Foundations of Experimental Economics

University of Kassel

Winter Term 2025/2026

Lecture 6: Online Experiments, RCTs, Survey Experiments and AI

Introduction

Goal for Today

- **Online vs. Lab Experiments: Examine** the advantages and disadvantages of conducting experiments online and discuss comparative studies of data collected in laboratory versus online settings.

- **Randomized Controlled Trials (RCTs): Briefly overview** RCTs as a specialized form of framed field experiments.

- **Survey Experiments**: **Explore** how survey experiments serve as tools for collecting data and estimating causal effects.

- **AI and Experimental Economics: Discuss** emerging applications at the intersection of artificial intelligence and experimental economics.

Recommended Readings

Online experiments

- Arechar, A.A., S. Gächter and L. Molleman (2018): Conducting interactive experiments online. *Experimental Economics*, 21, 99-131

RCTs

- Banerjee, A. V., & E. Duflo (2009). The experimental approach to development economics. *Annual Review of Economics*, 1, 151–178.

Survey experiments

- Stantcheva, S. (2022a): How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible. *NBER Working Paper 30527*. (Section 6)

AI

- Charness, G., Jabarian, B. and List, J.A. (2023): Generation next: Experimentation with AI. NBER Working Paper Series, Working Paper 31679, https://www.nber.org/papers/w31679.

Online Experiments

Online Experiments

- An online experiment is an experiment conducted over the Internet.
  - Participants complete tasks or make decisions remotely via digital platforms.

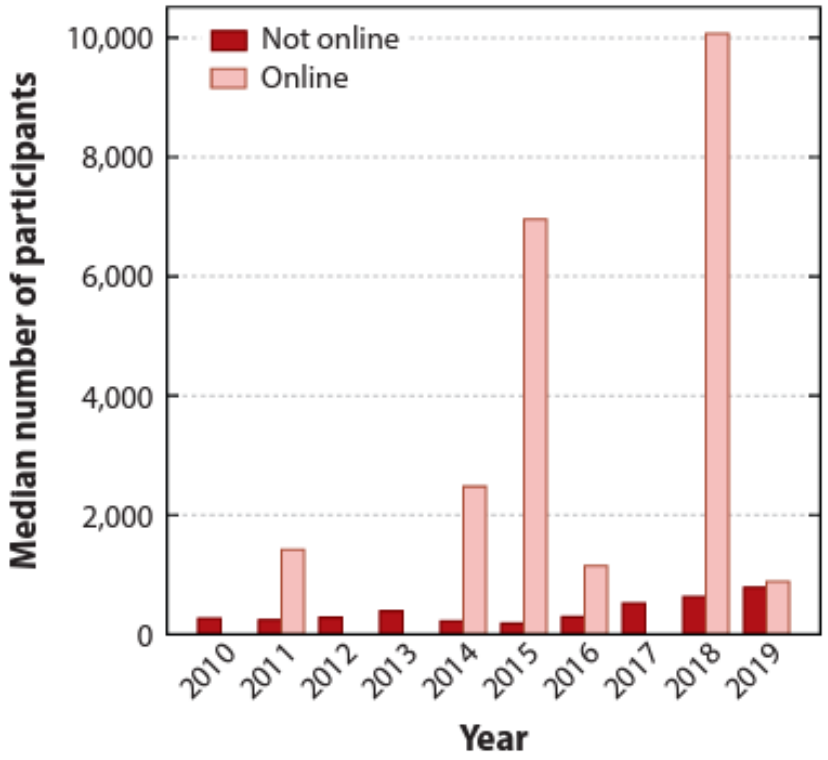
- Online experiments do not constitute a specific kind of experiment; all kinds of economic experiments (lab or field) can be conducted online.
  - In the following, we will focus on how to conduct lab/artefactual field experiments online.

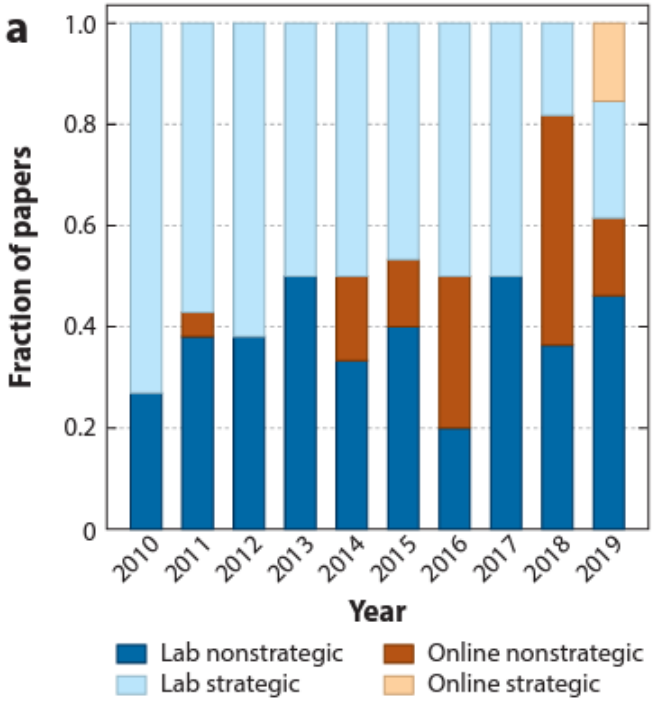
Trends in Experimental Economics (1/2)

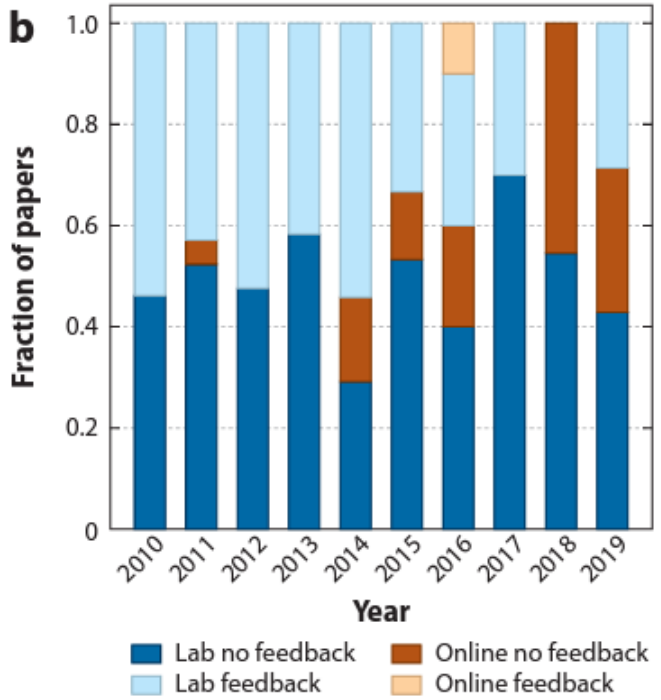
- Fréchette et al. (2022) look at trends in 164 experimental papers published in the top-five journals between 2010 and 2019:
  - Between 2010-2014 only 5% of experiments included an online treatment.
  - Between 2015-2019 that number increased to 23%.
  - This share is likely to be higher for the period since 2020.

- They further show that online experiments include much more subjects, but less complex experimental designs compared to experiments in physical labs

Trends in Experimental Economics (2/2)







Source: Fréchette et al. (2022, p. 783)

Crowdsourcing Platforms: MTurk and Prolific

- **MTurk** is an online labor market where offers are posted as Human Intelligence Tasks (HITs). HITs may include tasks like writing text, identifying specific content in an image, answering surveys for market research but also academic studies.
  - Majority of users are from the US and India.
  - Used for a very large number of online experiments, especially in Psychology.

- **Prolific** is a platform specifically for academic, opinion and market research studies:
  - Majority of users are from the UK and the US.

Example: Prolific

A study about productivity

By Jesper Akesson

£0,84 • £5,00/hr | 18 mins | 252 places

In this study, we will ask you some questions, and ask you to complete a brief task. You will be able to earn additional rewards in the survey.

Devices you can use to take this study:

Desktop Mobile Tablet

Table 1 Methodological differences in conducting interactive experiments in the laboratory and on MTurk

Phase/challenge	Laboratory	Online (MTurk)
Recruitment		
Show-up fees	Typically a small part of total payoffs. Guaranteed when participant shows up to the session	Relatively large show-up fees promote recruitment rates, thereby facilitating prompt group formation. Experimenter can approve or reject the task submitted; if rejected no fee is paid
Inviting participants	Invitations sent well in advance, participants commit to a session. Recruitment often from a pre-existing database	Sessions advertised online as HITs and can be completed immediately
Selection into the experiment	At sign-up, participants know very little about the experiment. Details of the task are communicated once participants are in the laboratory	Experiments are typically advertised as HITs with a brief task description. 'Workers' browse available HITs and accept those of their preference
Experienced participants	Invitation conditioned on well-defined criteria of the laboratory's records	HITs targeted at subsets of MTurk workers; experimenter can specify exclusion criteria. Many MTurk workers will have participated in many prior studies
Session start-up		
Duplicate participants	Registration protocols usually prevent duplicate participation	Amazon acts against multiple worker accounts, but they exist
Comprehension	Participants can ask questions; comprehension questions ensure understanding	Experimenter is physically absent and cannot answer questions directly. Compulsory comprehension questions can be added but may make experiment (too) long for some participants

Source: Arechar et al. (2018, p. 101)

Experimental interactions		
Forming groups	Easy to guess how many participants will attend; group settings can be pre-defined	Hard to guess how many participants will attend; groups can be constructed 'on the fly'
Deception	In experimental economics deception is prohibited and laboratories foster reputations for non-deception	Because all requesters use the same subject pool, some participants may have experienced deception because requesters from other disciplines may use it
Communication	Hardly an issue; experimenter can restrict communication between subjects	Participants may in principle collude through external channels though this is difficult in practice
Experimental flow	Closed form software like z-Tree specifies session progress	Scripted browser navigation specifies progress
Attrition ('dropout')	Hardly an issue; participants that start a session usually finish it	Major challenge to internal validity, if dropout rates vary with treatment, selection bias may arise
Payment		
Payments	Cash usually paid upon completion	Automatic transfer through Amazon
Cost per participant	Relatively high but predictable	Relatively low but varies with attrition

Source: Arechar et al. (2018, p. 101)

Advantages of Online Experiments

- **Wider reach:** Access to large and diverse participant pools worldwide.
- **Lower costs:** Reduced expenses compared to physical labs (space, equipment, staffing, incentivization).
- **Speed:** Faster recruitment and data collection.
- **Scalability**: Easily increase sample size without major logistical hurdles.

Limitations of Online Experiments (1/2)

- **Participant attention & engagement:** Risk of distractions and limited control over participants' environments may reduce data quality.
- **Communication between subjects:** Participants may use phones, messaging apps, or online forums to share information, compromising experimental control.
- **Participant attrition:** High dropout rates during the experiment can lead to incomplete datasets and biased results.
- **Deception:** Subjects on online platforms usually have been subject to deception.

