
# Design and Implementation of Online Behavioral Experiments

**nodeGame.org**

Stefano Balietti

*MZES and Heidelberg*

# Introduction:

# From Lab to Online

@balietti
@nodegameorg
stefanobalietti.com@gmail.com

# Your Instructor: Stefano Balietti

http://stefanobalietti.com

**Currently**

- Fellow in Sociology Mannheim Center for European Social Research (MZES)
- Postdoc at the Alfred Weber Institute of Economics at Heidelberg University

**Previously**

- Microsoft Research - Computational Social Science New York City
- Postdoc Network Science Institute, Northeastern University
- Fellow IQSS, Harvard University
- PhD, Postdoc, Computational Social Science, ETH Zurich

# Your Instructor: Stefano Balietti

http://stefanobalietti.com

**Main Research Interests:**

- Consensus formation, social influence, and polarization
- Inequality, redistribution, and preferences thereof
- Incentives schemes for peer review systems
- Optimal experimental design
- Philosophy of science, in particular Paul Feyerabend

# Goals of the Workshop

1. Review *fundamental* concepts of experimental design
2. Understand *challenges* of *opportunities* for online experiments

# Goals of the Workshop

1. Review *fundamental* concepts of experimental design
2. Understand *challenges* of *opportunities* for online experiments

# Goals of the Workshop

1. Review *fundamental* concepts of experimental design
2. Understand *challenges* of *opportunities* for online experiments
3. Acquire *solid* knowledge of JavaScript and Node.JS programming

# Goals of the Workshop

1. Review *fundamental* concepts of experimental design
2. Understand *challenges* of *opportunities* for online experiments
3. Acquire *solid* knowledge of JavaScript and Node.JS programming
4. Implement your own experiment with the nodeGame framework

# Goals of the Workshop

1. Review *fundamental* concepts of experimental design
2. Understand *challenges* of *opportunities* for online experiments
3. Acquire *solid* knowledge of JavaScript and Node.JS programming
4. Implement your own experiment in the nodeGame framework
5. Learn how to run your experiment with digital clouds and online labor markets (Amazon Mechanical Turk)

# Experimentation: from Lab to Online

# What Is an Experiment?

# What Is an Experiment?

- An experiment is a ***methodological procedure*** carried out with the goal of verifying, falsifying, or establishing the validity of a hypothesis.

- A ***test*** under ***controlled conditions*** that is made to demonstrate a known truth, examine the validity of a hypothesis, or determine the efficacy of something previously untried.

- An experiment is an empirical method that ***arbitrates between competing hypotheses***.

# What Is a Synchronous Experiment?

- It is an experiment where participants
  - (i) interact in a **common environment**,
  - (ii) at the **same time**,
  - (iii) and their actions have an ***immediate*** **effect** on the decision process, and experimental outcome (e.g., monetary payoff)

- Synchronous experiments can be **turn-based** or **real-time**

# Internal Validity

The extent to which a causal inference based on a study is warranted by an experiment.

# 3 Pillars of Experimental Design

**Control**

**Independence**

**Randomization**

# 3 Pillars of Experimental Design

## Control

Deliberate choice of the variables that will be kept constants and those that will be varied: **treatment variables**

