TO SCREEN OR NOT TO SCREEN: THE INFERENCE COST OF POLICIES

Gonzalo R. Arrieta
Maxim Bakhtin
Stanford University

December 4, 2023

MOTIVATION

Screening is important and prevalent

- Hiring decisions: Distinguish good and bad candidates
- Insurance: Distinguish high- vs. low-risk customers
- Trade: Distinguish good- vs bad-quality seller

But whether to screen and how much is not always obvious

- Should we assign an easy or a difficult task to a new worker with unknown ability?

MOTIVATION

Policy choices often affect screening ability

- Incentivizing certain actions pools behaviors and reduces how much we can learn from people's behavior
  - Strictness of law enforcement: little apparent crime leads to little information about people's character
  - Assessing recidivism among prison inmates: little misbehavior within prisons leads to little information about inmates' character (and probability of recidivism)

MOTIVATION

Screening trades off **immediate cost vs delayed benefits**

- **Cost**: chance of worse short-run outcomes
    - E.g., failed task, delay, lower profit
- **Benefit**: information for future decisions
    - E.g., worker's ability, client's quality, seller's reliability

People may screen too little if they underweight the benefits from information

RESEARCH QUESTIONS

Screening trades off **immediate cost vs delayed benefits**

This paper experimentally answers:

1. Does this trade-off lead to **suboptimal screening**?

2. What **mechanisms** drive this?

RESEARCH QUESTIONS

Screening trades off **immediate cost vs delayed benefits**

This paper experimentally answers:

1. Does this trade-off lead to **suboptimal screening**?
    - Yes, people screen too little
2. What **mechanisms** drive this?
    - Failure to anticipate inference
    - Failure to plan ahead

HYPOTHESIS

In practice, screening choices may be affected by:

- Time preferences: delayed benefits vs. immediate costs
    - Worker failing a task now vs. knowing their ability for future promotion decisions
- Risk preferences: uncertainty in payoff from different types
    - Uncertainty whether the worker will succeed or fail
- Strategic reasoning: multiple players
    - Need to account for the worker's strategic response and possible learning

HYPOTHESIS

Hypothesis: People screen too little, even controlling for these effects, because they underestimate the informational benefit:

- Assigning a hard task to a new hire can reveal their ability at the cost of potentially failing the task
- Employers minimize the 'direct' cost of failing the task
- Employers do not assign hard tasks to new workers
- They effectively 'pool' different worker types together ⇒ Lose useful information for promotion decisions

OUR CONTRIBUTION

- Learning and bandits Anderson (2012); Hudja and Woods (2021); Banovetz (2020); Kwon (2020); Hoelzemann and Klein (2021); Merlo and Schotter (1999, 2003)
  - We study a more natural setting
  - We focus on identifying mechanisms
- Failures to optimize Esponda and Vespa (2014); Martínez-Marquina et al. (2019); Dal Bó et al. (2018); Eyster (2019)
  - We explore a non-strategic and deterministic setting
- Unintended consequences of policies Bitler and Karoly (2015); Nandi and Laxminarayan (2016)
  - Introduce the limits of inference as a potential unintended consequence of policies

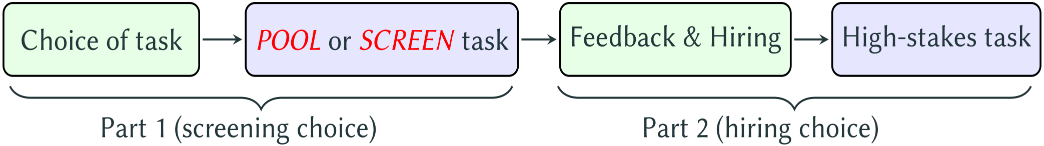
Experiment

EXPERIMENT DESIGN

- Two computers solve tasks
- Tasks can be **Pooling**, *POOL*, or **Screening**, *SCREEN*
- **Good Quality computer:** Produces $0.05 in *POOL* and $0.05 in *SCREEN*
- **Bad Quality computer:** Produces $0.05 in *POOL* and $0 in *SCREEN*
- *SCREEN* reveals quality from output, *POOL* does not

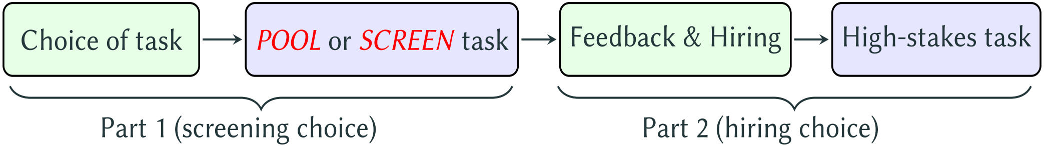
EXPERIMENT DESIGN

- **Good Quality computer:** Produces \$0.05 in *POOL* and \$0.05 in *SCREEN*
- **Bad Quality computer:** Produces \$0.05 in *POOL* and \$0 in *SCREEN*