Limitations of Online Experiments (2/2)

- **Sample bias:** Self-selection and unequal internet access can limit representativeness of the sample.
- **Verification of identity:** Difficult to confirm participants' true idenities or prevent multiple entries from the same person.
- **Reduced design complexity:** Synchronous decision-making is often challenging online, so experiments tend to focus on simpler, non-strategic interactions.
- **Difficulty handling questions:** Harder to answer questions from participants.

Comparison of Results from Physical Lab and Online Experiments

- In the following, we look at some studies that compare behavior in the physical lab environment with behavior in online experiments.

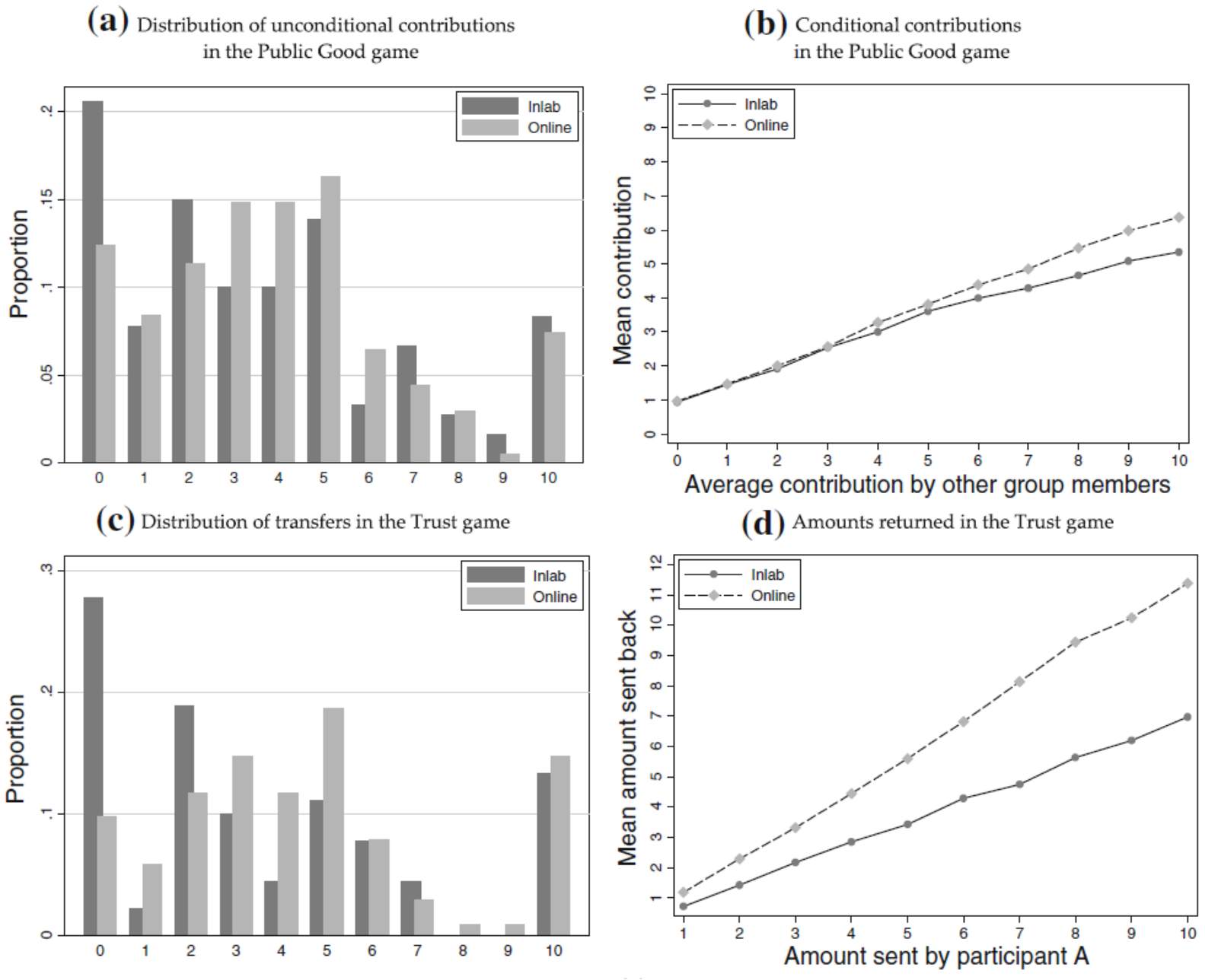
- Some of these studies keep the subject pool constant (same standard student pool) and only change the environment (lab vs. online), others change both the environment and the subject pool (e.g., student pool in the lab vs. MTurk workers online).

Hergueux and Jacquemet (2015) (1/5)

- Hergueux and Jacquemet (2015) compare behavior in the physical lab and an online environment for the following games or tasks:
  - Public Good game, Dictator game, Ultimatum game, Trust game, Risk aversion elicitation, and a Social value survey.

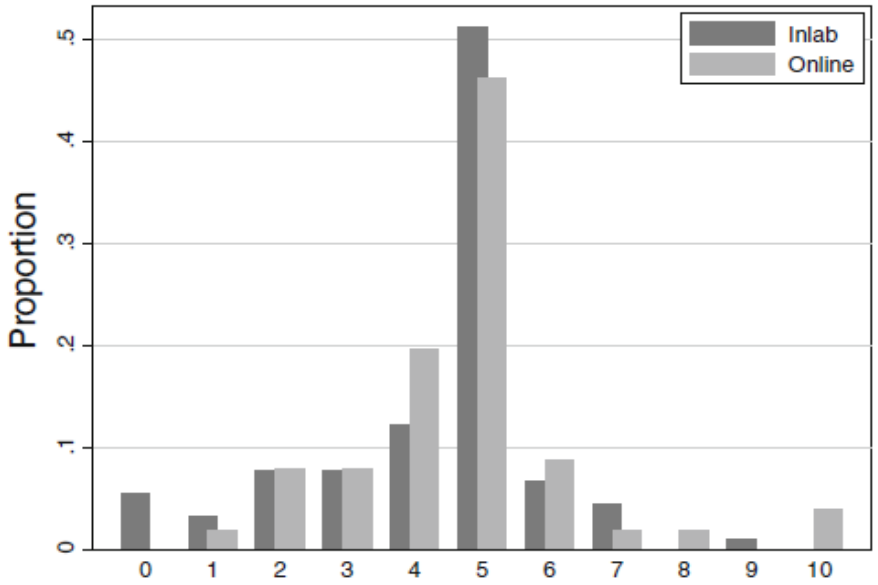
- Subjects were recruited from the pool of the experimental lab of the University Paris 1; after signing up for a session they are randomly allocated to the physical lab or the online treatment.

- Subjects in the lab get paid in cash, online subjects via PayPal.

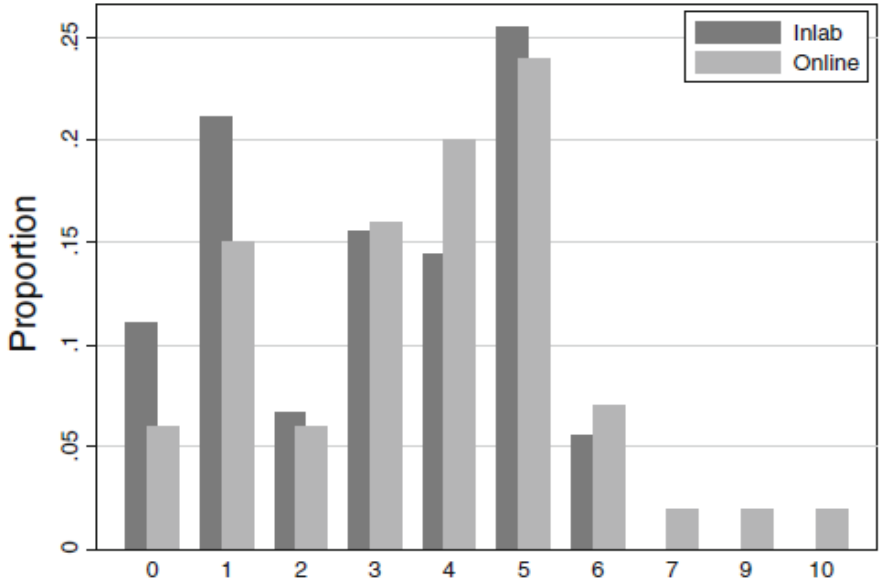


Source: Hergueux and Jacquemet (2015)

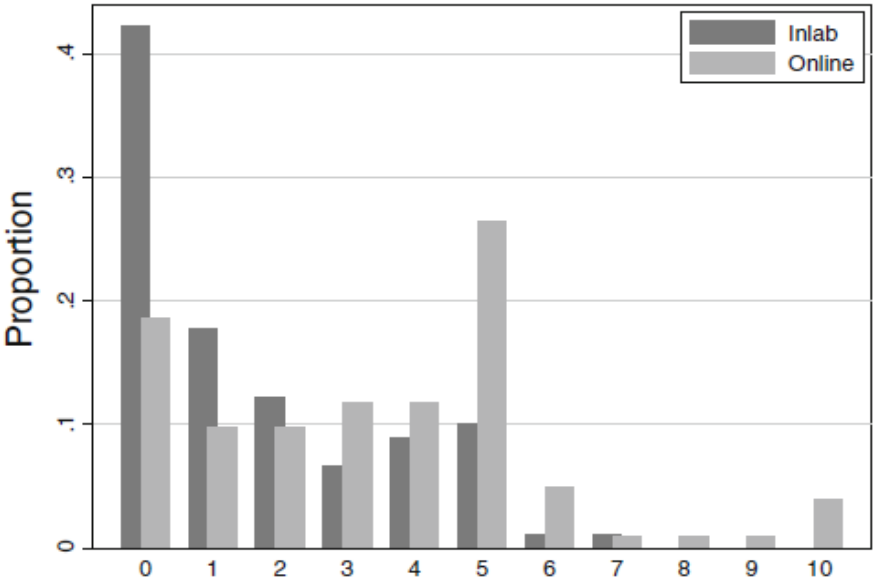
(e) Distribution of transfers in the Ultimatum game



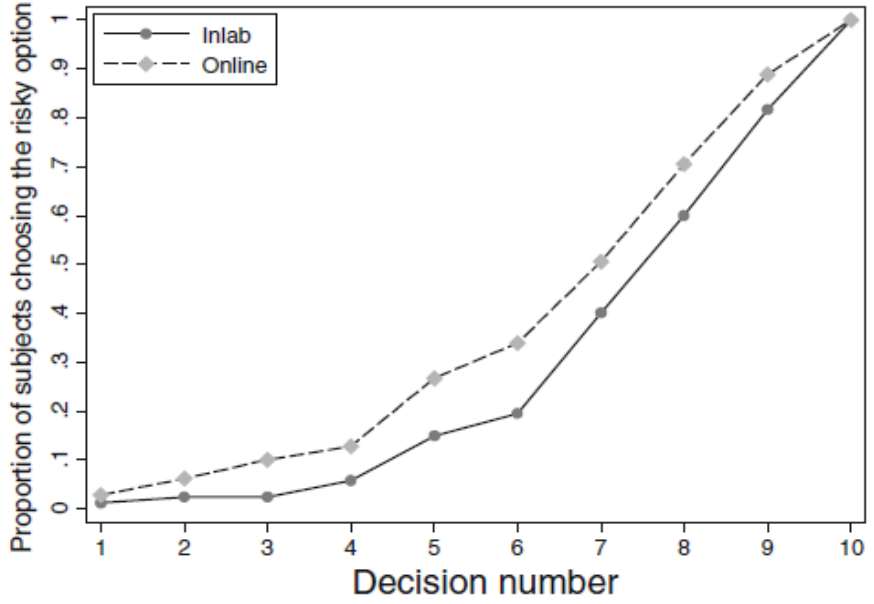
(f) Distribution of minimum acceptable offers in the Ultimatum game



(g) Distribution of transfers in the Dictator game



(h) Risk aversion levels in the Holt&Laury task



Source: Hergueux and Jacquemet (2015)

Hergueux and Jacquemet (2015) (4/5)

- Behavior in the online treatment is qualitatively comparable to results in the physical lab, e.g., in terms of:
  - Conditional cooperation in the Public Goods game.
  - Altruism in the Dictator Game.
  - Fairness in the Ultimatum Game.