# 3 Pillars of Experimental Design

**Independence**

Treatment variables must *not* be correlated

# 3 Pillars of Experimental Design

## Randomization

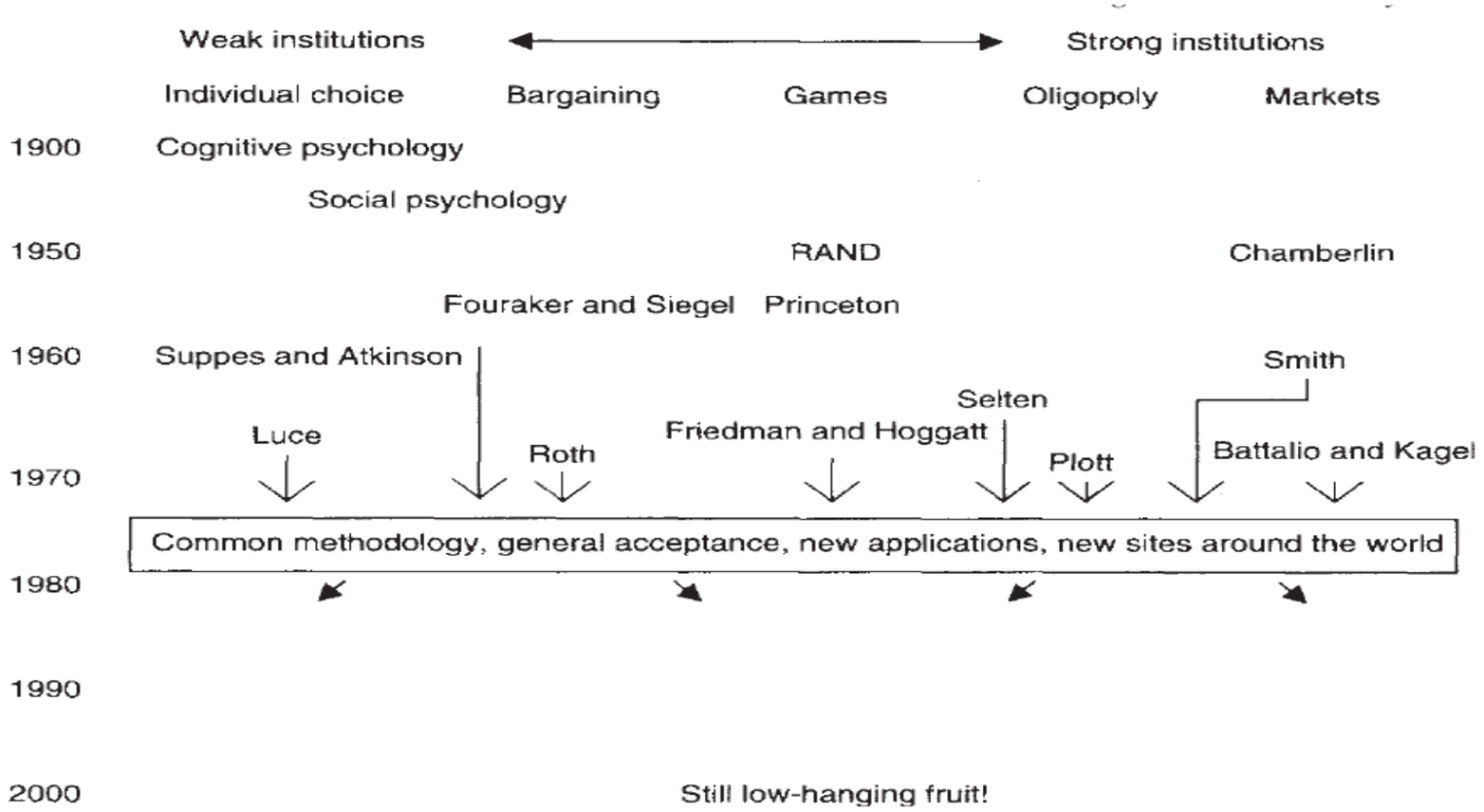
- Participants are **randomly** assigned to treatments
- Treatments are executed in controlled order
- **Stratification** might be required

# Short History of Experimental Research

"There is a property common to almost all the moral sciences, and by which they are distinguished from many of the physical; that is, that ***it is seldom in our power to make experiments with them***"

**John Stuart Mill, 1836**

# Short History of Experimental Research

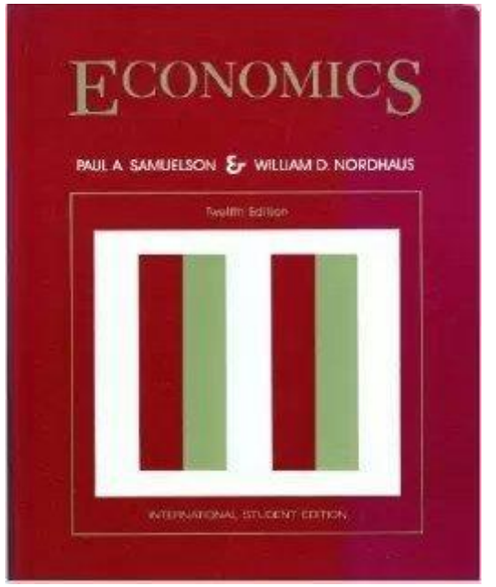



Friedman, Cassar (2004) Economists go to the laboratory. Who, what, when and why.

