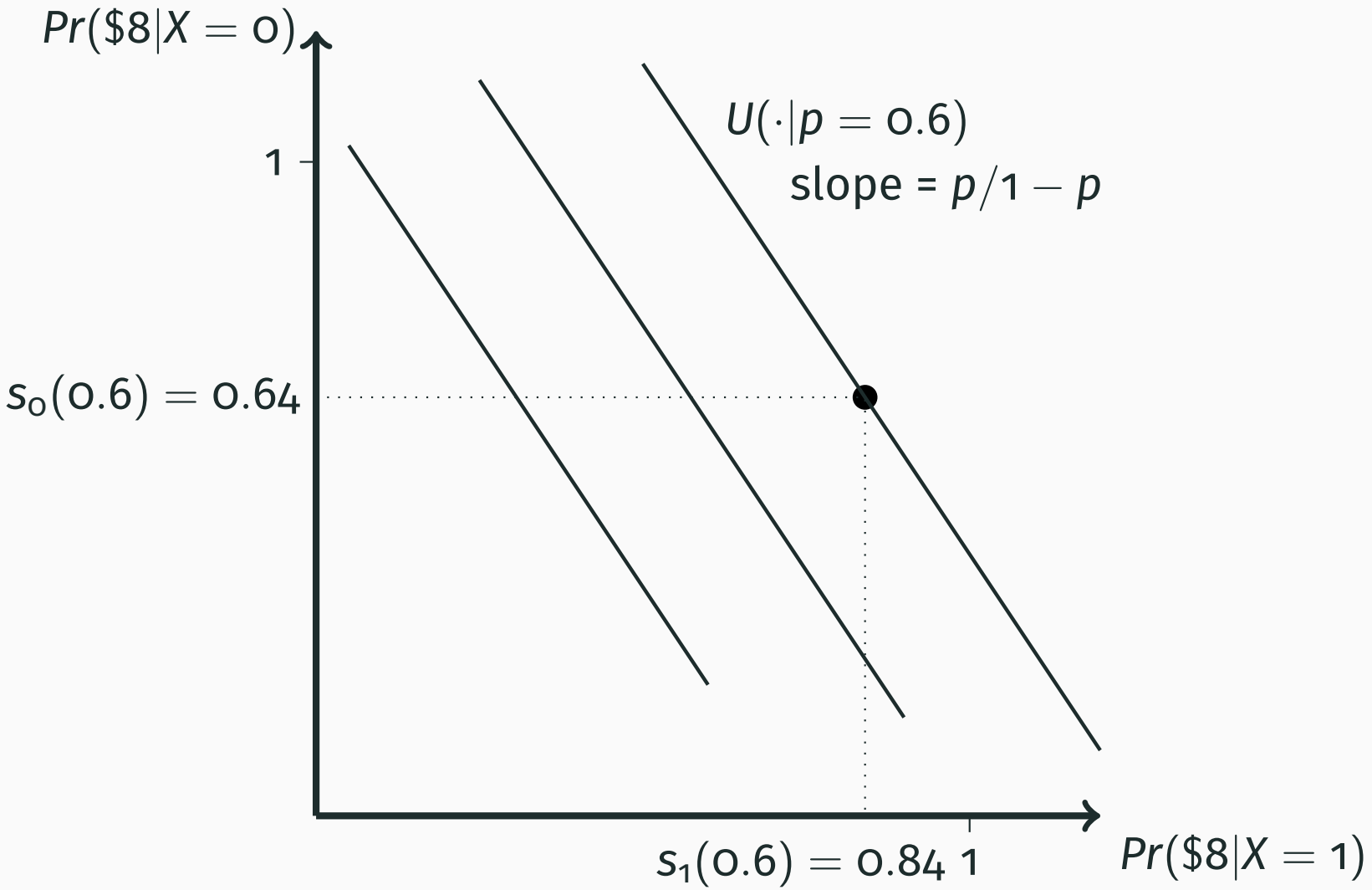
$$= \underbrace{p \cdot [1-(1-q)^2] + (1-p) \cdot [1-(0-q)^2]}_{\text{Plugging in BQSR formulas}}$$

"Subjective-Objective Reduction"

- "Probabilistic sophistication" (Machina & Schmeidler 1995)
- Weakening of ROCL (binary lotteries only)
- But, requires integration of subj. & obj. uncertainty

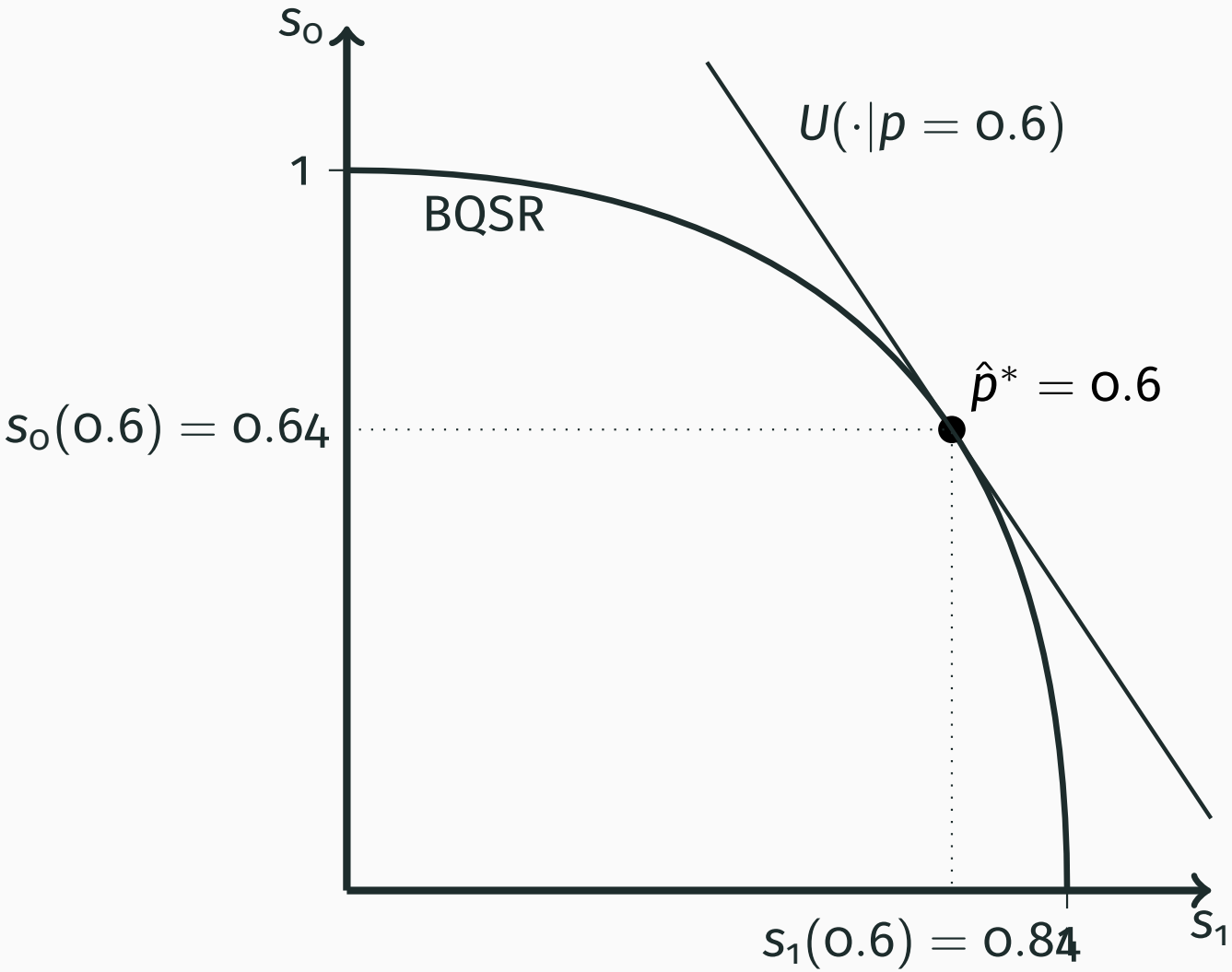
IC under S-O reduction because $U(q|p)$ is max'd at $q^* = p$