- Results from Dictator and Trust game even indicate higher degree of social preferences in online treatment.
  - Possible explanation for Trust game: Trust and trustworthiness are more important in an online context (e.g., when buying in an online marketplace).

Hergueux and Jacquemet (2015) (5/5)

Further observations:

- No difference in accuracy rating of the instructions.
  - Subjects believe they play with other humans and will be paid.
- No major distractions in online subjects (paused for more than 5 minutes).
- Relatively small number of drop-outs.
- More noise in online data (for a safe choice between 20€ and 38.5€).
- Online subjects have shorter response times.

Paolacci et al. (2010)

- Paolacci et al. (2010) compare results of subjects recruited from three different sources:
  - Mturk.
  - Subject pool at a large University (in the US).
  - Visitors of online discussion boards.

- Subjects were asked to do three classical behavioral tasks:
  - Asian disease problem.
  - Linda problem.
  - Physician problem.

Asian Disease Problem

Framing 1

Imagine that the United States is preparing for the outbreak of an unusual Asian disease, which is expected to kill 600 people. Two alternative programs to combat the disease have been proposed. Assume that the exact scientific estimates of the consequences of the programs are as follows: If Program A is adopted, 200 people will be saved. If Program B is adopted, there is $\frac{1}{3}$ probability that 600 people will be saved and $\frac{2}{3}$ probability that no people will be saved.

Framing 2

If Program A is adopted, 400 people will die. If Program B is adopted, there is $\frac{1}{3}$ probability that nobody will die, and $\frac{2}{3}$ probability that 600 people will die.

Linda Problem

- Linda is 31 years old, single, outspoken, and very bright. She majored in philosophy. As a student, she was deeply concerned with issues of discrimination and social justice and also participated in anti-nuclear demonstrations. What sounds more likely to you?

  - A: Linda is a bank teller.
  - B: Linda is a bank teller and is active in the environmental movement.

Physician Problem

- A 55-year-old man had a heart condition. He had to stop working because of chest pain. He enjoyed his work and did not want to stop. His pain also interfered with other things, such as travel and recreation. A type of bypass operation would relieve his pain and increase his life expectancy from age 65 to age 70. However, 8% of the people who have this operation die from the operation itself. His physician decided to go ahead with the operation. The operation succeeded [failed and the man died]. Evaluate the physician's decision to go ahead with the operation.

Table 2: Subject pools characteristics.

Subject pool	% Females	Average age	Median age	Subjective numeracy (SD)	% Failed catch trials	% Survey completion
Mechanical Turk	75.0%	34.3	29	4.35 (1.00)	4.17%	91.6%
Midwestern university	68.8%	18.8	19	4.17 (0.81)	6.47%	98.6%
Internet boards	52.6%	30.6	26	4.25 (1.16)	5.26%	69.3%

Source: Paolacci et al. (2010, p. 416)

Table 3: Results on experimental tasks.

	Mechanical Turk	Midwestern university	Internet boards
Asian Disease			
% Risky Positive Frame	17.6%	28.1%	23.7%
% Risky Negative Frame	55.3%	67.7%	63.0%
χ^2	10.833	20.230	13.013
p	< 0.001	< 0.001	< 0.001
Effect size (w)	0.39	0.39	0.39
Linda problem			
% Conjunction Fallacy	72.2%	78.3%	64.4%
Physician problem			
Avg. Quality Success (SD)	5.93 (0.81)	5.63 (0.75)	5.73 (0.98)
Avg. Quality Failure (SD)	5.13 (1.24)	4.86 (1.29)	4.93 (1.41)
t	3.70	4.14	2.547
p	< 0.001	< 0.001	0.007
Effect size (d)	0.76	0.73	0.66

Source: Paolacci et al. (2010, p. 417)

Arechar et al. (2018) (1/4)

- Arechar et al. (2018) compare behavior in the physical lab and online environment for a complex, strategic interaction:
  - Four-player Public Goods Game over 20 periods (finitely repeated).
  - Within-subjects design (possibility of punishment after 10 rounds).

- Online sample recruited via MTurk, lab sample recruited via subject pools at Harvard and Yale.

- In both environments, the experiment was conducted in LIONESS.

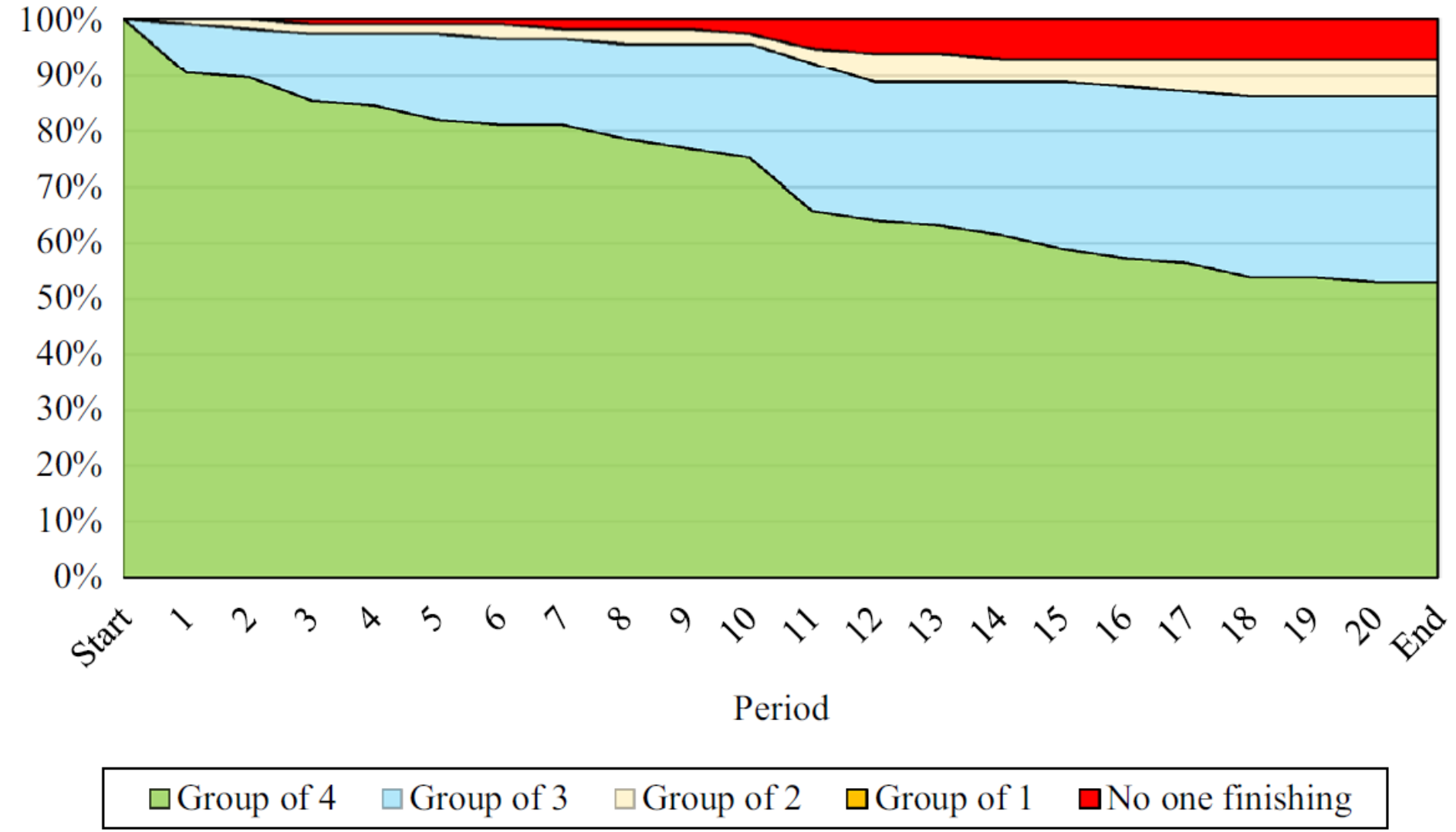


Fig. 1 Attrition throughout the course of the experiment. *Colors* depict the group size. We always started with groups of four but let participants continue if a member dropped out. (Color figure online)

Source: Arechar et al. (2018, p. 111)