EXPERIMENT DESIGN

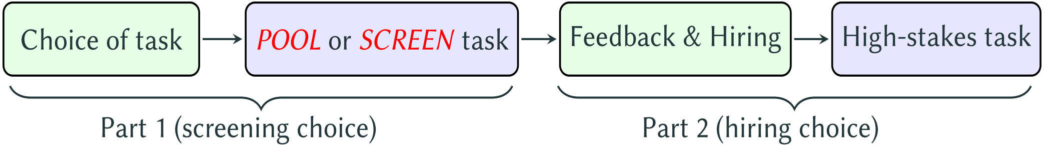
- **Good Quality computer:** Produces \$0.05 in *POOL* and \$0.05 in *SCREEN*
- **Bad Quality computer:** Produces \$0.05 in *POOL* and \$0 in *SCREEN*



- **Part 1:** Participant chooses between *POOL* and *SCREEN* for part 1 → Focus
  - This is the decision to screen or not to screen
  - Part 1 bonus: amount computers produce in the chosen task (\$0.10 or \$0.05)

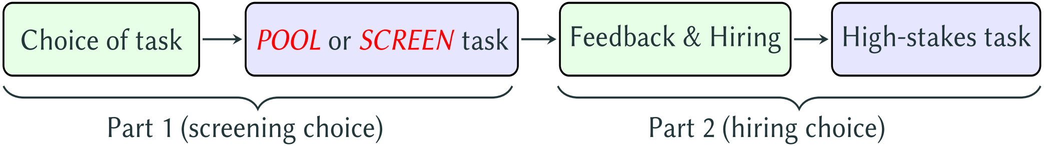
EXPERIMENT DESIGN

- **Good Quality computer:** Produces \$0.05 in *POOL* and \$0.05 in *SCREEN*
- **Bad Quality computer:** Produces \$0.05 in *POOL* and \$0 in *SCREEN*

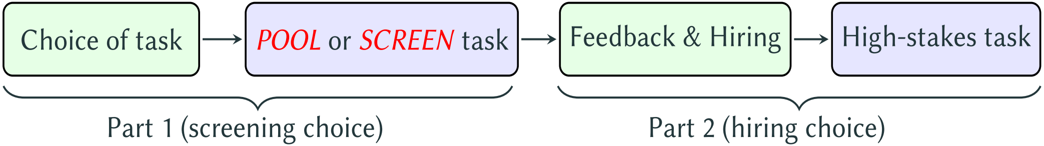


- **Part 1:** Participant chooses between *POOL* and *SCREEN* for part 1
  - This is the decision to screen or not to screen
  - Part 1 bonus: amount computers produce in the chosen task (\$0.10 or \$0.05)
- **Part 2:** Participant sees part 1 bonus and chooses part 2 computer:
  - Part 2 bonus: amount chosen computer produces in high-stakes *SCREEN* task
  - **Good quality:** \$4.3; **Bad quality:** \$0.05

SCREENING TRADE-OFF

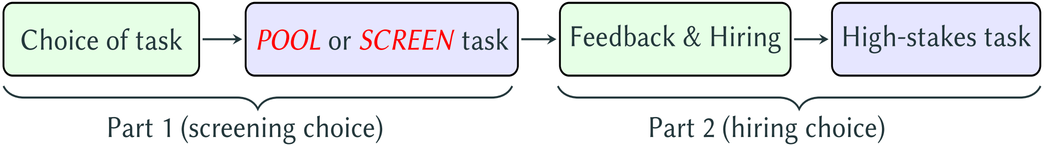


SCREENING TRADE-OFF



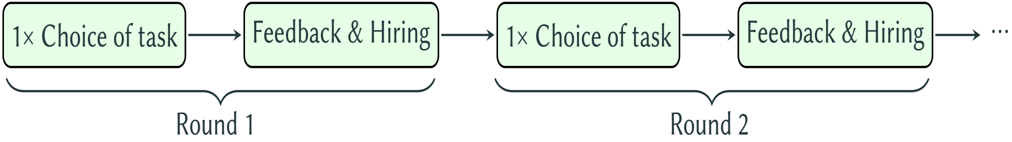
- Screening is costly in part 1: *SCREEN* earns \$0.05 less than *POOL*
- Screening is beneficial for part 2: Good computer earns \$4.25 more than Bad

SCREENING TRADE-OFF



- Screening is costly in part 1: *SCREEN* earns \$0.05 less than *POOL*
- Screening is beneficial for part 2: Good computer earns \$4.25 more than Bad
- Screening is optimal: *SCREEN*: \$4.35; *POOL*: (50% \$4.4, 50% \$0.15)
    - Time preferences are irrelevant because all payoffs are at the end
    - Risk preferences cannot justify *POOL* because require extreme love of risk
        - Rule out by directly eliciting preferences over induced lotteries ▸ lotteries
- No need for strategic reasoning

EXPERIMENT STRUCTURE



- Ten rounds
- Ten different parameterizations in random order ▸ Parameters
  - Rounds 1 and 10 are fixed — *SCREEN* is always optimal
  - Rounds 2–9 — *SCREEN* is optimal in half of rounds, *POOL* — in the other half
- Instead of *SCREEN* and *POOL*, we use colors as labels for the tasks in each round (e.g., Brown and Blue tasks)

MAIN TREATMENTS: BASELINE

Two treatments labeled **Baseline** and Strategy method

1. Main treatment: **Baseline** ($N = 251$)
    - Sequential elicitation: part 1 → part 2 with feedback in between
    - Part 1:

Which task do you want the computers to solve in part 1?