# Short History of Experimental Research

"Economics unfortunately cannot perform the controlled experiments of chemists or biologists because [it] cannot easily control other important factors. Like astronomers or meteorologists, [it] *generally must be content largely to observe.*"

Samuelson and Nordhaus 1985, Economics Textbook (italics added)




ENCYCLOPÆDIA Britannica

"There is no laboratory in which economists can test their hypotheses"

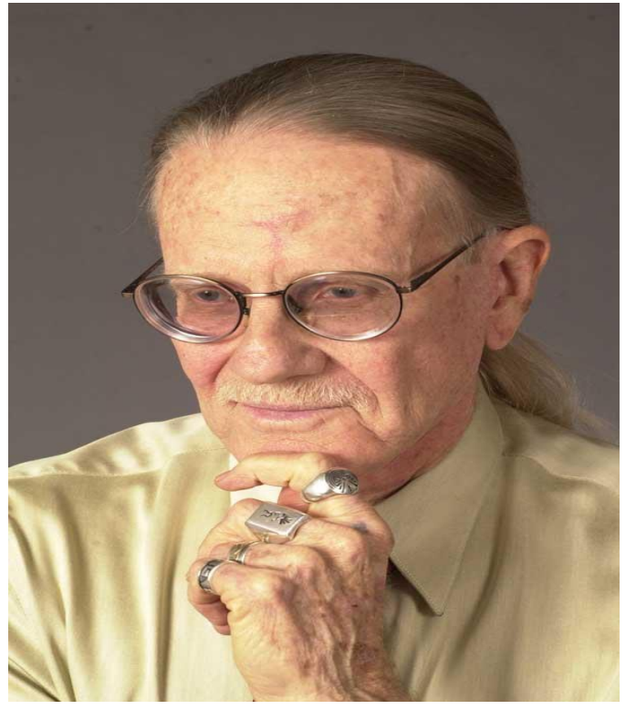
**1993**

# Short History of Experimental Research

**Nobel prize for Vernon Smith**

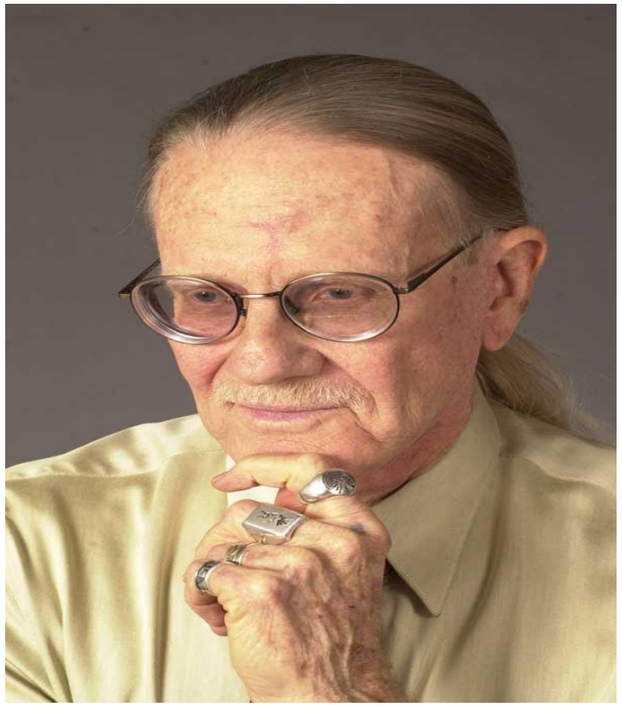
“for having established laboratory experiments as a tool in empirical economic analysis, especially in the study of alternative market mechanisms”