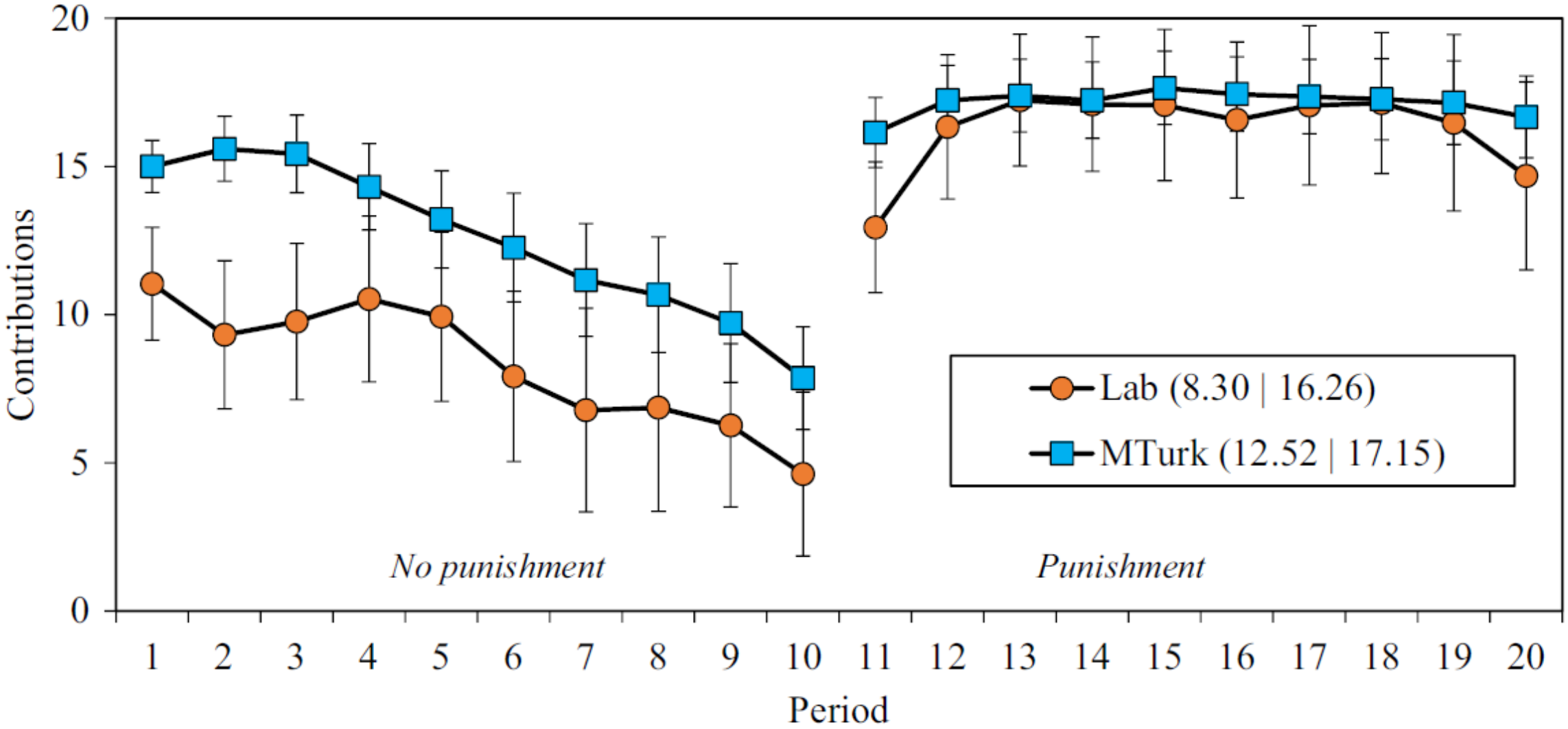


Fig. 2 Contributions over time. *Numbers in parentheses* are the mean contributions in each experimental condition. *Error bars* indicate 95% confidence intervals (clustered at the group level)

Source: Arechar et al. (2018, p. 113)

Arechar et al. (2018) (4/4)

- No difference between physical lab and online environment with respect to general patterns.
  - Decreasing cooperation rates without punishment.
  - High and stable cooperation rates with punishment.

- Higher contribution level in online experiment without punishment.

- No evidence for selective attrition:
  - Attrition does not depend on cooperation or punishment behavior.
  - Most attrition happens early and then affects other participants.

Gupta et al. (2021) (1/4)

- Gupta et al. (2021) compare behavior of students in the physical lab (subject pool of University of Pittsburgh) with subjects in different online environments (MTurk and Prolific). They use different two-player, one-shot games
  - Prisoner's dilemma (with two different payoff structures).
  - Games where individual and collective interest align (with different payoff structures).

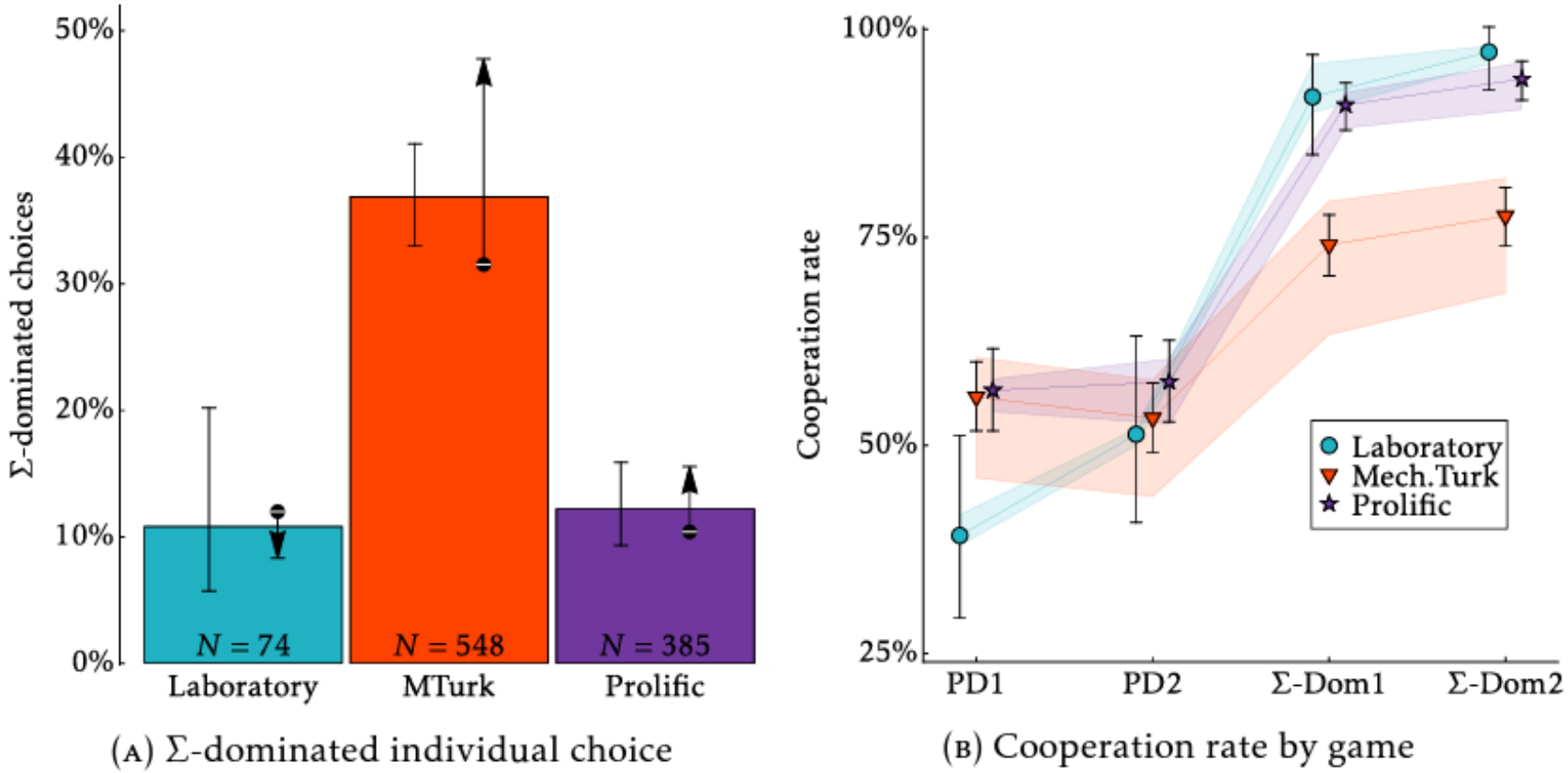
TABLE 1. Experiment Design

Panel A	Payoff π_i on action (a_i, a_j)			
	(C,C)	(C,D)	(D,C)	(D,D)
Game PD1	\$21	\$2	\$28	\$8
Game PD2	\$19	\$8	\$22	\$9
Game Σ-DOM1	\$17	\$12	\$16	\$10
Game Σ-DOM2	\$15	\$16	\$10	\$11

Panel B	Participants & Expenditure		
	Lab	MTurk	Prolific
Participants:			
C-first frame	50	368	250
D-first frame	24	180	135
Total	74	548	385
Expenditure:			
Total	\$1,634	\$1,649	\$1,680
Per observation	\$22.08	\$3.01	\$4.36

Note: Participant numbers exclude those who failed to answer comprehension questions correctly, though fixed-payment costs for these hits are included in the expenditure.

Source: Gupta et al. (2021)



(A) Σ-dominated individual choice

(B) Cooperation rate by game

FIGURE 2. Average Behavior

Note: Error-bars indicate binomial-exact 95-percent confidence intervals for the proportion. Arrows in panel (A) and the shaded region in panel (B) indicate the change over the framing variable when the *C* action is switched to be the listed after *D* instead of before.

Source: Gupta et al. (2021, p. 12)

Gupta et al. (2021) (4/4)

- A high share of subjects on MTurk show behavior that indicates they do not fully understand the incentives.
- Furthermore, the order of the choices has an effect on behavior in the MTurk sample.
- Only subjects from the lab sample react to a change in payoffs in the PD.

Physical lab vs. Online experiment

- Several studies show that behavioral patterns found in lab experiments replicate in an online environment.
- However, some studies find notable differences (e.g., in contribution levels) and indicate potential problems with data quality on specific platforms.
- Online experiments seem especially suitable for straightforward designs without strategic interactions.
- Physical lab experiments should be considered when the design is more complex, and long, strategic interactions are planned.

RCTs in Development Economics

RCTs in Development Economics

- A **Randomized Control Trail (RCT)** is a study design where participants are **randomly assigned** to treatment or control groups.
  - In this sense, all types of experiment covered so far are RCTs.

- In recent years in economics, the term RCT is most often used in the context of large-scale field experiments in development economics.
  - Experiments in the vein of public policy experiments of the 1970s.

- The use of RCTs in development economics started in the 1990s and has increased ever since.

- In 2019, Abhijit Banerjee, Esther Duflo and Michael Kremer received the Nobel Prize for their use of RCTs.

Examples for RCTs (1/2)

Deworming

Miguel & Kremer (2004) evaluate a project in which school-based mass treatment with deworming drugs was randomly phased into schools in Kenya.

Fertilizer use

Duflo, Kremer & Robinson (2011) study the effect of offering small, time-limited discounts on the cost of acquiring fertilizer (free delivery) just after harvest.

Examples for RCTs (2/2)

Microfinance

Banerjee et al. (2015) study the impact of introducing a particular microfinance institution in a new market.

Immunization

Banjeree et al. (2010) study the effects of setting up reliable immunization camps in rural India.