○ Yellow task.

○ Green task.

MAIN TREATMENTS: BASELINE

Two treatments labeled **Baseline** and Strategy method

1. Main treatment: **Baseline** (N = 251)
    - Sequential elicitation: part 1 → part 2 with feedback in between
    - Part 1:

      Which task do you want the computers to solve in part 1?

      ◯ Yellow task.

      ◯ Green task.

    - Part 2:

      How do you want your bonus for part 2 to be determined (according to the reminder above)?

      ◯ This computer produced $0.05 in part 1. I want to get what it produces in part 2 as my bonus.

      ◯ This computer produced $0.00 in part 1. I want to get what it produces in part 2 as my bonus.

MAIN TREATMENTS: STRATEGY METHOD

Two main treatments: Baseline and **Strategy method**

2. Control: **Strategy Method** (N = 244)
   - Control for noise
   - Reverse order of elicitation: part 2 → part 1

MAIN TREATMENTS: STRATEGY METHOD

Two main treatments: Baseline and **Strategy method**

2. Control: **Strategy Method** ($N = 244$)
    - Control for noise
    - Reverse order of elicitation: part 2 → part 1
    - **In part 2, make inference for participants**

If the computers solve the Yellow task in part 1 (and hence you will know their quality):

- ○ This computer is Bad. I want to get what it produces in part 2 as my bonus.
- ○ This computer is Good. I want to get what it produces in part 2 as my bonus.

If the computers solve the Green task in part 1 (and hence you will not know their quality):

- ○ This computer is of unknown quality. I want to get what it produces in part 2 as my bonus.
- ○ This computer is of unknown quality. I want to get what it produces in part 2 as my bonus.

MAIN TREATMENTS: STRATEGY METHOD

Two main treatments: Baseline and **Strategy method**

2. Control: **Strategy Method** (N = 244)
    - Control for noise
    - Reverse order of elicitation: part 2 → part 1
    - In part 2, make inference for participants
    - **In part 1, display payoff consequences**

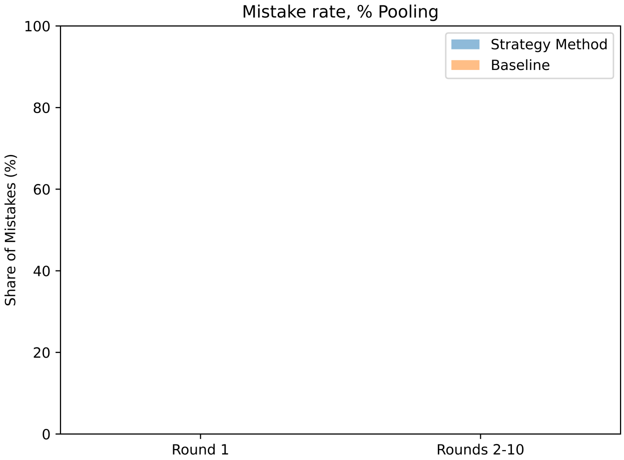
Which task do you want the computers to solve in part 1?

- ○ Yellow task. You get a $0.05 bonus in part 1, and a $4.30 bonus in part 2.
- ○ Green task. You get a $0.10 bonus in part 1. If the unknown quality computer is Good, you get a $4.30 bonus in part 2. If the unknown quality computer is Bad, you get a $0.05 bonus in part 2.

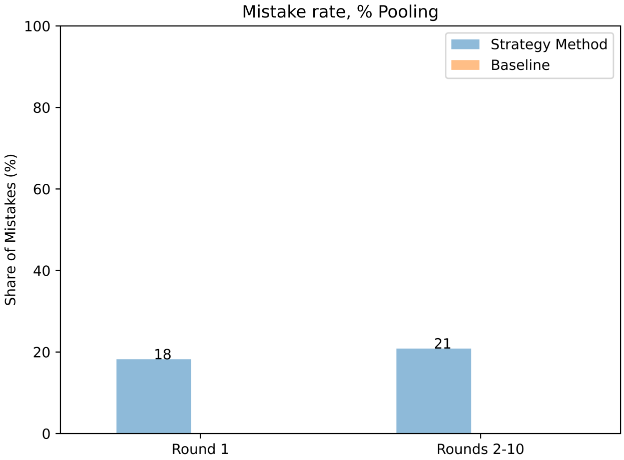
LOGISTICS

- Ran on Prolific in October 2023
- All treatments: $N = 781$
    - Gender: 50% female
    - Average age: 43 years
    - Race: 75% white
    - Education: 16% up to high school, 85% up to college
- Median time: 21 min
- Average payoff: $6.79