Visualizing IC of BSR



S-O Reduction $\Rightarrow$ linear indiff curves, slope $p/1-p$

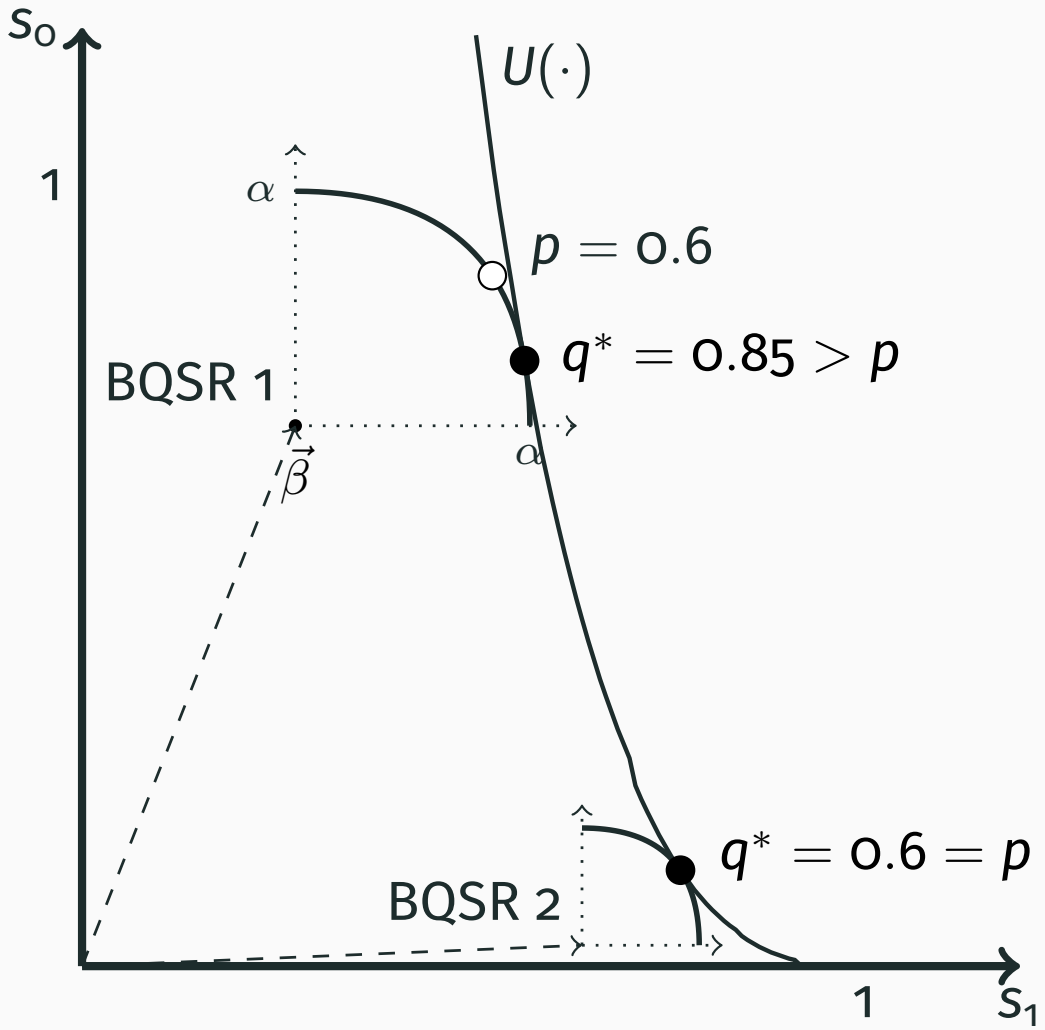
Visualizing IC of BSR



Slope of scoring rule at p is $p/1-p$

Necessity of S-O Reduction

Scaled BQSR: $\vec{\beta} + \alpha(1 - (X - q)^2)$



Violate SO Reduction $\Rightarrow$ Non-linear $U(\cdot) \Rightarrow \exists$ scaled BSR that's not IC

A Useful Lemma

Lemma
*If **every** scaled BSR is IC then S-O Reduction must be satisfied.*

A Reason for Non-Linearity

Asymmetric responsiveness (Danz, Vesterlund & Wilson 2022)

q	$Pr(\$8 \mid X{=}0)$	$Pr(\$8 \mid X{=}1)$
0.05	99.75%	9.75%
0.10	99%	19%
0.15	97.75%	27.75%
0.20	96%	36%
0.25	93.75%	43.75%

- Data: push towards $q = 50$
  - $\Rightarrow$ convex $U(\cdot) \Rightarrow$ violation of S-O Reduction
- But... *any* IC mechanism must have this asymmetry
  - FOC: $p\, s_1'(p) + (1-p)\, s_0'(p) = 0 \quad \Rightarrow \quad -\frac{s_0'(p)}{s_1'(p)} = \frac{p}{1-p}$

Multiple Price Lists (MPL)

Row#	Option A	OR	Option B
1	\$8 if $X = 1$	or	\$8 w/ prob 1%
2	\$8 if $X = 1$	or	\$8 w/ prob 2%
⋮	⋮	⋮	⋮
q	\$8 if $X = 1$	or	\$8 w/ prob q%
$q+1$	\$8 if $X = 1$	or	\$8 w/ prob $q+1$%
$q+2$	\$8 if $X = 1$	or	\$8 w/ prob $q+2$%
$q+3$	\$8 if $X = 1$	or	\$8 w/ prob $q+3$%
⋮	⋮	⋮	⋮
99	\$8 if $X = 1$	or	\$8 w/ prob 99%
100	\$8 if $X = 1$	or	\$8 w/ prob 100%

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

One row randomly selected for payment

Multiple Price Lists (MPL)

Row#	Option A	OR	Option B
1	\$8 if $X = 1$	or	\$8 w/ prob 1%
2	\$8 if $X = 1$	or	\$8 w/ prob 2%
⋮	⋮	⋮	⋮
q	\$8 if $X = 1$	or	\$8 w/ prob q%
$q + 1$	\$8 if $X = 1$	or	\$8 w/ prob $q + 1$%
$q + 2$	\$8 if $X = 1$	or	\$8 w/ prob $q + 2$%
$q + 3$	\$8 if $X = 1$	or	\$8 w/ prob $q + 3$%
⋮	⋮	⋮	⋮
99	\$8 if $X = 1$	or	\$8 w/ prob 99%
100	\$8 if $X = 1$	or	\$8 w/ prob 100%

"Multiple Price List" (MPL) version of BDM for probabilities
Holt & Smith (2016), Healy (2018)

Multiple Price Lists (MPL)

Row#	Option A	OR	Option B
1	\$8 if $X = 1$	or	\$8 w/ prob 1%
2	\$8 if $X = 1$	or	\$8 w/ prob 2%
⋮	⋮	⋮	⋮
q	\$8 if $X = 1$	or	\$8 w/ prob q%
$q+1$	\$8 if $X = 1$	or	\$ w/ prob $q+1$%
$q+2$	\$8 if $X = 1$	or	\$ w/ prob $q+2$%
$q+3$	\$8 if $X = 1$	or	\$8 w/ prob $q+3$%
⋮	⋮	⋮	⋮
99	\$8 if $X = 1$	or	\$8 w/ prob 99%
100	\$8 if $X = 1$	or	\$8 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

MPL vs BSR

Theorem:

All Scaled BQSRs are I.C.

Subjective-Objective Reduction

Statewise Dominance

MPL is I.C.

*Assume if reduction occurs, it occurs at all levels

RDU

What about Rank-Dependent Utility / Probability Weighting?

BSR:

$$w_1[p]w_2[(1-(1-p)^2)] + w_1[1-p]w_2[(1-(0-p)^2)] \quad \times$$

vs.

$$w_0[p(1-(1-p)^2) + (1-p)(1-(0-p)^2)] \quad \checkmark$$

MPL:

$$w_1(p) \neq w_2(p) \quad \times$$

vs.

$$w_0(p) = w_0(p) \quad \checkmark$$

More Than Two States

- What if X can take more values?
  - Ex: score on a quiz, GDP next quarter
- Could elicit $Pr(X = x)$ for every possible x... but that's a lot!
- The BQSR elicits the subject's **mean** for X
  - Announce mean m
  - $Pr(\$8) = \left(1 - (x - m)^2\right)$
  - Still paying in probabilities (rescale X to $[0, 1]$)
  - Still requiring S-O Reduction:

$$Pr(\$8|m) = \sum_x Pr(X = x)(1 - (x - m)^2)$$

- *Note:* with two states, mean = probability
- Is there an MPL for the mean?

MPL for The Mean of X

Row#	Option A	OR	Option B
1	X% chance of \$8	or	1% chance of \$8
2	X% chance of \$8	or	2% chance of \$8
⋮	⋮	⋮	⋮
m	X% chance of \$8	or	m% chance of \$8
m+1	X% chance of \$8	or	m+1% chance of \$8
⋮	⋮	⋮	⋮
99	X% chance of \$8	or	99% chance of \$8
100	X% chance of \$8	or	100% chance of \$8

Requires S-O Reduction: "X% chance" $\sim$ "$E[X]$% chance"

Eliciting the Median

- BSR elicits the mean... can we elicit the median?
- **Linear** scoring rule elicits the median!
- BLSR:

$$Pr(\$8|m) = (1 - |x - m|)$$

- *Note:* with two states, median = mode
- Is there an MPL?

MPL for The Median of X

Row#	Option A	OR	Option B
1	\$8 if $X \geq 1$	or	50% chance of \$8
2	\$8 if $X \geq 2$	or	50% chance of \$8
⋮	⋮	⋮	⋮
m	\$8 if $X \geq m$	or	50% chance of \$8
m+1	\$8 if $X \geq m$+1	or	50% chance of \$8
⋮	⋮	⋮	⋮
99	\$8 if $X \geq 99$	or	50% chance of \$8
100	\$8 if $X \geq 100$	or	50% chance of \$8

Does *NOT* require S-O Reduction

Easily altered to elicit any quantile

Summary

- Summary:

  Probability: BQSR vs. Probability MPL

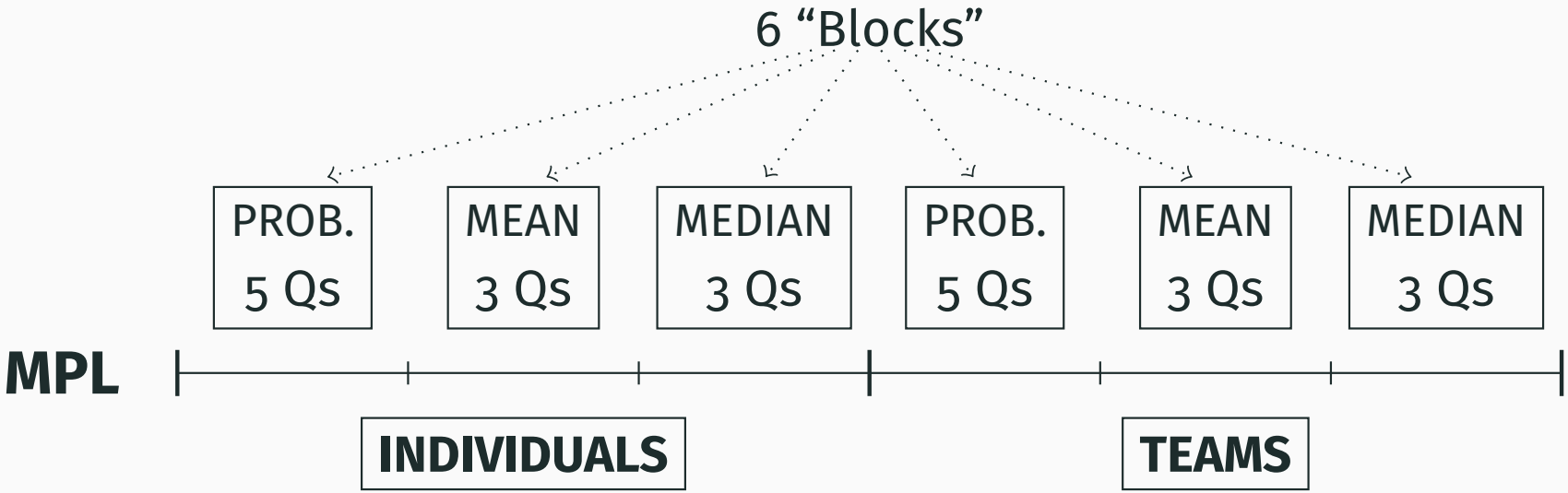
  Mean: BQSR vs. Mean MPL

  Median: BLSR vs. Median MPL

- MPL: weaker assumption for IC (except for the mean)