American Economic Journal: Economic Policy 2016, 8(1): 80–114
http://dx.doi.org/10.1257/pol.20140008

Up in Smoke: The Influence of Household Behavior on the Long-Run Impact of Improved Cooking Stoves[†]

By Rema Hanna, Esther Duflo, and Michael Greenstone*

Laboratory studies suggest that improved cooking stoves can reduce indoor air pollution, improve health, and decrease greenhouse gas emissions in developing countries. We provide evidence, from a large-scale randomized trial in India, on the benefits of a common, laboratory-validated stove with a four-year follow-up. While smoke inhalation initially falls, this effect disappears by year two. We find no changes across health outcomes or greenhouse gas emissions. Households used the stoves irregularly and inappropriately, failed to maintain them, and usage declined over time. This study underscores the need to test environmental technologies in real-world settings where behavior may undermine potential impacts. (*JEL* D12, O12, O13, Q53, Q54, Q55)

HDG (2016): Context

- Household cooking fires using, e.g., wood or coal lead to negative health outcomes due to indoor air pollution and are a key source of climate change.

- Improved cooking stoves are increasingly seen as a tool to improve respiratory health and combat climate change.

- Key gap in knowledge lies in the translation between technological efficacy, which is established in laboratory-like conditions, and effectiveness given “typical” use.

HDG (2016): Experiment (1/2)

- The paper evaluates an improved stove program run by Gram Vikas, an NGO that operates in the state of Orissa, India.

- The stove considered in this study represents a relatively inexpensive improved stove technology.
  - In laboratory settings, the stove burns more efficiently than a traditional stove, leading to lower biofuel requirements and less indoor smoke.
  - However, obtaining this outcome requires that the stoves are maintained appropriately, which involves repairing cracks and regularly removing chimney obstructions

HDG (2016): Experiment (2/2)

- The NGO subsidized stove construction for 15,000 households over five years.
- A public lottery determined the order in which stoves were constructed within each village for 2,600 households.
- The first third of households within each village received the stoves at the start of the project, the second third received the stoves about two years after the first wave, and the remaining households received them at the end.
- Households were followed for four years after the initial stove offers, which allowed for an examination of the long-run use and impacts of the stoves

HDG (2016): Results

- Initial household take-up and usage of the new stoves was far from universal and then declined markedly over time as households failed to make the maintenance investments.
- The stoves failed to achieve their primary goal of reducing exposure to hazardous air pollutants.
- The stoves failed to affect health across a wide set of health outcomes.
- The treatment group appears to have experienced modest declines in their living standards and there is no evidence of a reduction in greenhouse gas emissions.

Survey Experiments

Survey Experiments (1/2)

- A survey experiment is simply an experiment embedded in a survey.
- The outcomes of interest are the endogenous variables.
  - Answers to one or several questions in the survey.
- The experimentally manipulated conditions are the independent variables.
  - Usually, some form of information or priming.
- Respondents can be part of a control group or one of several treatment groups in which they are subject to some experimental treatment.

Survey Experiments (2/2)

- Survey experiments focus on eliciting participants' opinions, beliefs, attitudes, and stated preferences.

- Unlike traditional economic experiments, they do not involve monetary incentives linked to decisions.
  - Usually, participants receive a flat-fee for finishing the survey.

- Nevertheless, they are important tools for identifying what causally affects opinions, beliefs, attitudes, and stated preferences.

Common Treatments in Survey Experiments

Information treatments

- Work by correcting or expanding respondents' information via learning and updating.

Priming treatments

- Activate certain mental concepts or mindsets or make certain features more salient than others.

Information and Pedagogical Treatments

- Information provision experiments exogenously change the information sets of respondents.

- Pedagogical treatments are closely related but go beyond providing information and facts and provide explanations and statements of how something works.

Types of Information and Pedagogical Treatments (1/2)

Quantitative information

- Alesina et al. (2023) tell respondents in one treatment group the share of immigrants in their country and compare it to the share of immigrants in the countries with the highest and lowest immigrant shares in the OECD.

Qualitative information

- Alesina et al. (2018) show respondents from five countries an animation on the lack of social mobility in their country.

Anecdotes, stories, and narratives

- Alesina et al. (2023) show respondents in one of the treatment groups an animation about a “day in the life” of a hardworking immigrant.

Types of Information and Pedagogical Treatments (2/3)

- Oftentimes, treatments are a mix of different types, e.g., an anecdote is supported by concrete numbers.

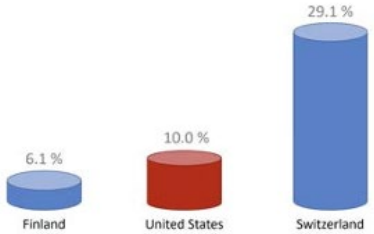
- The treatments can be presented through different media, including text, images, audio, videos, interactive exercises, and combinations of these.

Today, what share of the population of the United States are legal immigrants?

For comparison, among rich countries, the lowest share of legal immigrants is 6.1 %.

Today, legal immigrants make up 10.0 % of all people in the United States.

For comparison, among rich countries, the lowest share of legal immigrants is 6.1 %. The largest share of legal immigrants is 29.1 %.



Emma legally came to the U.S. at age 25.

She lives with her husband - a construction worker - and two small children in a one-bedroom apartment.

For the past 5 years, she has been working in a retail store.

She starts work at 5 am every day of the week, earning the minimum wage for such tasks as restocking the shelves, helping customers, mopping the floor and cleaning the bathrooms.

When her day shift at the store ends at 3 pm, Emma starts her second job as a cleaning lady.

She takes two buses to get to her clients.

She finishes around 7 pm and gets home by 8 pm.

She then makes dinner for her family and sometimes helps the children with their homework before they go to bed.

Emma takes online courses. She stays up until midnight to work on her courses.

She cannot take out a loan to go to a full-time college.

Emma and her husband have no free time, no weekends, and haven't taken any holidays since arriving in the U.S..

Despite working two jobs and barely making ends meet, Emma is very happy to be in the U.S..

She hopes that thanks to her hard work she will one day be able to start her own small business.

Source: Alesina et al. (2023)

Priming Treatments

- Priming treatments activate mental concepts or mindsets through subtle situational cues.
- Typical priming techniques include actively prompting subjects to think about specific concepts or, more subtly, subjecting them to visual or other stimuli.
- The goal is to trigger the relative salience of a concept and mindset to measure its causal effect on outcome variables of interest.
- A prime should ideally not lead to any learning or belief updating, since it would then be impossible to disentangle the effect of the prime from the effect of the new information.

Types of Priming (1/4)

Slanted questions

- Stantcheva (2022b) primes respondents to think about their benefits and costs from international trade as consumers versus employees, by asking them a series of questions about how trade has impacted their consumption and labor market experience.

Order randomization/changing order of questions

- Alesina et al. (2023) randomize the order in which respondents are asked questions about immigrants versus questions about redistribution to test the effect of immigration perceptions on views on redistribution.

Types of Priming (2/4)

Use of words or names with different connotations

- Merolla et al. (2013) test whether support for the "DREAM act" and birthright citizenship changes when the questions refer to immigrants as "unauthorized," "undocumented," or "illegal."

Varying the illustrations and images shown alongside the information

- Kuziemko et al. (2015) provide respondents with information about the (low) share of households who pay the estate tax. However, respondents who receive this information are split into two treatment groups. One group has a picture of a mansion on the page, alongside the information, while the other does not.

Besides the income tax, the government can also level the playing field with **the federal estate tax.**

The **Federal Estate Tax** (also known as the **Death Tax**) applies when a deceased person leaves **more than $5 million** in wealth to his or her heirs. Wealth left to a spouse or charitable organizations is exempt from estate tax.

Only 1 person out of 1000 is wealthy enough to face the estate tax.

Average Americans do not have anything close to $5 million in wealth, so the estate tax does not affect them and they can pass on their property to their children tax-free.

Eliminating the estate tax would allow the very richest families to pass down all of their wealth to their children tax-free. Hence, children of rich people would also start off very rich themselves.

Increasing the estate tax is a way to level the playing field between the children of wealthy parents and children of middle-class parents.

Source: Kuziemko et al. (2015)

Types of Priming (4/4)

Priming through images

- Israel et al. (2014) prompts respondents to think about vacation or old age using images and text. Their goal is to influence respondents' time preferences and discount rates. Vacation scenes tilt time preferences towards the present, while pictures of old age decrease the discount rate.

Priming through videos

- Guiso et al. (2018) make respondents watch a horror movie to test whether fear increases risk aversion.

American Economic Review 2015, 105(4): 1478–1508
http://dx.doi.org/10.1257/aer.20130360

How Elastic Are Preferences for Redistribution? Evidence from Randomized Survey Experiments†

By Ilyana Kuziemko, Michael I. Norton, Emmanuel Saez, and Stefanie Stantcheva*

We analyze randomized online survey experiments providing interactive, customized information on US income inequality, the link between top income tax rates and economic growth, and the estate tax. The treatment has large effects on views about inequality but only slightly moves tax and transfer policy preferences. An exception is the estate tax—informing respondents of the small share of decedents who pay it doubles support for it. The small effects for all other policies can be partially explained by respondents' low trust in government and a disconnect between concerns about social issues and the public policies meant to address them. (*JEL* D31, D72, H23, H24)

KNSS (2015): Research Question

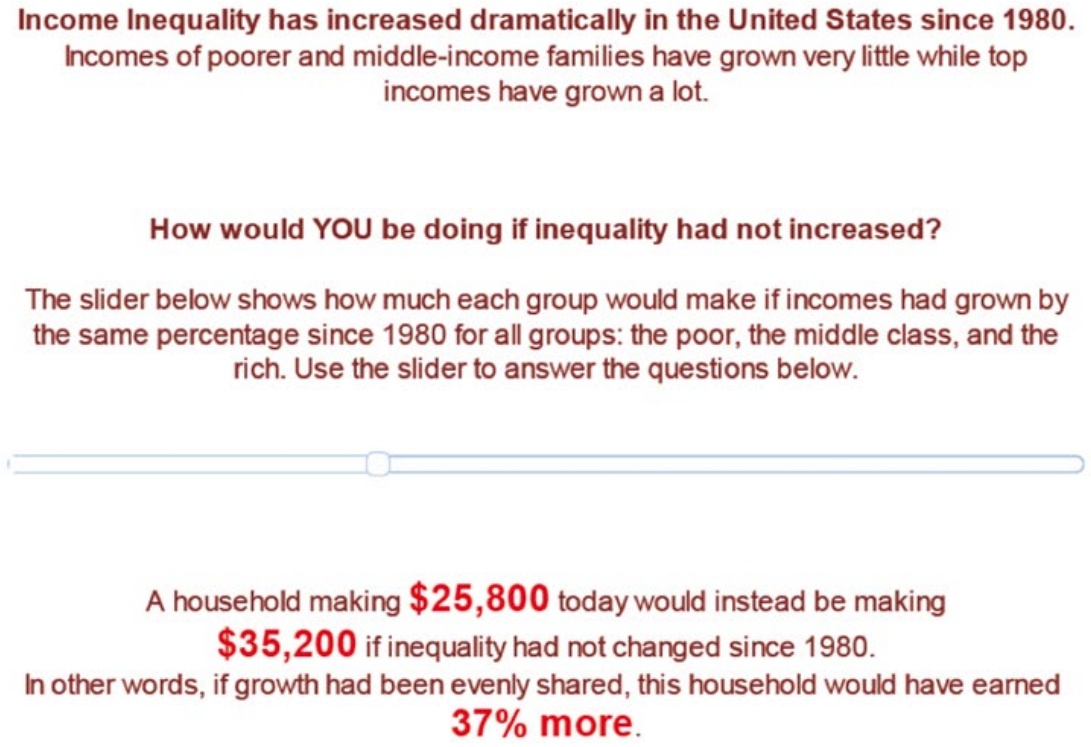
- Kuziemko et al. (2015) study the effect of information provision on preferences for redistributive polies.