RESULTS FOR ROUNDS WHEN SCREENING IS OPTIMAL

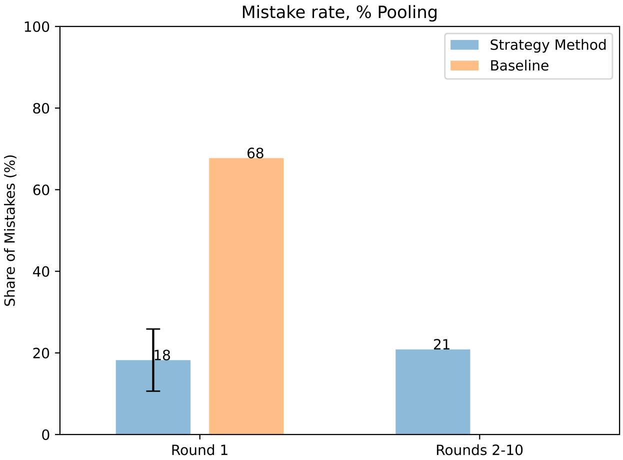


RESULTS FOR ROUNDS WHEN SCREENING IS OPTIMAL



- Few mistakes in Strategy Method, constant over time

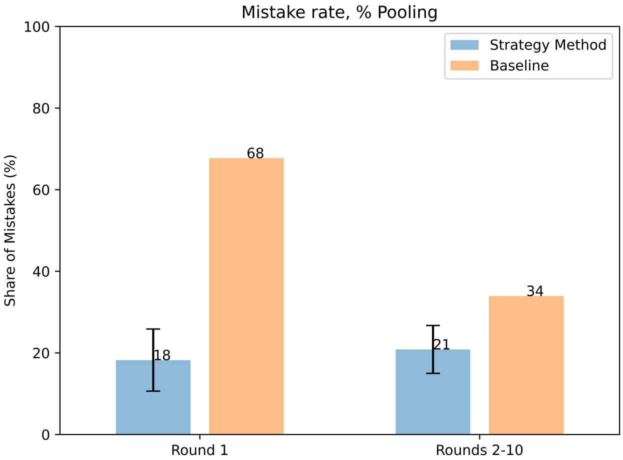
RESULTS FOR ROUNDS WHEN SCREENING IS OPTIMAL



Error bars: 95% CI for difference from Baseline

- Most participants make mistakes in Baseline in round 1

RESULTS FOR ROUNDS WHEN SCREENING IS OPTIMAL



Error bars: 95% CI for difference from Baseline

- Limited learning even after 10 rounds

WHAT CAN CAUSE THE MISTAKE?

This cannot be:

- Time or risk preferences
- Failures of strategic reasoning

WHAT CAN CAUSE THE MISTAKE?

This cannot be:

- Time or risk preferences
- Failures of strategic reasoning

What can cause the mistake?

1. Participants **fail to anticipate inference** from their observations
2. Participants **fail to plan ahead**

We test these with two extra treatments

FAILURE TO ANTICIPATE INFERENCE: DESIGN

Do people **fail to anticipate that they will be able to infer the computers' quality** from their observations?

FAILURE TO ANTICIPATE INFERENCE: DESIGN

Do people **fail to anticipate that they will be able to infer the computers' quality** from their observations?

Automatic Inference treatment (N = 244):

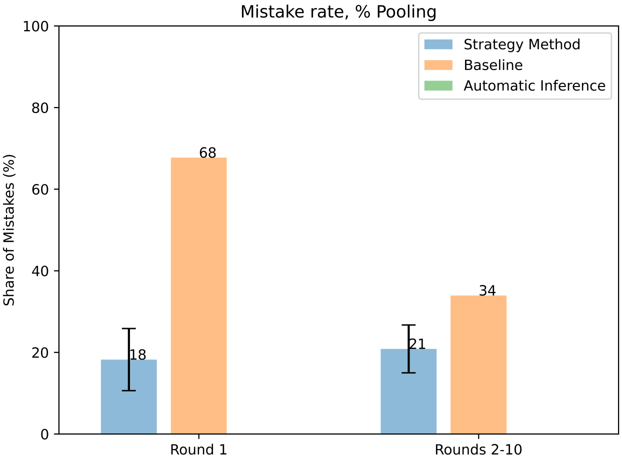
- Highlight the information produced by each part 1 choice

Which task do you want the computers to solve in part 1?