**2002**



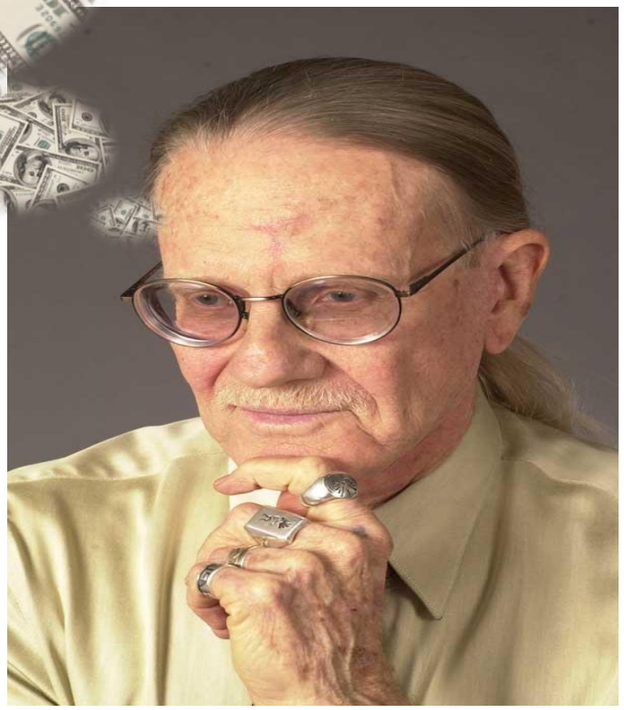
# Induced Value Theory

“The key idea of induced value theory is that proper use of a reward medium allows an experimenter to induce pre-specified characteristics, and the subjects innate characteristics become largely irrelevant. ”
(Vernon Smith, 1976)

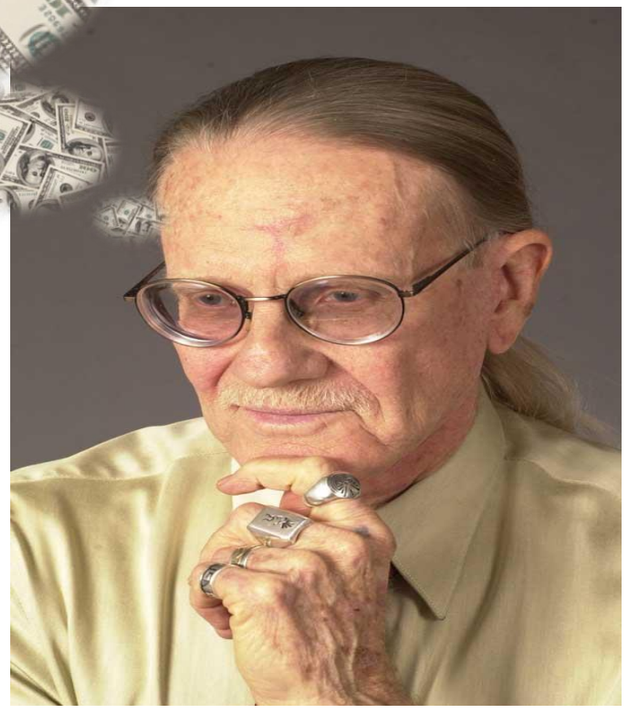


# Induced Value Theory

“The key idea of induced value theory is that proper use
medium allows an experimenter to induce pre-specified chara
and the subjects innate characteristics become largely irrelevant. ”
(Vernon Smith, 1976)