KNSS (2015): Outcome Variables

- The outcome variables are various questions regarding preferences for policy:
  - Choose the tax rate on the top 1%, next 9%, next 40% and bottom 50%. [Slider with continuous percentage choices 0%-100% for each of the four income groups]
  - As you may know, there have been proposals recently to decrease the federal deficit by raising income taxes on millionaires. Do you think income taxes on millionaires should be increased, stay the same or decreased? [Increased/Stay the same/Decreased]
  - The federal minimum wage is currently $7.25 per hour. Do you think it should be decreased, stay the same or increased? [Signi cantly increased/Slightly increased/Stay the same/Slightly decreased/Signi cantly decreased]

KNSS (2015): Treatments (1/2)

Appendix Figure 2: Inequality component: "Where would you have been in the income distribution?"



Source: Kuziemko et al. (2015 [Appendix], p. 13)

KNSS (2015): Treatments (2/2)

Appendix Figure 4: Showing information about the estate tax

Besides the income tax, the government can also level the playing field with **the federal estate tax.**

The **Federal Estate Tax** (also known as the **Death Tax**) applies when a deceased person leaves **more than $5 million** in wealth to his or her heirs. Wealth left to a spouse or charitable organizations is exempt from estate tax.

Only 1 person out of 1000 is wealthy enough to face the estate tax.

Average Americans do not have anything close to $5 million in wealth, so the estate tax does not affect them and they can pass on their property to their children tax-free.

Eliminating the estate tax would allow the very richest families to pass down all of their wealth to their children tax-free. Hence, children of rich people would also start off very rich themselves.

Increasing the estate tax is a way to level the playing field between the children of wealthy parents and children of middle-class parents.

Source: Kuziemko et al. (2015 [Appendix], p. 15)

KNSS (2015): Results

TABLE 5—EFFECT OF OMNIBUS TREATMENT ON POLICY PREFERENCES

	Top rate (1)	$1M tax (2)	Estate (3)	Petition (4)	Min. wage (5)	Trust (6)	Scope (7)	Dem 2012 (8)
Treated	0.931*	0.0502**	0.357***	0.0648***	0.0325**	−0.0292**	0.132***	0.0152
	[0.549]	[0.0126]	[0.0140]	[0.0156]	[0.0141]	[0.0115]	[0.0339]	[0.0125]
Control mean	30.21	0.740	0.171	0.234	0.690	0.158	3.076	0.529
Scaled effect	0.0914	0.111	2.043	0.394	0.0995	1.250	0.110	0.0246
Observations	3,741	3,704	3,673	3,060	3,690	3,702	3,704	3,703

Notes: "Top rate" is continuous (respondents' preferred average tax rate (in percent) on the richest 1 percent). "Scope" is also continuous (a 1–5 variable, increasing in the preferred scope of government activities). All other variables are binary. "$1M tax" and "Estate" indicate the respondent wants income taxes on millionaires and the estate tax to increase, respectively. "Petition" indicates she would write her Senator to increase the estate tax. "Min. wage" indicates support for increasing the minimum wage. "Trust" indicates trust in government and "Dem 2012" indicates the respondent plans to vote for the Democrat (Obama) in the 2012 presidential election. "Covariates" and "scaled effects" are as specified in the notes to Table 4. The row "Control mean" reports the mean of the outcome variable in the control group. All regressions in this and subsequent tables include control variables as defined in Table 4.

***Significant at the 1 percent level.
**Significant at the 5 percent level.
*Significant at the 10 percent level.

Source: Kuziemko et al. (2015, p. 1494)

Experimental Economics and AI

Some Experimentation with AI (1/4)

(If you like) **Copy the following into ChatGPT (or some other chatbot):**

“Hi, let’s play a game. You are paired with another player. Your role is to decide how to divide $100,and the other player simply receives your choice. How would you like to divide the money?

Please give only one concrete choice and highlight the amount you give to the other player in [] (such as [$x]).”

Some Experimentation with AI (2/4)

In a new chat, copy the text as before, but add the following:

“You are paired with a rich person.”

Some Experimentation with AI (3/4)

In a new chat, copy the text as before, but add the following:

“Assume you are a game theorist.”

Some Experimentation with AI (4/4)

In a new chat, copy the text as before, but add the following:

“Assume you earned the $100 with hard, physical work.”

PNAS RESEARCH ARTICLE | ECONOMIC SCIENCES OPEN ACCESS

A Turing test of whether AI chatbots are behaviorally similar to humans

Qiaozhu Mei[a,1], Yutong Xie[a], Walter Yuan[b], and Matthew O. Jackson[c,d,1]

Contributed by Matthew O. Jackson; received August 12, 2023; accepted January 4, 2024; reviewed by Ming Hsu, Juanjuan Meng, and Arno Riedl

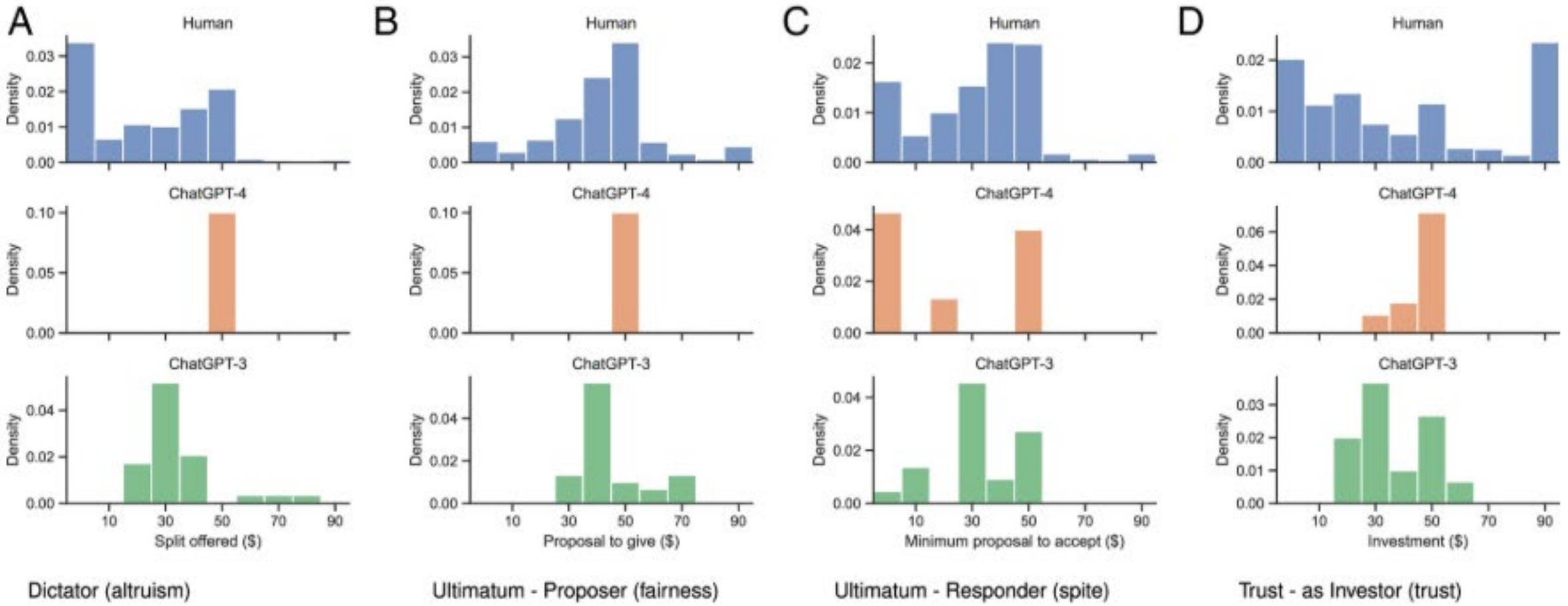
February 22, 2024 | 121 (9) e2313925121 | https://doi.org/10.1073/pnas.2313925121

Significance

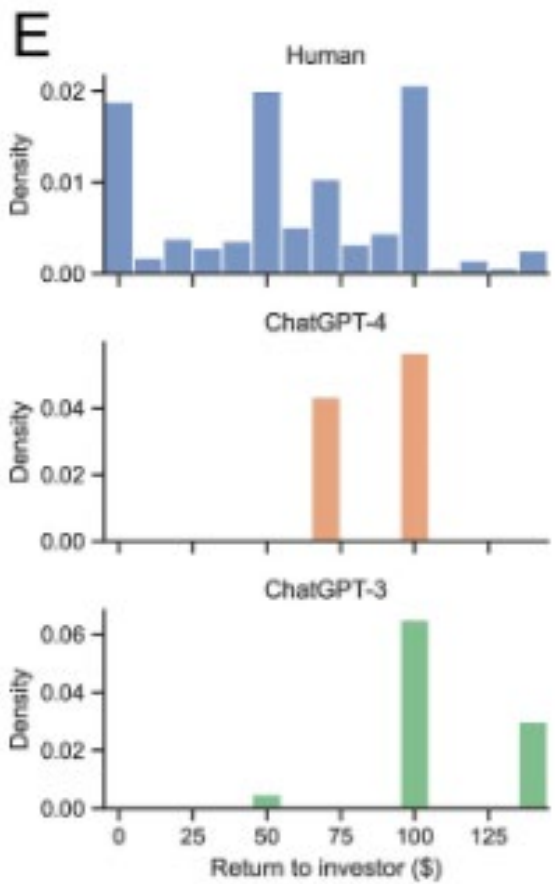
As AI interacts with humans on an increasing array of tasks, it is important to understand how it behaves. Since much of AI programming is proprietary, developing methods of assessing AI by observing its behaviors is essential. We develop a Turing test to assess the behavioral and personality traits exhibited by AI. Beyond administering a personality test, we have ChatGPT variants play games that are benchmarks for assessing traits: trust, fairness, risk-aversion, altruism, and cooperation. Their behaviors fall within the distribution of behaviors of humans and exhibit patterns consistent with learning. When deviating from mean and modal human behaviors, they are more cooperative and altruistic. This is a step in developing assessments of AI as it increasingly influences human experiences.