Experimental Design

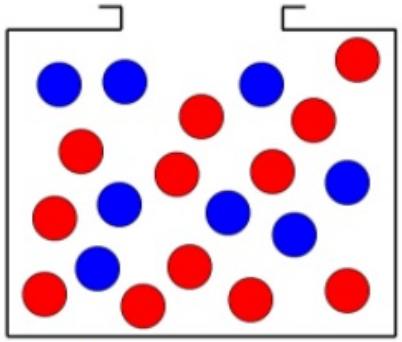
Experimental Design



- Each block has 3 or 5 questions of the same type
- Instructions before each block
- Order of blocks randomized within INDIV and TEAM
- Order of questions randomized within each block
- Three mechanisms: **MPL**, **BSR**, **NoInfo**
    - Each subject sees only one mechanism
- INDIV first vs TEAMS first: no difference

The 11 Questions

This jar contains red and blue marbles.



The computer will randomly draw *one* marble from this jar.

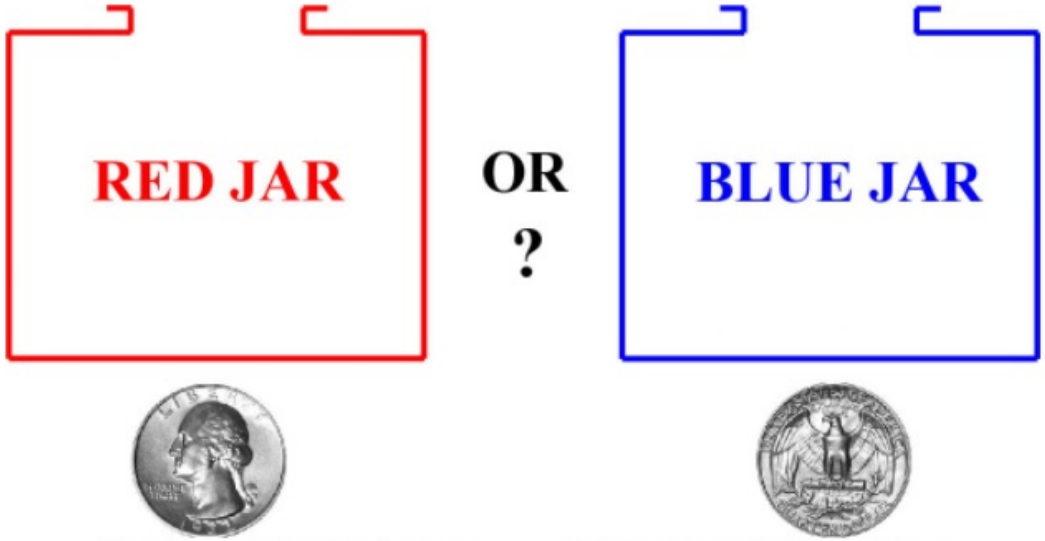
Q1: How many RED marbles are there in the jar? [ ] ($ if correct)

Q2: How many total marbles (of either color) are there in the jar? [ ] ($ if correct)

Q3: What do you think is the probability (from 0% to 100%) that a RED marble will be drawn? [ ] %

The 11 Questions

The computer will flip a coin to choose one of these two jars:



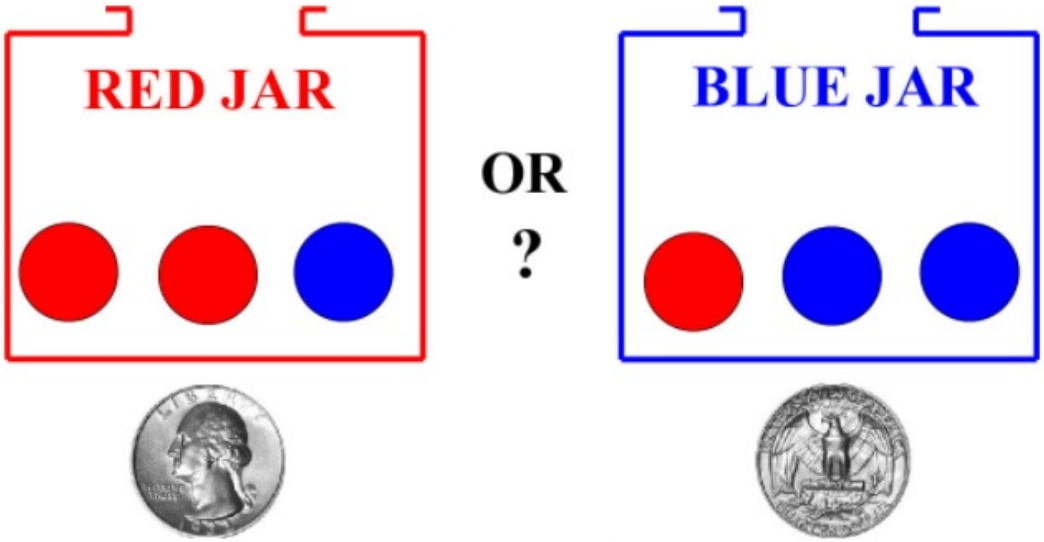
Heads: red jar is chosen.

Tails: blue jar is chosen.

Q1: What do you think is the probability (from 0% to 100%) that the RED JAR was chosen? ____ %

The 11 Questions

Again, one of two jars is chosen by a coin flip. But now the jars contain 3 marbles:



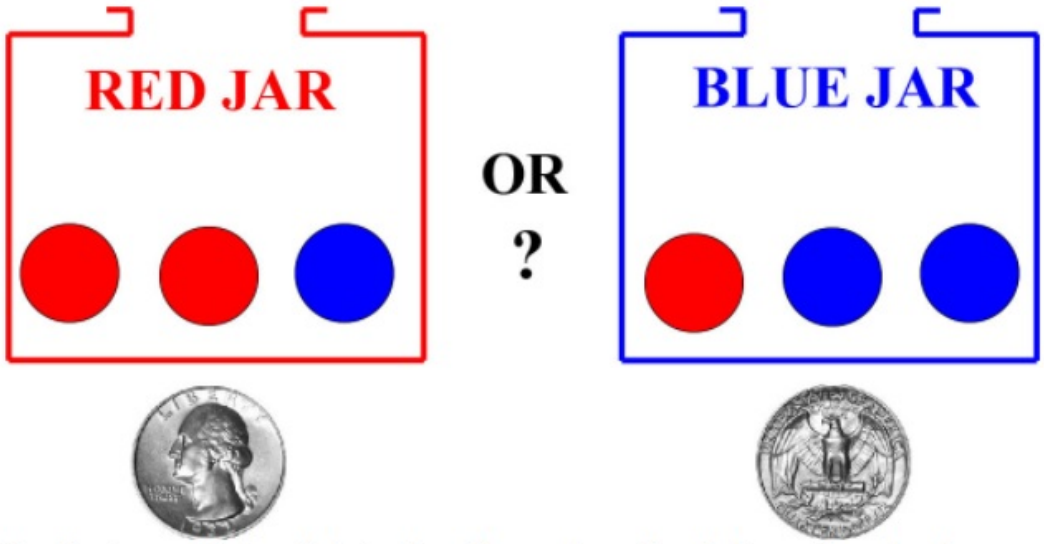
To give you a clue of which jar was chosen, we drew a marble from the chosen jar.

The marble drawn was a **BLUE** marble.

Q1: Now what do you think is the probability (from 0% to 100%) that the RED JAR was chosen? ____ %

The 11 Questions

Continuing on with the same chosen jar:



We put the first marble back into the chosen jar, shook it, and again drew a marble.

The second marble was also **BLUE**

(Thus, two **BLUE** marbles were drawn).

Q1: Now what do you think is the probability (from 0% to 100%) that the RED JAR was chosen? ____ %

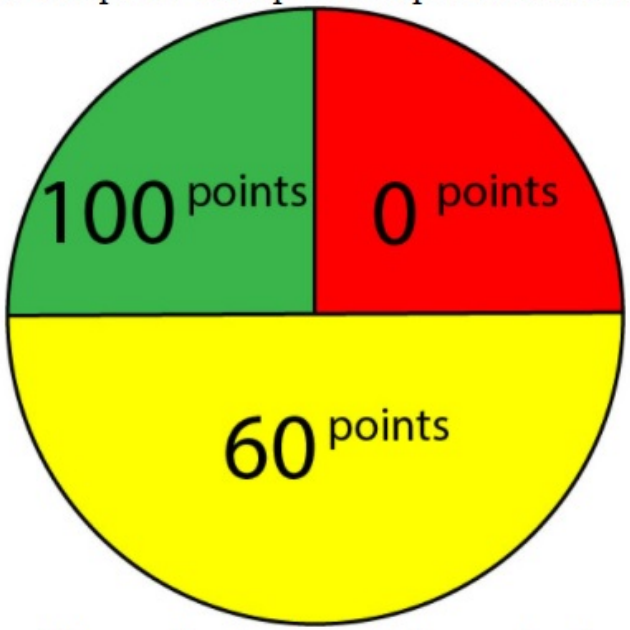
The 11 Questions

In 2005 we asked a Carnegie Mellon undergraduate this question:
What is the capital of Australia?