# Induced Value Theory



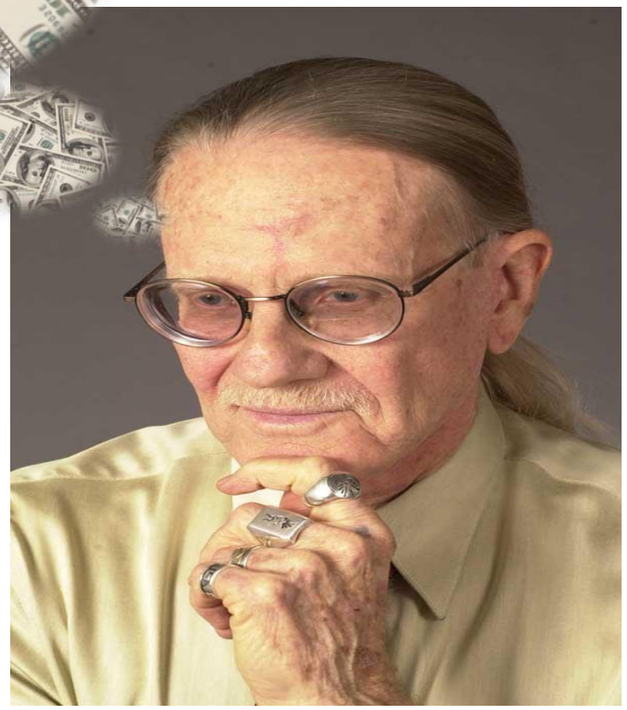
“The key idea of induced value theory is that proper use
medium allows an experimenter to induce pre-specified chara
and the subjects innate characteristics become largely irrelevant. ”
(Vernon Smith, 1976)

**Monotonicity**: Participants always prefer more of the reward medium (*m*)

**Salience**: Changes in *m* are directly related to a participant’s actions

**Dominance**: changes in a subject’s utility from participating in the experiment are primarily due to changes in m.

# Induced Value Theory



"The key idea of induced value theory is that proper use of a reward medium allows an experimenter to induce pre-specified characteristics, and the subjects innate characteristics become largely irrelevant. " (Vernon Smith, 1976)

**Monotonicity**: Participants always prefer more of the reward medium (*m*)

**Salience**: Changes in *m* are directly related to a participant's actions

**Dominance**: changes in a subject's utility from participating in the experiment are primarily due to changes in m.

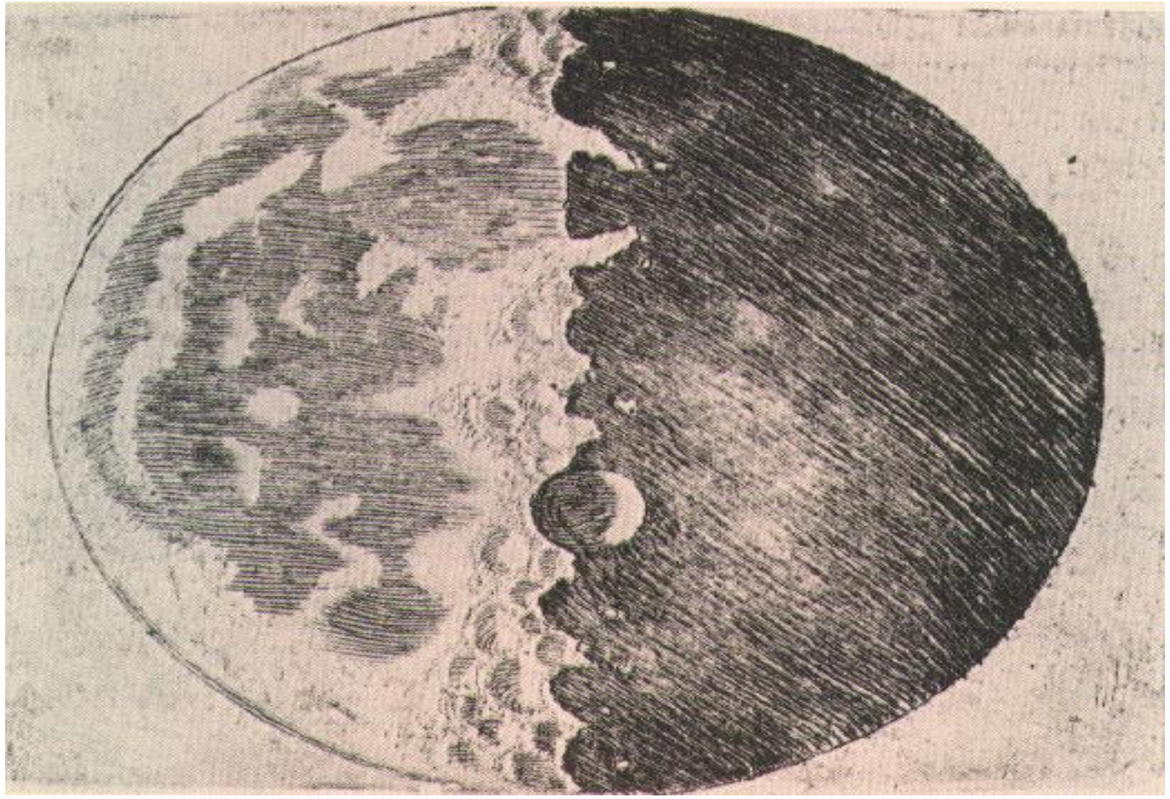
**Can you think of examples where failing to induce values can affect experimental results?**