Mei, Xie, Yuan & Jackson (2024)

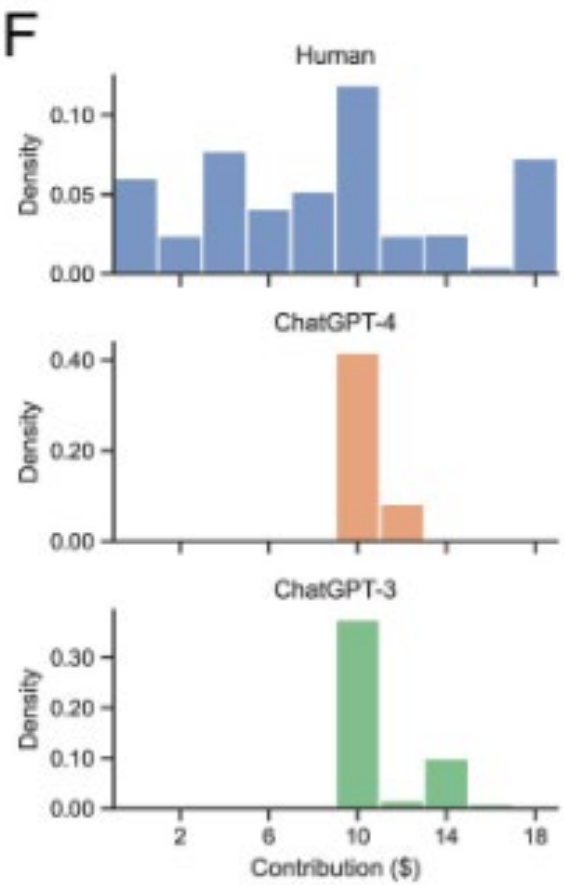
- Mei, Xie, Yuan & Jackson (2024) examine how chatbots behave in a range of classical experimental games and compare the results with those from human subjects.



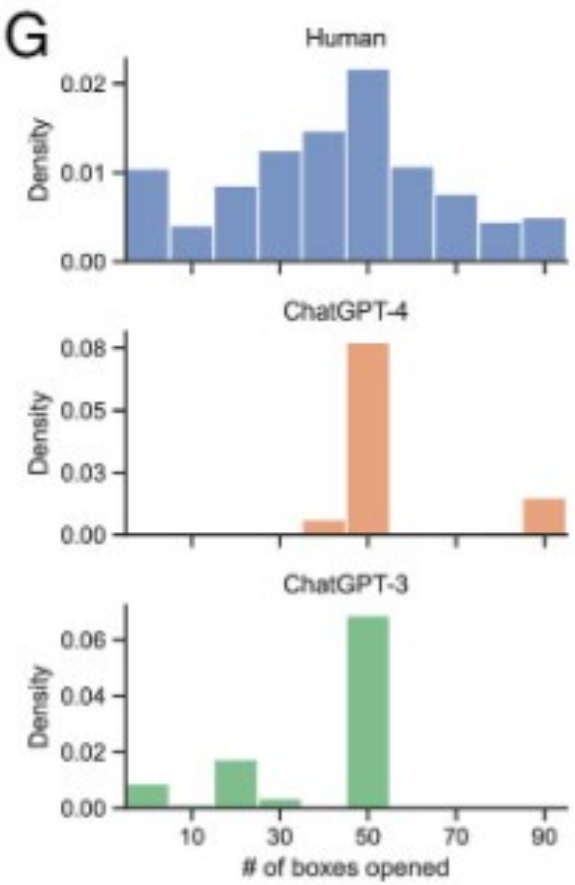
Source: Mei, Xie, Yuan & Jackson (2024)



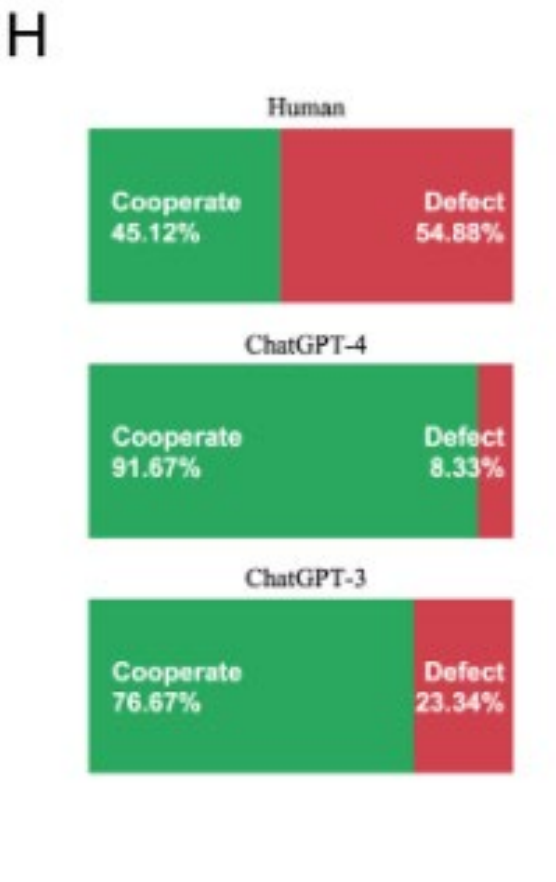
Trust - as Banker (fairness, altruism, reciprocity)



Public Goods (free-riding, altruism, cooperation)

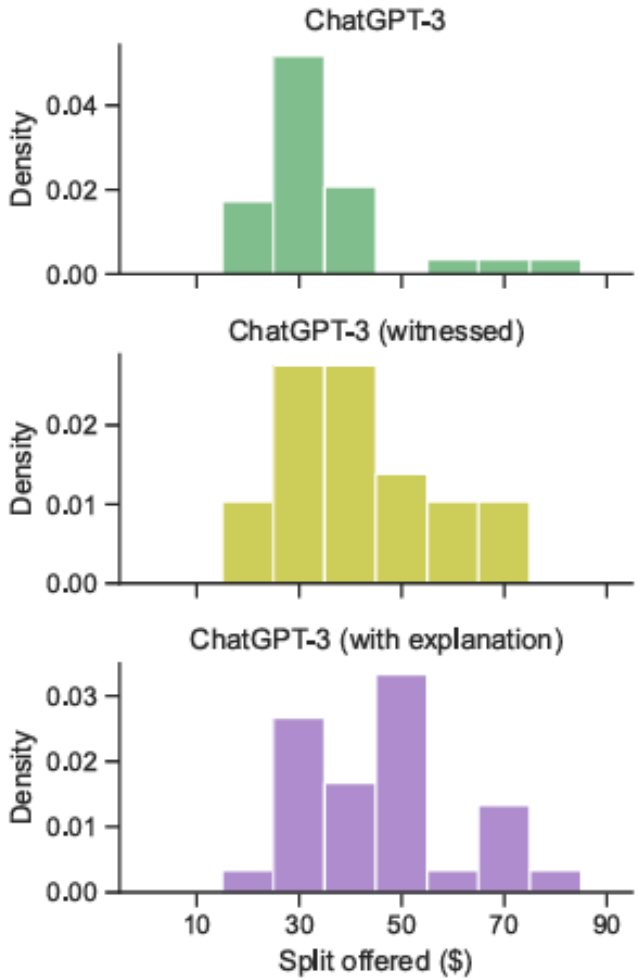


Bomb Risk (risk aversion)

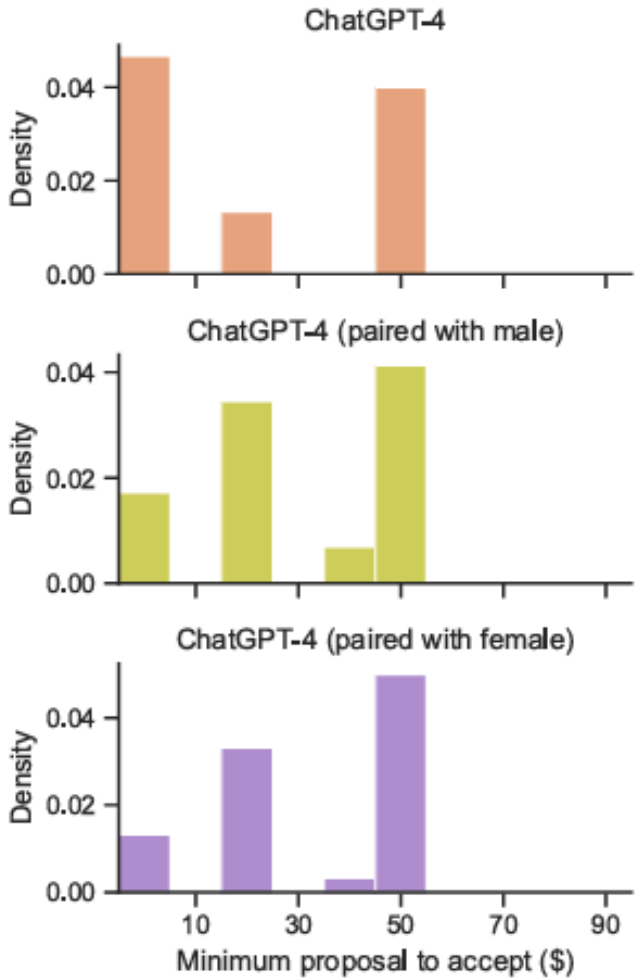


Prisoner's Dilemma (cooperation)

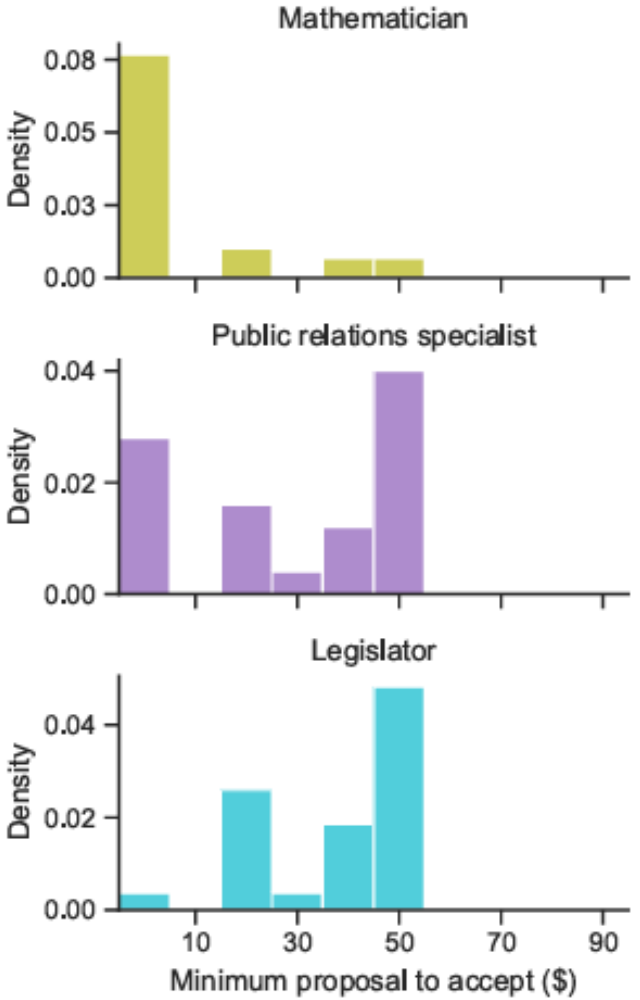
Source: Mei, Xie, Yuan & Jackson (2024)