Q1: What do you think is the probability (from 0% to 100%) that they got this question right? ____ %

The 11 Questions

The computer will spin this spinner one time:

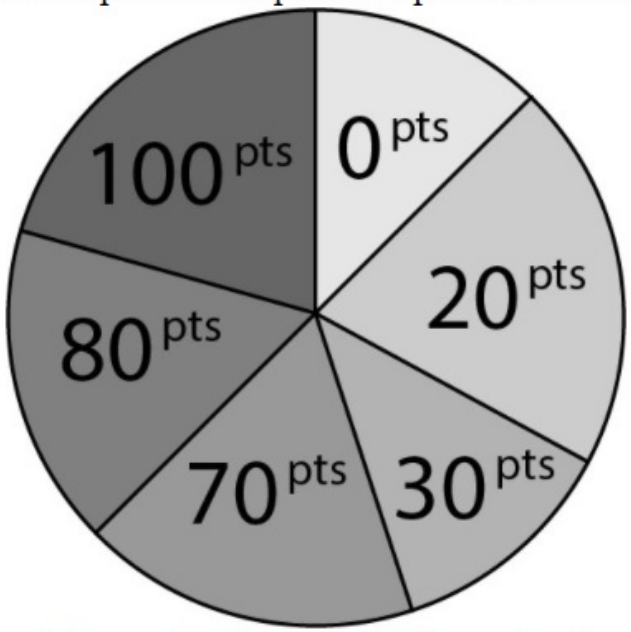


The *median* is the 'middle number.'
If the median is M, then you have ≥50% chance of getting ≥M points, *and* ≥50% chance of getting ≤M points.

Q1: I think the median # of points for this spinner is ____ pts

The 11 Questions

The computer will spin this spinner one time:



The *median* is the 'middle number.'
If the median is M, then you have ≥50% chance of getting ≥M points, *and* ≥50% chance of getting ≤M points.

Q1: I think the median # of points for this spinner is ____ pts

The 11 Questions

In 2005 we gave a Carnegie Mellon undergraduate student this quiz:

1. Who is credited with inventing the wristwatch in 1904?
2. Laudanum is a form of what drug?
3. The psychoactive ingredient in marijuana is THC. What does THC stand for?
4. What chemical element has the atomic number five?
5. The study of the structural and functional changes in cells, tissues and organs that underlie disease is called what?
6. What does the suffix -itis mean?
7. The bilby, bandicoot, and quokka are all representatives of what mammalian subclass?
8. Which one of the 50 United States is the only one never to have experienced an earthquake?
9. What evolutionary biologists wrote: *'Creation science' has not entered the curriculum for a reason so simple and so basic that we often mention it: because it is false.*?
10. What is the single most diverse phylum within the animal kingdom?

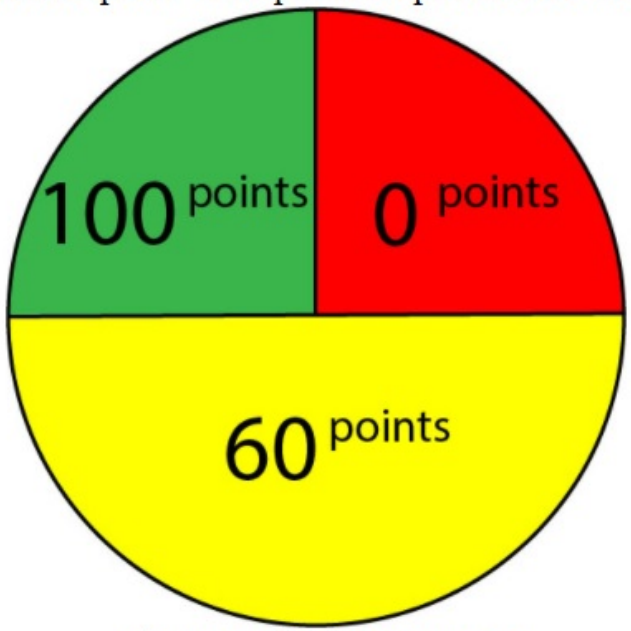
Each question was worth 10 points, for a total of 100.
The *median* is the 'middle number.'
If the median is M, then you have ≥50% chance of getting ≥M points, *and* ≥50% chance of getting ≤M poi

Q1: I think the median score for this person (from 0 to 100) is

[ ] pts

The 11 Questions

The computer will spin this spinner one time:

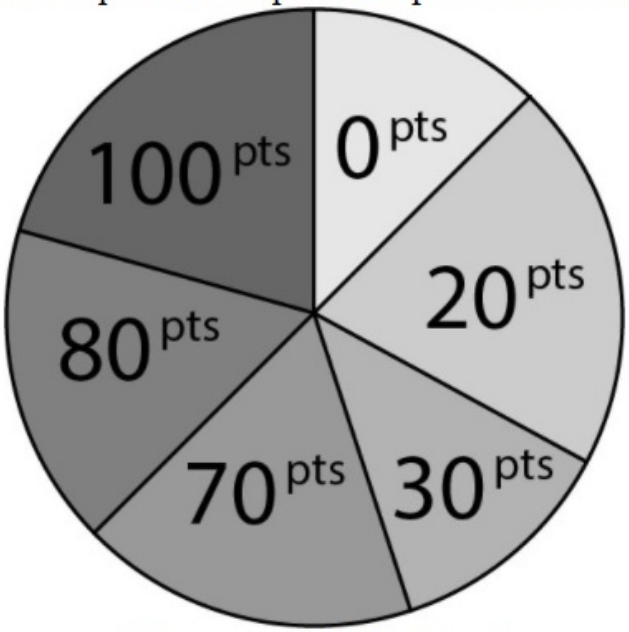


The *mean* is the 'avearge.'
If you multiply each number by its probability and add them up, you get the mean.

Q1: I think the mean # of points for this spinner is _____ pts

The 11 Questions

The computer will spin this spinner one time:



The *mean* is the 'avearge.'
If you multiply each number by its probability and add them up, you get the mean.

Q1: I think the mean # of points for this spinner is _____ pts

The 11 Questions

In 2005 we gave a Carnegie Mellon undergraduate student this quiz:

1. Who is credited with inventing the wristwatch in 1904?
2. Laudanum is a form of what drug?
3. The psychoactive ingredient in marijuana is THC. What does THC stand for?
4. What chemical element has the atomic number five?
5. The study of the structural and functional changes in cells, tissues and organs that underlie disease is called what?
6. What does the suffix -itis mean?
7. The bilby, bandicoot, and quokka are all representatives of what mammalian subclass?
8. Which one of the 50 United States is the only one never to have experienced an earthquake?
9. What evolutionary biologists wrote: *'Creation science' has not entered the curriculum for a reason so simple and so basic that we often mention it: because it is false.*?
10. What is the single most diverse phylum within the animal kingdom?

Each question was worth 10 points, for a total of 100.
The *mean* of their score is the 'avearge.'
If you multiply each possible score by the probability they got that score and add them up, you get the mea

Q1: I think the mean of their score (from 0 to 100) is [ ] pts

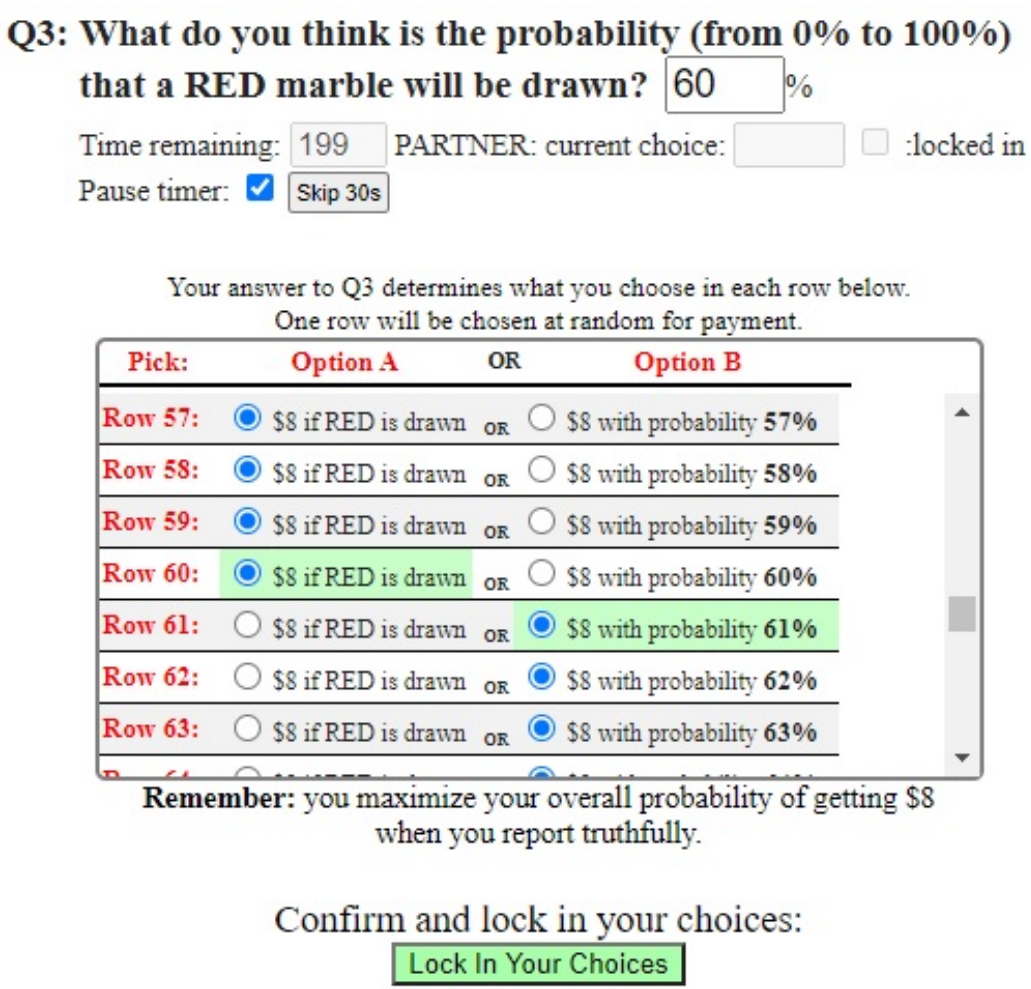
How To Present the Mechanisms

"In the first place, the subject must understand the scoring rule... This is an important reason to present the rule through some vivid tabular or graphic device..."