○ Yellow task. We will tell you the computers' quality

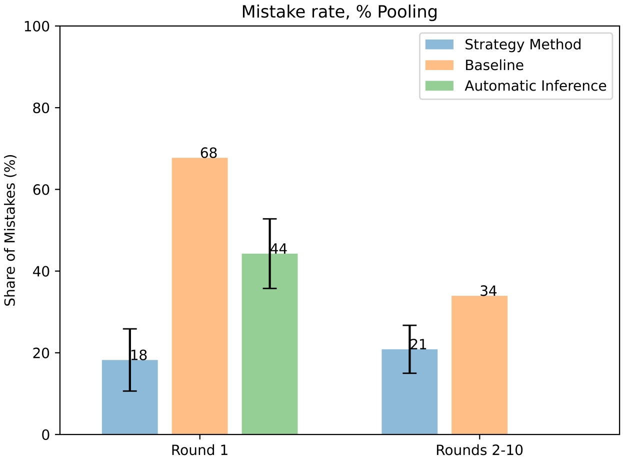
○ Green task. We will not tell you the computers' quality

FAILURE TO ANTICIPATE INFERENCE: RESULTS



Error bars: 95% CI for difference from Baseline

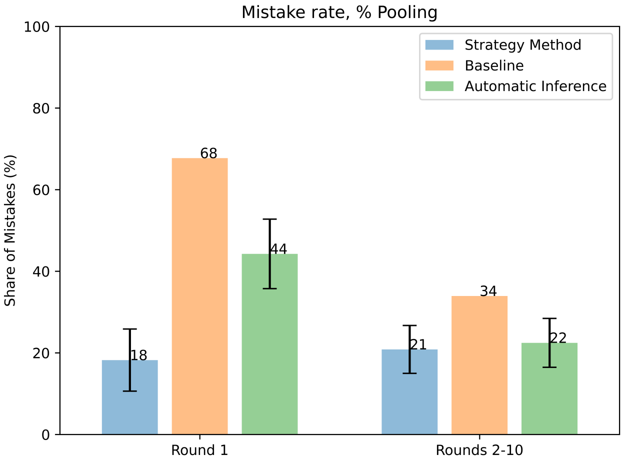
FAILURE TO ANTICIPATE INFERENCE: RESULTS



Error bars: 95% CI for difference from Baseline

- Failure to anticipate inference accounts for half of the mistakes

FAILURE TO ANTICIPATE INFERENCE: RESULTS



Error bars: 95% CI for difference from Baseline

- Full learning

FAILURE TO PLAN AHEAD: DESIGN

Do people **fail to plan ahead**?

FAILURE TO PLAN AHEAD: DESIGN

Do people **fail to plan ahead**?

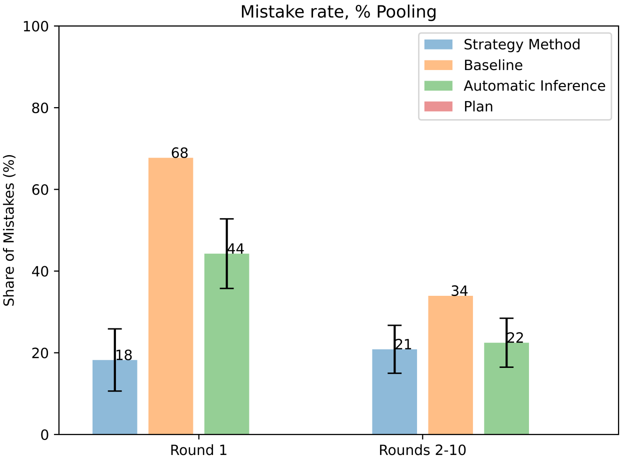
Plan treatment ($N = 50$):

- Combine part 1 and part 2 choices in a single plan

Which task do you want to choose for part 1 and which computer do you want to choose to determine your part 2 bonus?

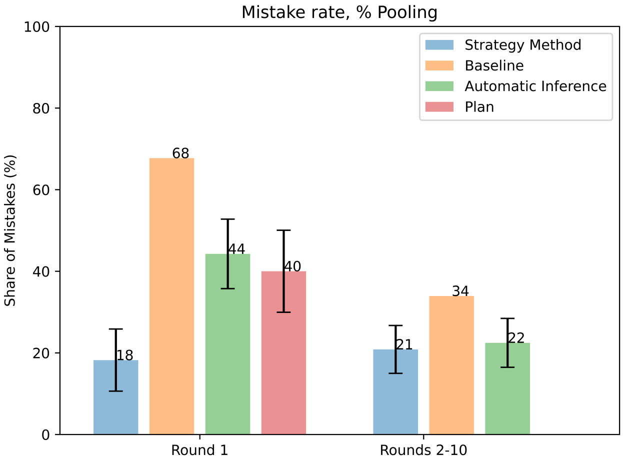
- ◯ In part 1, Green task. In part 2, one of the computers that produce $0.05 in part 1 Green task, chosen randomly.
- ◯ In part 1, Yellow task. In part 2, the computer that produces $0.05 in part 1 Yellow task.
- ◯ In part 1, Yellow task. In part 2, the computer that produces $0.00 in part 1 Yellow task.