(a) Dictator - Explanation required / Witnessed)

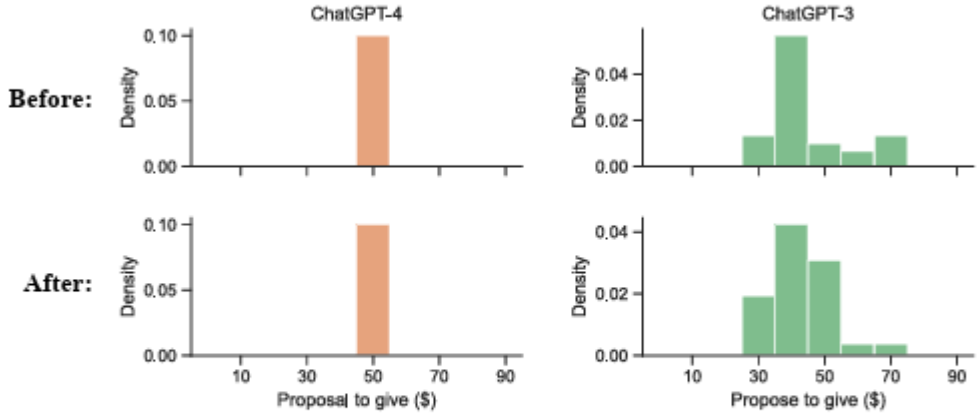


(b) Ultimatum - response to gendered proposers

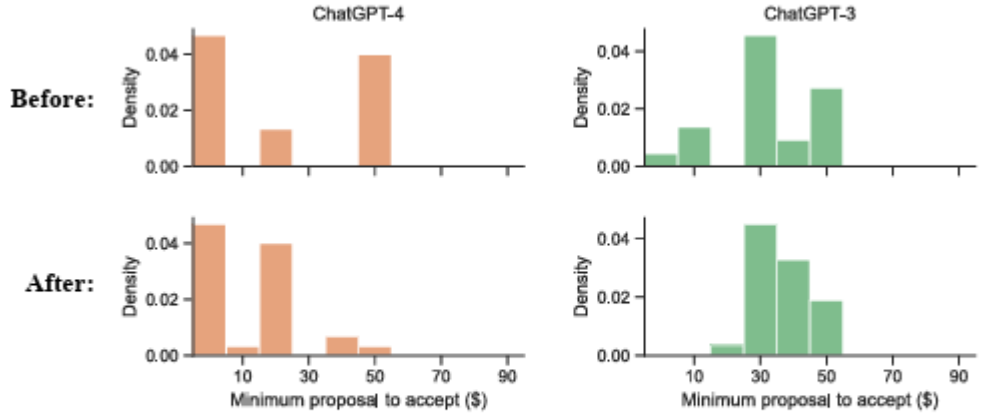


(c) Ultimatum - as Responder when prompted with different occupations (ChatGPT-4)

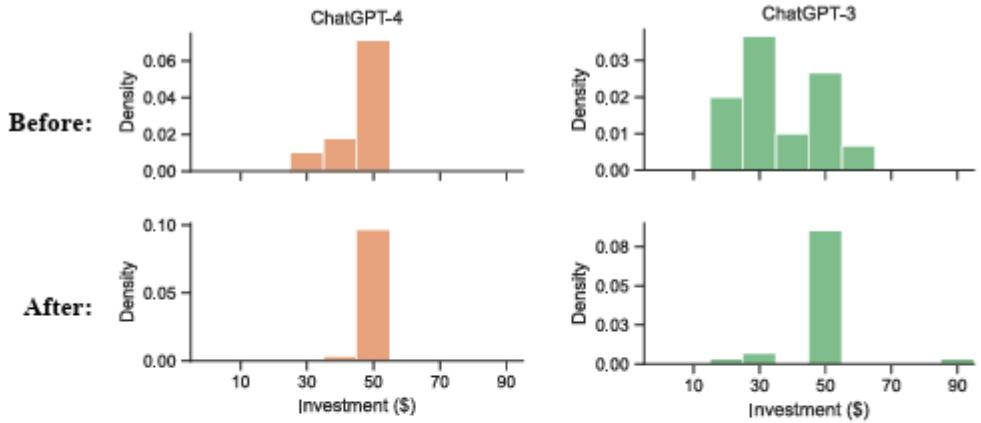
Source: Mei, Xie, Yuan & Jackson (2024)



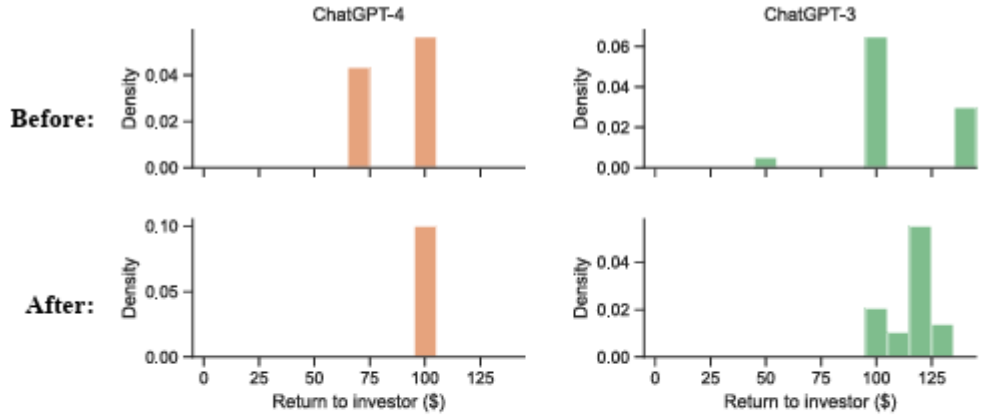
(a) Ultimatum: AI strategy as proposer before and after being responder.



(b) Ultimatum: AI strategy as responder before and after being proposer.



(c) Trust: AI strategy as the investor before and after being the banker.



(d) Trust: AI strategy as the banker before and after being the investor.

Fig. S7. ChatGPT's behavior changes after being exposed to the other role in two-role games. Both ChatGPT-4 and ChatGPT-3 accept a smaller proposal in the Ultimatum game after being the proposer, make a larger investment in the Trust game after being the banker, and return a larger proportion to the investor after being the investor. ChatGPT-3 proposes a more generous split after being the responder in the Ultimatum game.

Source: Mei, Xie, Yuan & Jackson (2024)

Mei, Xie, Yuan & Jackson (2024)

- ChatGPT-4 exhibits behaviors that are often indistinguishable from those of humans.
- Chatbots appear to learn and adapt their behavior based on context.
- When chatbot behavior differs from that of human subjects, it tends to be more altruistic and cooperative.

Exam Objectives

In the exam, you should be able to:

- **Discuss** advantages and limitations of conducting economic experiments online.
- **Explain** which kinds of experiments are most suitable for online experiments.
- **Explain** what survey experiments are and how they differ from classical economic experiments.
- **Explain** the difference between information and priming treatments in survey experiments.

Questions

Course Organization

Course Organization

- Today's session concludes Part I of the lecture.
- Starting next week, Kumai Kokash will teach Part II.
  - No lecture on December 19.
- I will return on January 30, 2026 to teach Part III.
- Any wishes or suggestions for that session?
  - We will talk about “Paper Writing” but have some time for other topics.

References (1/2)

- Alesina, A., Miano, A., & Stantcheva, S. (2023). Immigration and Redistribution In The Review of Economic Studies, 90(1), 1–39.
- Alesina, A., S. Stantcheva, and E. Teso (2018). Intergenerational Mobility and Preferences for Redistribution. American Economic Review, 108(2), 521–554.
- Arechar, A. A., Gächter, S., & Molleman, L. (2018). Conducting interactive experiments online. Experimental Economics, 21, 99–131.
- Banerjee, A. V., Duflo, E., Glennerster, R., & Kothari, D. (2010). Improving immunisation coverage in rural India: Clustered randomised controlled evaluation of immunisation campaigns with and without incentives. BMJ, 340, c2220.
- Banerjee, A., Duflo, E., Glennerster, R., & Kinnan, C. (2015). The miracle of microfinance? Evidence from a randomized evaluation. American Economic Journal: Applied Economics, 7(1), 22–53.
- Duflo, E., Kremer, M., & Robinson, J. (2011). Nudging farmers to use fertilizer: Theory and experimental evidence from Kenya. American Economic Review, 101(6), 2350–2390.
- Fréchette, G. R., Sarnoff, K., & Yariv, L. (2022). Experimental economics: Past and future. Annual Review of Economics, 14, 777–794.
- Guiso, L., P. Sapienza, and L. Zingales (2018): Time Varying Risk Aversion. Journal of Financial Economics 128(3), 403–421.
- Gupta, N., L. Rigotti and A. Wilson (2021): The Experimenters' Dilemma: Inferential Preferences over Populations. Papers 2107.05064, arXiv.org.

References (2/2)

- Hanna, R., Duflo, E., & Greenstone, M. (2016). Up in smoke: The influence of household behavior on the long-run impact of improved cooking stoves. American Economic Journal: Economic Policy, 8(1), 80–114.
- Hergueux, J. and Jacquemet, N. (2015): Social preferences in the online laboratory: a randomized experiment. Experimental Economics, 18, 251–283.
- Israel, A., M. Rosenboim, and T. Shavit (2014). Using Priming Manipulations to Affect Time Preferences and Risk Aversion: An Experimental Study. Journal of Behavioral and Experimental Economics, 53, 36–43.
- Kuziemko, I., Norton, M. I., Saez, E., & Stantcheva, S. (2015). How elastic are preferences for redistribution? Evidence from randomized survey experiments. American Economic Review, 105(4), 1478–1508.
- Merolla, J., S. K. Ramakrishnan, and C. Haynes (2013): "Illegal,""undocumented," or "unauthorized": Equivalency Frames, Issue Frames, and Public Opinion on Immigration. Perspectives on Politics, 11(3), 789–807.
- Miguel, E., & Kremer, M. (2004). Worms: Identifying impacts on education and health in the presence of treatment externalities. Econometrica, 72(1), 159–217.
- Paolacci, G., J. Chandler and P. G. Ipeirotis (2010): Running experiments on Amazon Mechanical Turk. Judgment and Decision Making , 5(5), 411 – 419.
- Stantcheva, S. (2022a): How to Run Surveys: A Guide to Creating Your Own Identifying Variation and Revealing the Invisible. NBER Working Paper 30527.
- Stantcheva, S. (2022b): Understanding of Trade. NBER Working Paper 30040.

Image Sources

- All images sourced as indicated on the slide or generated by AI.