–Savage (1971)

- **BSR:** Wilson & Vespa (2019), Danz, Wilson & Vesterlund (2022)
- **MPL:** Holt & Smith (2016), Healy (2018)

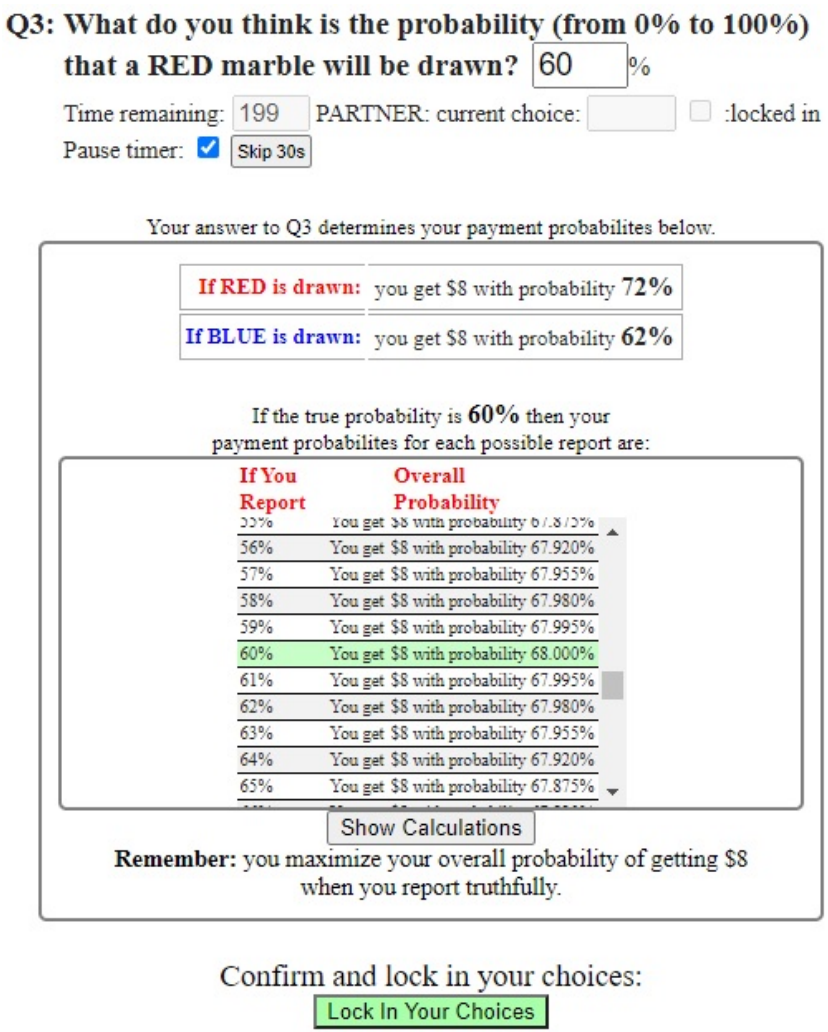
The Mechanism Interfaces: MPL



Link

Note: subjects saw the same phrase in all three treatments

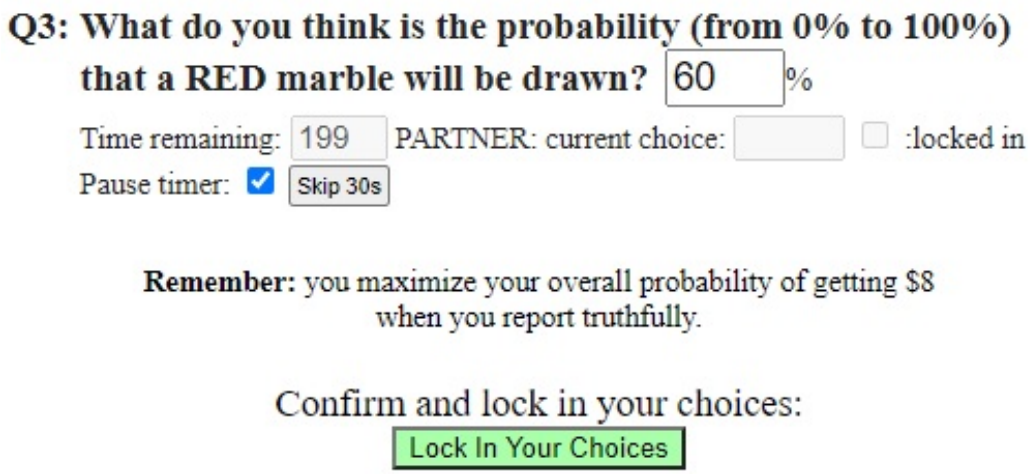
The Mechanism Interfaces: BSR



Link

Note: subjects saw the same phrase in all three treatments

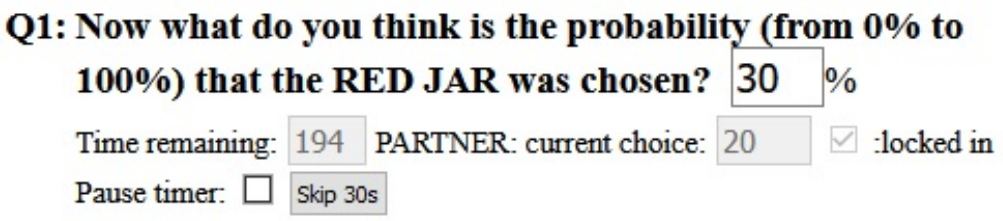
The Mechanism Interfaces: NoInfo

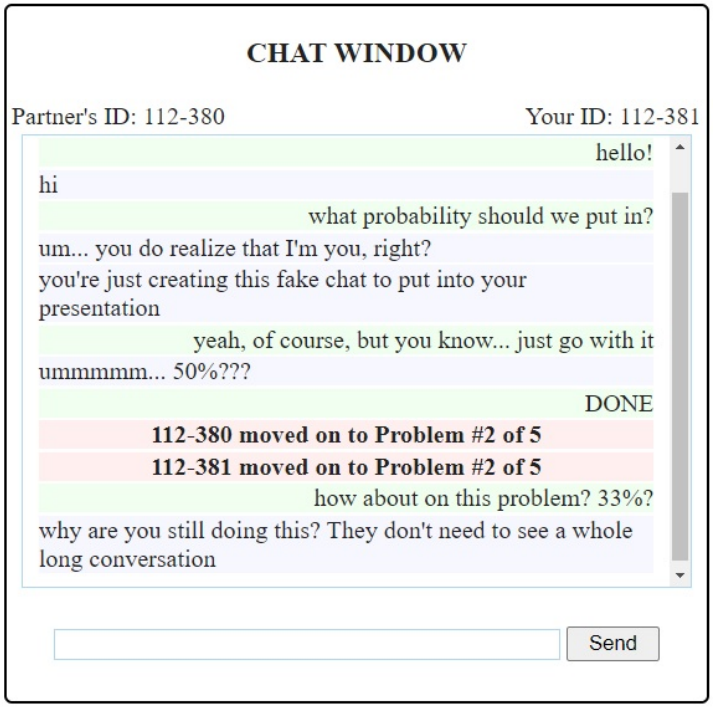


Link

Note: subjects saw the same phrase in all three treatments

Teams Interface





- Use chat window to communicate
- Must lock in the same number to proceed
- Can unlock & change $\Rightarrow$ “Silent agreement”
- If time runs out, one choice is randomly used

Logistics

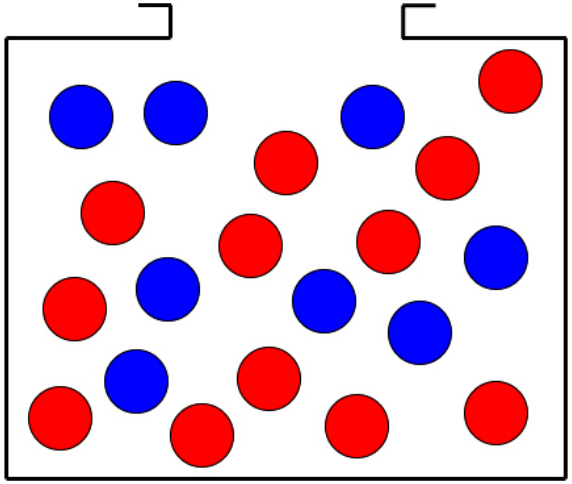
- Usual OSU subject pool (ORSEE)
- Zoom meeting
- Less control of software environment $\Rightarrow$ missing observations
  - INDIV: 1.7% TEAM: 8.3%
- Venmo payments (option for in-person)
- $12 show-up + possible $8 "bonus." (59% won the bonus)

Subjects:

Mechanism:	**MPL**	**BSR**	**NoInfo**
INDIV First:	68	68	63
TEAMS First:	54	54	0
Pooled:	122	122	63

Results

Objective-Easy #1: % Correct



Pr(Red) = 12/20 = 60%

% Correct:

	MPL	BSR	NoInfo
INDIV:	91.7%	96.6%	92.1%
TEAM:	94.8%	100%	96.4%

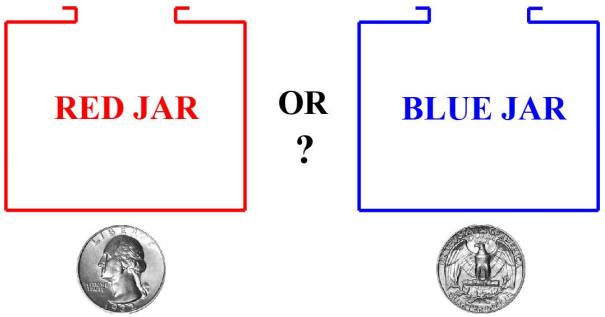
MPL seems a little worse. Are they trying to manipulate?