FAILURE TO PLAN AHEAD: RESULTS



Error bars: 95% CI for difference from Baseline

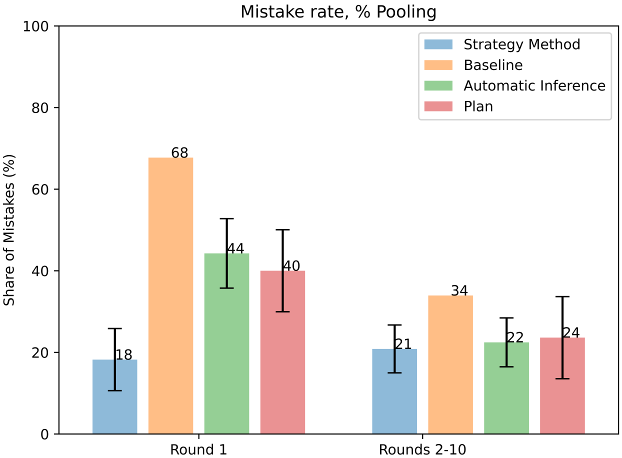
FAILURE TO PLAN AHEAD: RESULTS



Error bars: 95% CI for difference from Baseline

- Failure to plan ahead also accounts for half of the mistakes
- Mechanisms are complementary: Plan + A. I. ≈ Strategy Method

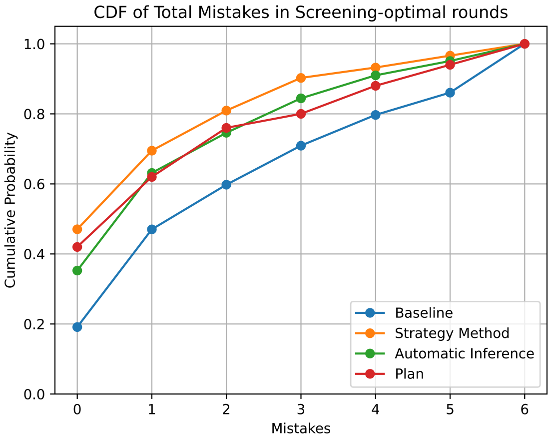
FAILURE TO PLAN AHEAD: RESULTS



Error bars: 95% CI for difference from Baseline

- Also full learning

DISTRIBUTION OF MISTAKES



- FOSD in mistakes: Baseline > A.I. and Plan > Strategy Method
- Big group of never-learners in Baseline

RESULTS: DETAILS AND ROBUSTNESS

- Part 2 choices: >90% optimal choices ▸ Part 2
  - Participants know what to do with the information once they get it
- Most learning happens in the first round ▸ By round
- 90% make at most one error on questions about instructions ▸ CQ
- Results are robust to:
  - Conditioning on zero mistakes in these questions ▸ Zero CQ mistakes
  - Conditioning on knowing that they will observe computers' output before part 2 (in Baseline) ▸ Observable output
    - Question after part 1 before part 2: 'Will you learn how much money each computer produced in part 1?'
  - Conditioning on optimal part 2 choices after screening ▸ Optimal hiring
  - Controlling for demographics and education ▸ With controls

CONCLUSION

In our setting, people screen too little, even with feedback

- Mistake is prevalent and persistent even without uncertainty, time preferences, and strategic interactions
- Two **mechanisms** contribute to the mistake:
  - Failure to anticipate inference and failure to plan ahead
- Practical **lessons**:
  - Important to consider screening when choosing policies or assignments
  - Planning ahead and highlighting inference are complementary interventions
  - Planning can be helpful even without time inconsistency
    - It forces people to consider the full strategy

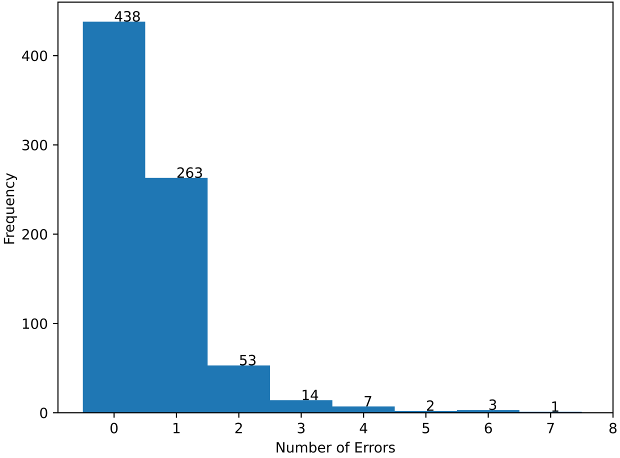
CONCLUSION

In our setting, people screen too little, even with feedback

- Mistake is prevalent and persistent even without uncertainty, time preferences, and strategic interactions
- Two **mechanisms** contribute to the mistake:
  - Failure to anticipate inference and failure to plan ahead
- Practical **lessons**:
  - Important to consider screening when choosing policies or assignments
  - Planning ahead and highlighting inference are complementary interventions
  - Planning can be helpful even without time inconsistency
    - It forces people to consider the full strategy

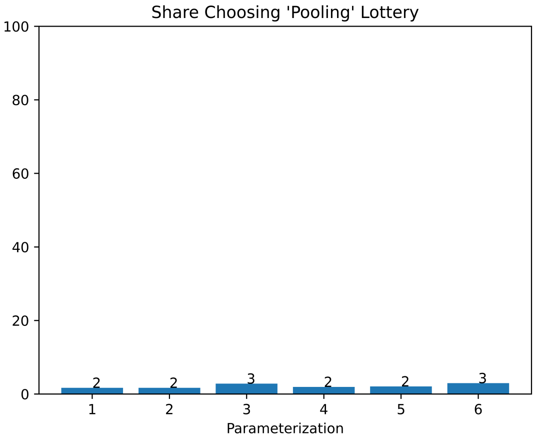
garrieta@stanford.edu & mbakhtin@stanford.edu