Objective-Easy #1: Chats

ID#181	MPL	ID#187
i have 12 for red and 8 for blue		
		12, 20, and 75%? yes
75 sounds good with me		
12|20|75%		12|20|75%

ID#289	MPL	ID#295
sorry I put wrong answer for 3		
12|20|50%		12|20|50%

Objective-Easy #2: % Correct



Pr(Red) = 50%

% Correct:

	MPL	BSR	NoInfo
INDIV:	91.5%	84.8%	93.7%
TEAM:	98.3%	93.1%	100%

Now BSR seems a little worse?

Objective-Easy #2: Chats

ID#390	MPL	ID#391
		50%
so theoretically it's 50 right but i think i said 48 last time just bc I'm in stats rn and we just did probability stuff about how smaller sample sizes are further from the probability so flipping it once might be 60-40 but 100 times is closer to 50-50 but ya I'm good w just 50		
		makes sense should we do 49%
sure		
49%		**49%**

Objective-Easy #2: Chats

ID#257	BSR	ID#260
		50 ?
id say 60		
		Why
cause heads is always more likely		
		Thats just false
55 is a compromise		
Which is also wrong but whatever		
55%		**55%**

ID#357	BSR	ID#365
(no chat)		
75%		**75%**

Error Rates By Question

For the 7 questions with an objectively-correct answer:

Question	BSR	MPL	NoInfo
60% Jar	**0.0%**	5.2%	3.6%
Coin Flip	6.9%	1.7%	**0.0%**
Median: 3 Slice	13.8%	25.4%	**7.4%**
Mean: 3 Slice	45.7%	28.2%	**17.9%**
Median: 6 Slice	**32.8%**	58.6%	47.2%
Bayes: 1 Signal	59.5%	**54.1%**	73.1%
Bayes: 2 Signal	94.7%	**93.5%**	100%

Chat Encoding

Two Types of Evidence of IC Failures:

Calculate Playing with the calculator

- May not end up deviating from their belief

Deviate Deviate from stated belief

- May not specify *why* they're deviating

Two independent chat encoders

Chat Encoding

Two Types of Evidence of IC Failures:

Calculate Playing with the calculator

- May not end up deviating from their belief

Deviate Deviate from stated belief

- May not specify *why* they're deviating

Team-level data:

Mechanism:	MPL	BSR	NoInfo
Calculate	3	16	0
Deviate	3	8	0
Both	1	6	0
Neither	56	43	31

Chat Encoding

Two Types of Evidence of IC Failures:

Calculate Playing with the calculator

- May not end up deviating from their belief

Deviate Deviate from stated belief

- May not specify *why* they're deviating

Question-level data:

Mechanism:	**MPL**			**BSR**			**NoInfo**
Question:	Obj-E	Obj-H	Subj	Obj-E	Obj-H	Subj	All
Calculate	1	1	1	3	14	12	0
Deviate	2	1	0	1	4	3	0
Both	1	0	0	0	4	2	0
Neither	242	242	182	240	230	170	341

Subjects use the BSR calculator when clueless!

Example: Calculate & Deviate w/ BSR

Capital of Australia

ID#591	BSR	ID#599
i said 90 bc Carnegie is a prestigious school and theyre smart kiddos so they hv to know this easy answer what do u think should we go higher than 90		
		I think we should go higher
95/ 100? 95? 100? **		
		seems 100 gets the higher probability
yea with 55.9		
		**highest should we do 100
yes		
100		**100**

The Story So Far

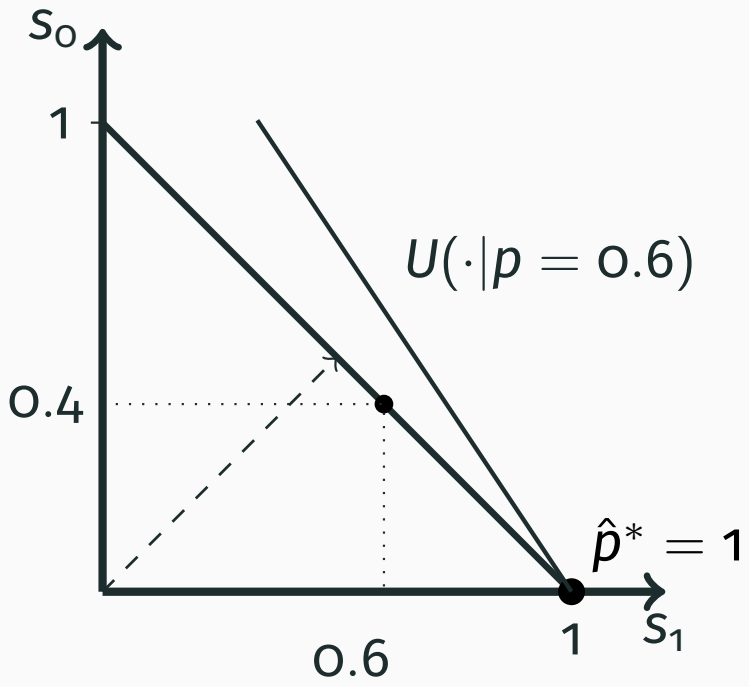
- Misreport rates on easy questions are very low
- BSR and MPL perform equally well
- NoInfo also performs well, but not strictly better (not needed)
- Chats conclude subjects are **not** intentionally manipulating
    - Somewhat more *attempts* in BSR

A Non-IC Mechanism

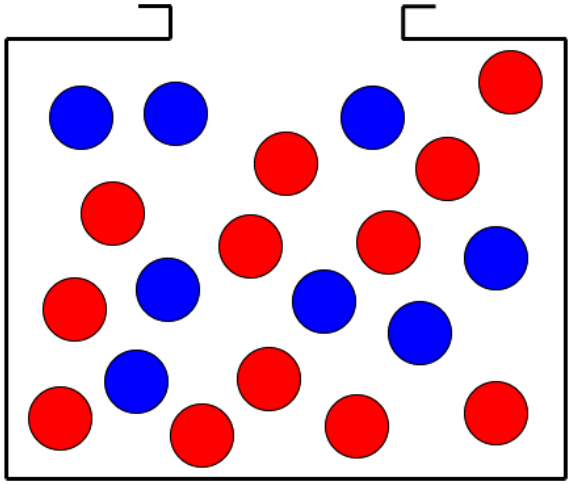
The Linear Scoring Rule

Binarized Linear Scoring Rule (LSR):

- Paid q if true, $1 - q$ if false
- $p < 1/2 \Rightarrow q^* = 0$
- $p > 1/2 \Rightarrow q^* = 1$
- No statement about what's optimal
- Same interface as BSR
- 60 subjects, 30 teams
- Only ran INDIV first



Objective-Easy #1: % Correct



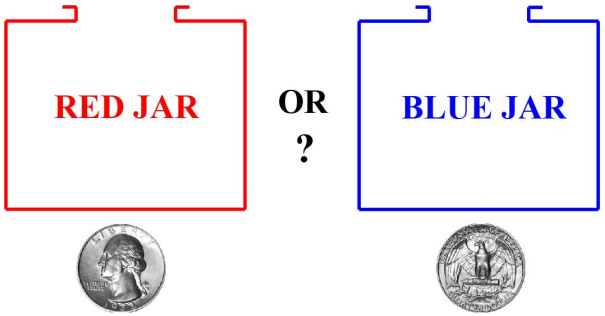
Pr(Red) = 12/20 = 60%

% Correct:

	MPL	BSR	NoInfo	LSR
INDIV:	91.7%	96.6%	92.1%	89.8%
TEAM:	94.8%	100%	96.4%	100%

Six INDIV misreports: (16.67, 40, 50, 62, 100, 100)
TEAMS don't misreport!

Objective-Easy #2: % Correct



Pr(Red) = 50%

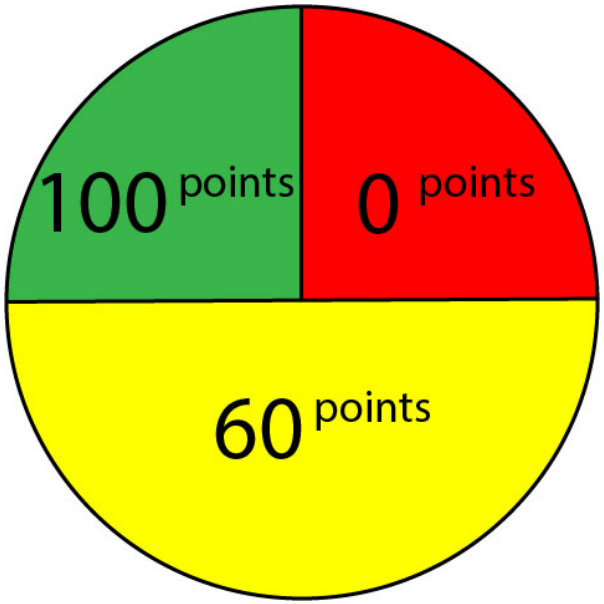
% Correct:

	MPL	BSR	NoInfo	LSR
INDIV:	91.5%	84.8%	93.7%	79.7%
TEAM:	98.3%	93.1%	100%	100%

INDIV misreports: 20,50,52,55,58,60,65,69,70,70,80,100,100

Again, TEAMS discourage misreports

Objective-Easy #3: % Correct



Median = 60pts

% Correct:

	MPL	BSR	NoInfo	LSR
INDIV:	69.2%	83.9%	74.2%	66.7%
TEAM:	74.6%	86.1%	92.6%	86.7%

Again, TEAMS reduce misreports

Objective-Easy: Chats

Subject 653 clearly figured out 100% is optimal, yet on the first question:

ID#651	LSR-Canberra	ID#653
30?		
		sounds good to me
ok		
30%		**30%**

Objective-Easy: Chats

A convincing subject:

ID#678	LSR-MedianEasy	ID#681
i think it will be 60		
	Yeah but look at the probabilities, if we think it is 60 there is a higher chance of money when we pick 100 it it goes from 52% to 60%	
okay, sounds good		
100		**100**

Chat Encoding

Chat Coding with the LSR:

Calculate Playing with the calculator

- May not end up deviating from their belief

Deviate Deviate from stated belief

- May not specify *why* they're deviating

Team-level data:

Mechanism:	MPL	BSR	NoInfo	LSR
Calculate	3	16	0	3
Deviate	3	8	0	4
Both	1	6	0	3
Neither	56	43	31	26

Chat Encoding

Chat Coding with the LSR:

Calculate Playing with the calculator

- May not end up deviating from their belief

Deviate Deviate from stated belief

- May not specify *why* they're deviating

Question-level data:

Mechanism:	**LSR**		
Question:	Obj-E	Obj-H	Subj
Calculate	2	2	1
Deviate	4	2	0
Both	2	2	0
Neither	116	118	89

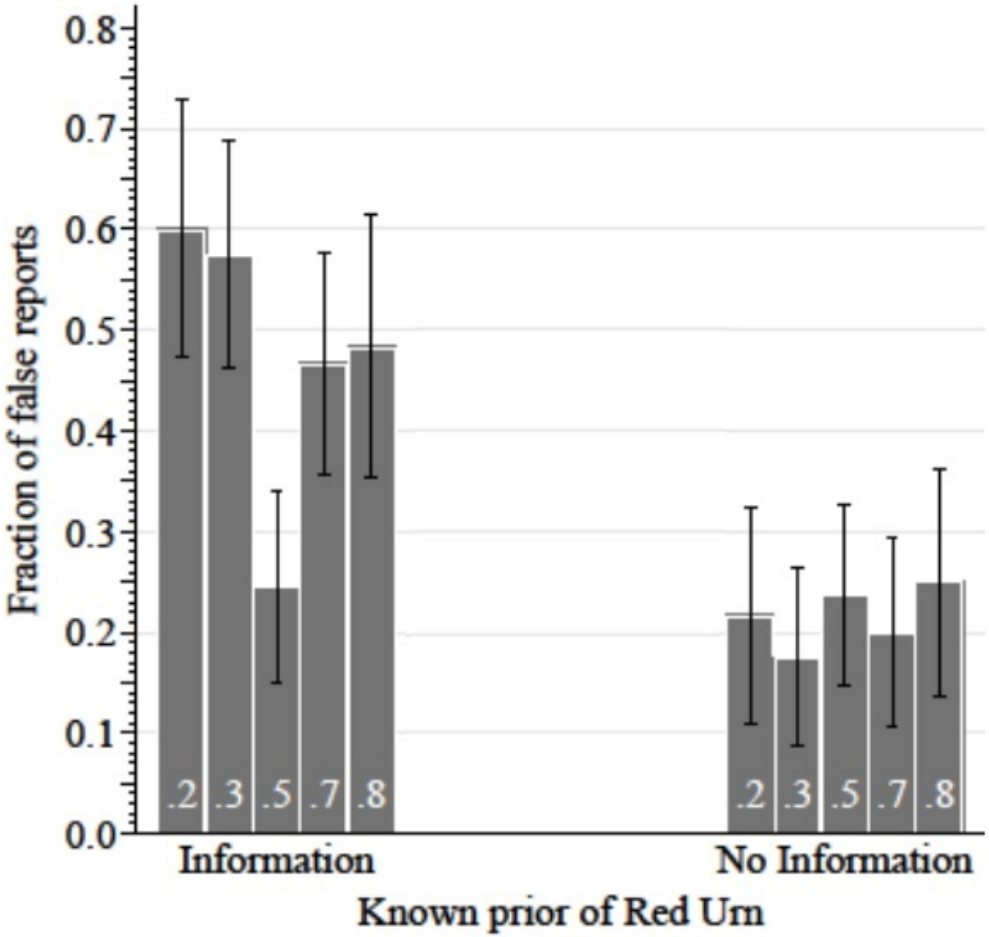
Summary:

- People don’t seem so sensitive to incentives!
- Apparent aversion to lying (helps IC)
    - Teams typically *reduce* (optimal) misreporting!

The Pittsburgh Paper

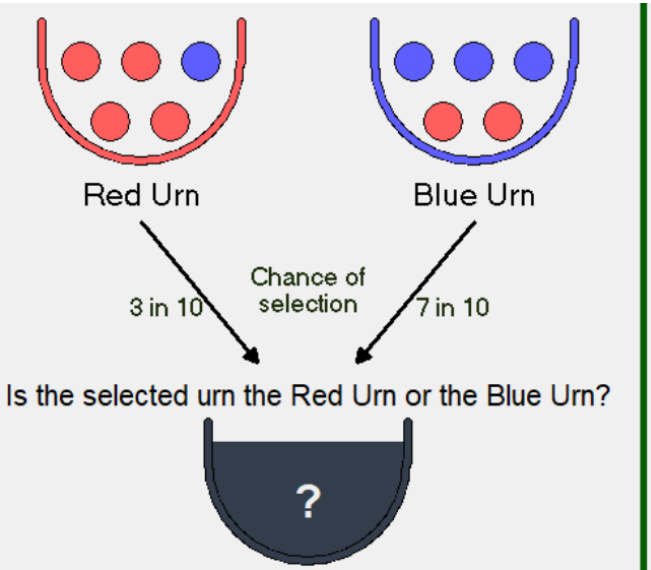
Danz, Vesterlund, & Wilson (*AER* 2022)

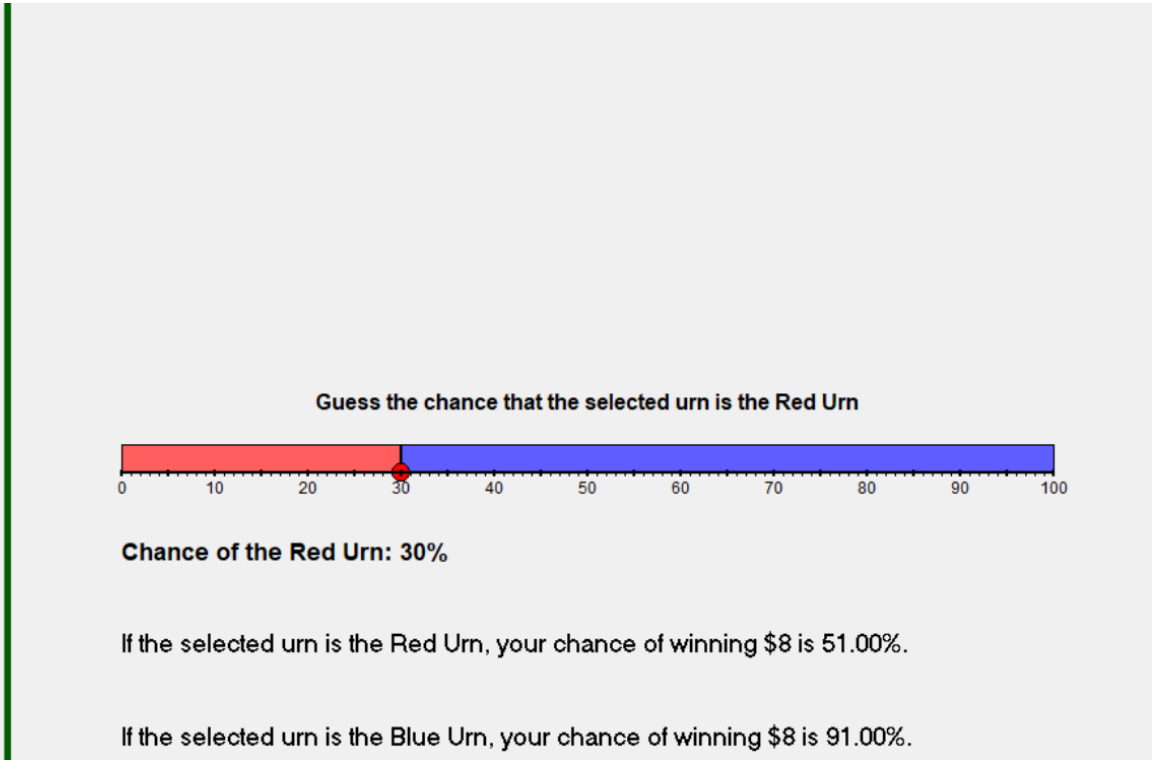
Easy Task misreport %'s:



- We had $<$ 10% at 0.5 and 0.6
- Why do they see misreporting & pull-to-center???

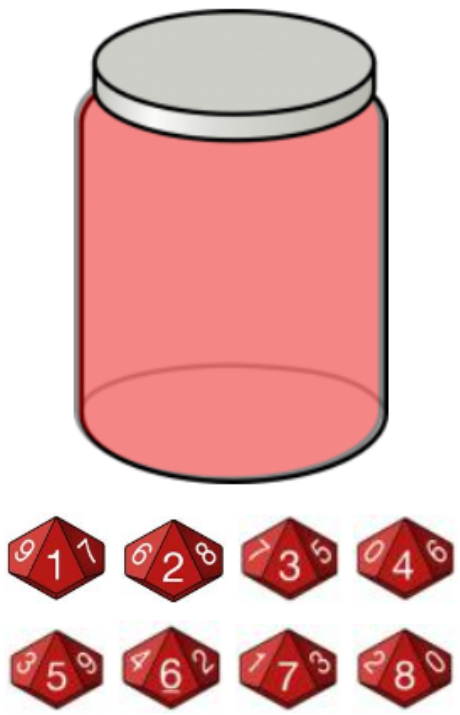
Danz Et Al. Choice Interface



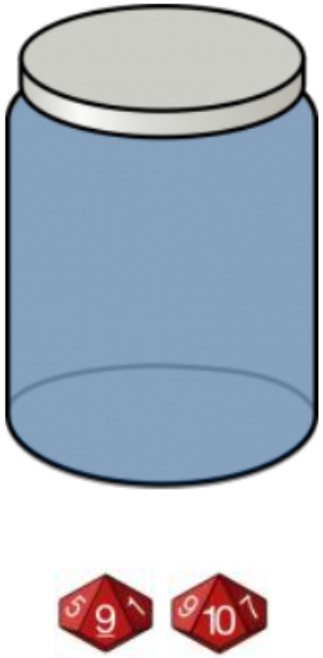


- Clickable slider $\Rightarrow$ inexact answers $\Rightarrow$ pull to center??
- True probability too small on the screen??
  - Changes on every screen
  - More susceptible to distraction by payment info?