COMPREHENSION QUESTIONS



- 90% make at most one error on CQs about experiment design

LOTTERY PREFERENCES

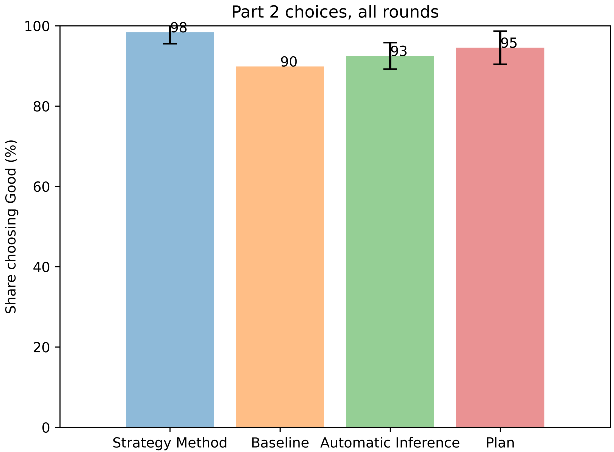


- Almost no one chooses the lottery associated with *POOL*

PARAMETERIZATIONS

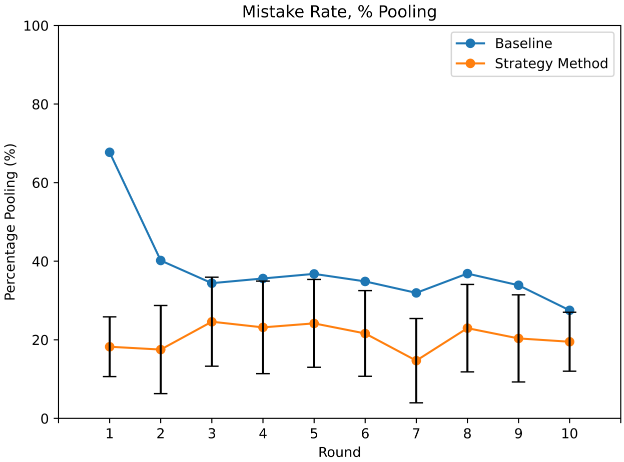
	Part 1			Part 2		
	POOL	*SCREEN*, Good	*SCREEN*, Bad	Good	Bad	Optimum
1	0.05	0.05	0.00	4.30	0.05	*SCREEN*
2	0.05	0.05	0.00	4.45	0.10	*SCREEN*
3	0.2	0.2	0.15	4.30	0.10	*SCREEN*
4	0.05	0.05	0.00	4.45	0.10	*SCREEN*
5	0.05	0.05	0.00	4.35	0.10	*SCREEN*
6	0.05	0.05	0.00	4.50	0.10	*SCREEN*
7	2.20	2.20	0.15	0.20	0.20	*POOL*
8	2.10	2.10	0.00	0.05	0.05	*POOL*
9	2.00	2.00	0.00	0.05	0.05	*POOL*
10	2.15	2.15	0.00	0.05	0.05	*POOL*

PART 2 CHOICES



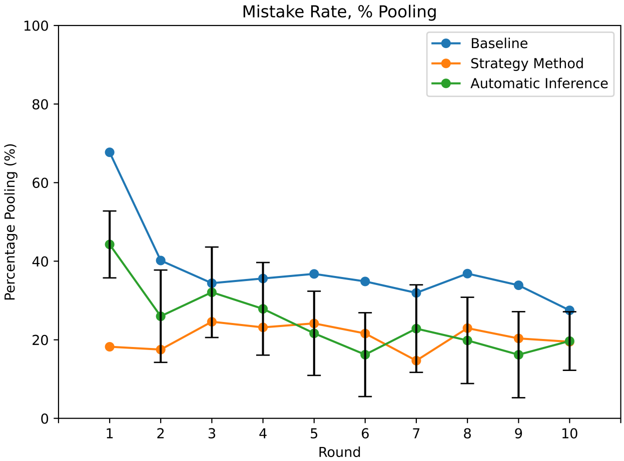
- Almost everyone chooses the Good computer

MISTAKES BY ROUND: BASELINE



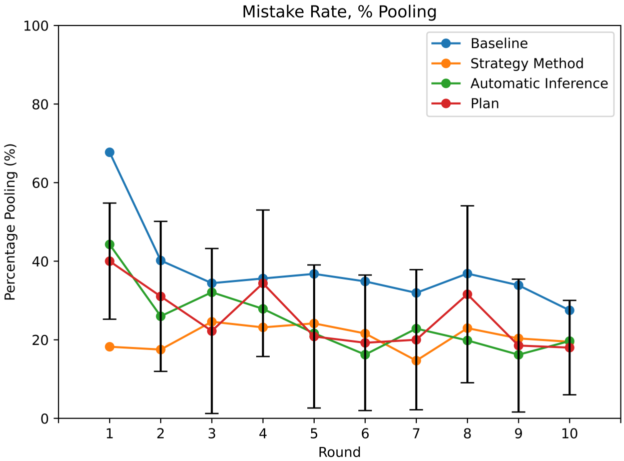
- Most learning happens in round 1

MISTAKES BY ROUND: AUTOMATIC INFERENCE



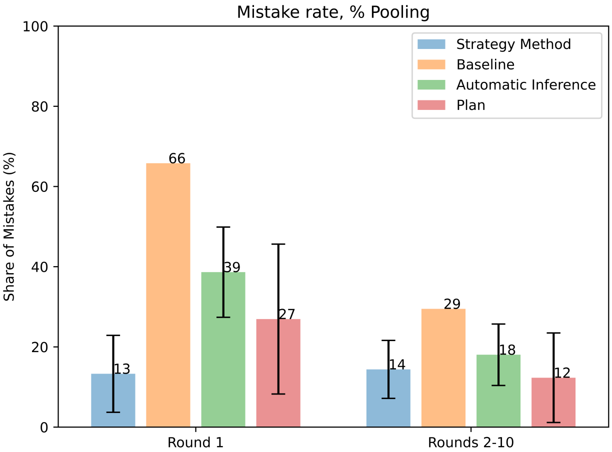
- Highlighting inference immediately and persistently reduces mistakes

MISTAKES BY ROUND: PLAN



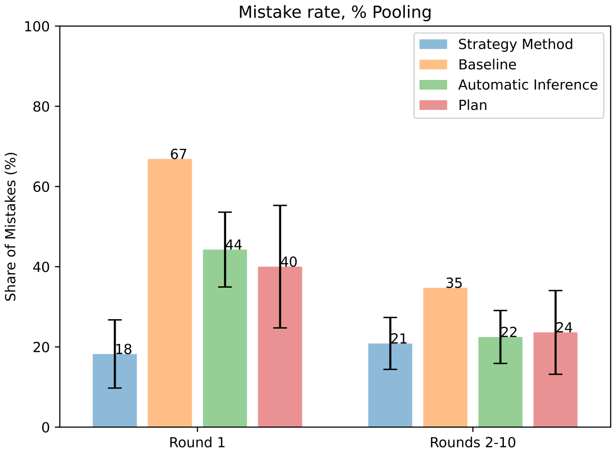
- Planning immediately and persistently reduces mistakes

ZERO CQ MISTAKES



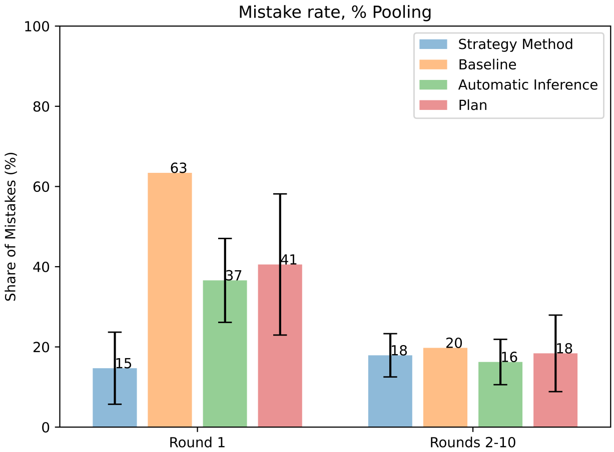
- Conditioning on making zero CQ mistakes

KNOWING COMPUTERS' OUTPUT IS OBSERVABLE



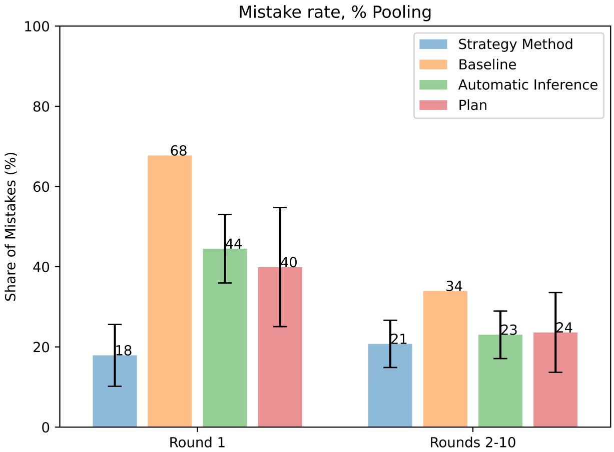
- Conditioning on Baseline participants who know that they will observe how much each computer produced before part 2

OPTIMAL PART 2 CHOICES



- Conditioning on always hiring Good computer after *SCREEN* ; excludes those who always choose *POOL*

CONTROLLING FOR CHARACTERISTICS



- Controlling for age, gender, race, and education

RESULTS FOR ROUNDS WHEN POOLING IS OPTIMAL