Our Choice Interface: NoInfo



8 in 10



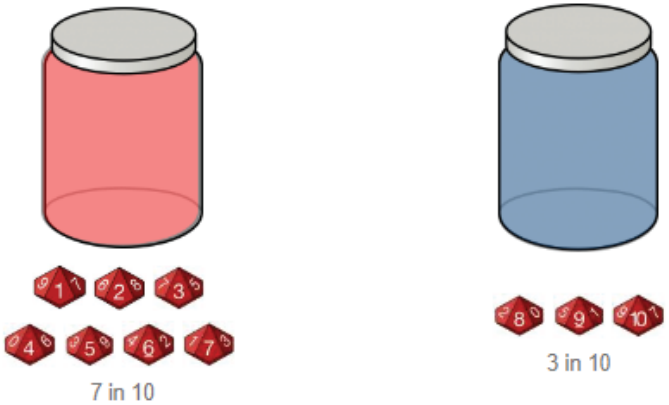
2 in 10

What do you think is the probability (from 0% to 100%) that the RED JAR was chosen?

[ ] %

Our Choice Interface: BQSR

The computer will roll a 10-sided die to choose one of these two jars. The Red Jar is chosen if the die comes up 1 through 7.



If I think the probability of the Red Jar is [ ] %
then my chances of getting $3 would be:

If Red Jar: you get $3 with probability __%

If Blue Jar: you get $3 with probability __%

If the true probability is ___% then your payment probabilites for each possible report are:

If You Report	Overall Probability
0%	You get $3 with probability __%
1%	You get $3 with probability __%
2%	You get $3 with probability __%
3%	You get $3 with probability __%
4%	You get $3 with probability __%
5%	You get $3 with probability __%
6%	You get $3 with probability __%
7%	You get $3 with probability __%
8%	You get $3 with probability __%
9%	You get $3 with probability __%

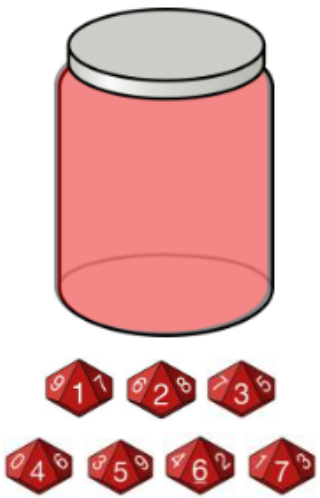
Show Calculations

Remember: you maximize your overall probability of getting $3 when you report truthfully.

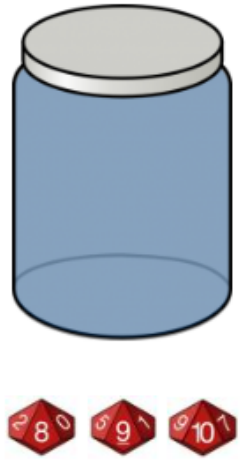
What do you think is the probability (from 0% to 100%) that the RED JAR was chosen?

[ ] %

Our Choice Interface: MPL



7 in 10



3 in 10

If I think the probability of the Red Jar is [ ] %
then my choices would be:

Pick:	Option A	OR	Option B
Row 0:	○ $3 if the Red Jar is chosen	OR	○ $3 with probability 0%
Row 1:	○ $3 if the Red Jar is chosen	OR	○ $3 with probability 1%
Row 2:	○ $3 if the Red Jar is chosen	OR	○ $3 with probability 2%
Row 3:	○ $3 if the Red Jar is chosen	OR	○ $3 with probability 3%
Row 4:	○ $3 if the Red Jar is chosen	OR	○ $3 with probability 4%
Row 5:	○ $3 if the Red Jar is chosen	OR	○ $3 with probability 5%
Row 6:	○ $3 if the Red Jar is chosen	OR	○ $3 with probability 6%
Row 7:	○ $3 if the Red Jar is chosen	OR	○ $3 with probability 7%

Remember: you maximize your overall probability of getting $3 when you report truthfully.

What do you think is the probability (from 0% to 100%) that the RED JAR was chosen?

Instructions Only

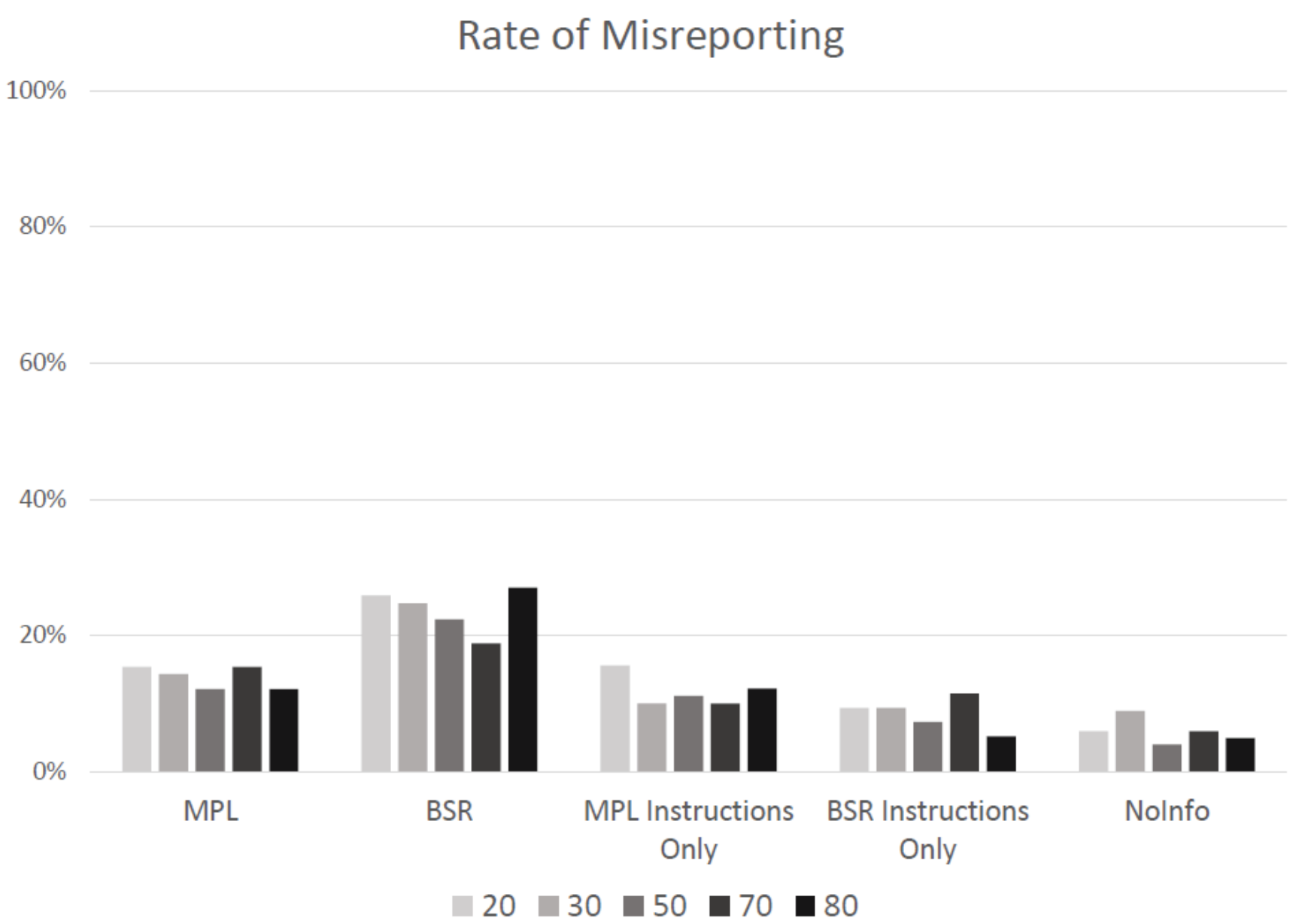
Prolific replication with 5 treatments ($n \approx 100$ ea.):

1. MPL
2. BQSR
3. NoInfo
4. MPL-InstructionsOnly
5. BQSR-InstructionsOnly

Instructions Only: How I would *actually* do elicitation:

- Mechanism details in instructions
- No details on decision screens

Robust Replication Results



Differences?

Please don’t say “their paper doesn’t replicate.”

This is a “robust replication” not an “exact replication”

Differences:

1. Pitt Lab adults vs. Prolific US adults
2. Clickable slider vs. text input
3. Different illustrations of the question
4. We scaled BQSR to make expected payment = MPL
5. Instructions similar, not the same
6. Different calculator interfaces

⋮

Discussion

Summary

- Theory:
  1. MPL has superior IC properties
- Empirics:
  1. MPL and BSR perform similarly
  2. NoInfo does fine, but isn't strictly better
  3. *Very little* evidence of manipulation
     - Even when they *should* manipulate!
     - Subjects are confused/overwhelmed, not manipulating

My recommendations:

1. Use the MPL
2. Explain it in the instructions only
3. Tell them truth max's expected earnings

An Advertisement

An Advertisement

"Ternary Price Lists" with Greg Leo

- MPL is IC under weaker assumptions than BSR
- BSR has *double* the marginal incentives of the MPL
- New mechanism: Ternary Price List (TPL)
  - IC under same assumptions as MPL
  - Has same marginal incentives as BSR

Row#	Option A	OR	Option B	OR	Option C
50	\$8 if $X = 0$	or	\$8 if $X = 1$	or	\$8 w/ prob 50%
51	\$8 if $X = 0$	or	\$8 if $X = 1$	or	\$8 w/ prob 51%
⋮	⋮	⋮	⋮	⋮	⋮
q	\$8 if $X = 0$	or	\$8 if $X = 1$	or	\$8 w/ prob q%
$q + 1$	\$8 if $X = 0$	or	\$8 if $X = 1$	or	\$8 w/ prob $q + 1$%
⋮	⋮	⋮	⋮	⋮	⋮
100	\$8 if $X = 0$	or	\$8 if $X = 1$	or	\$8 w/ prob 100%

Data coming very soon...

That’s It!