# External Validity

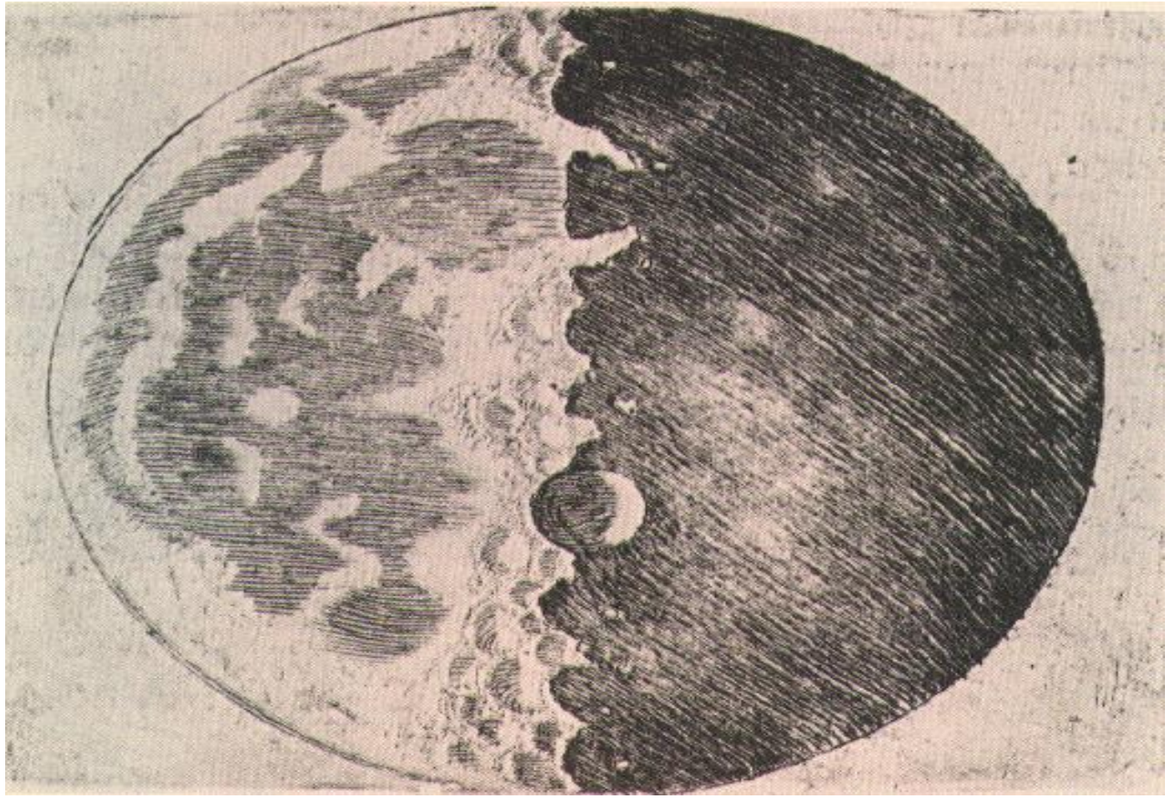
# External Validity

The extent to which the results of a study can be generalized or extended to others.

# External Validity



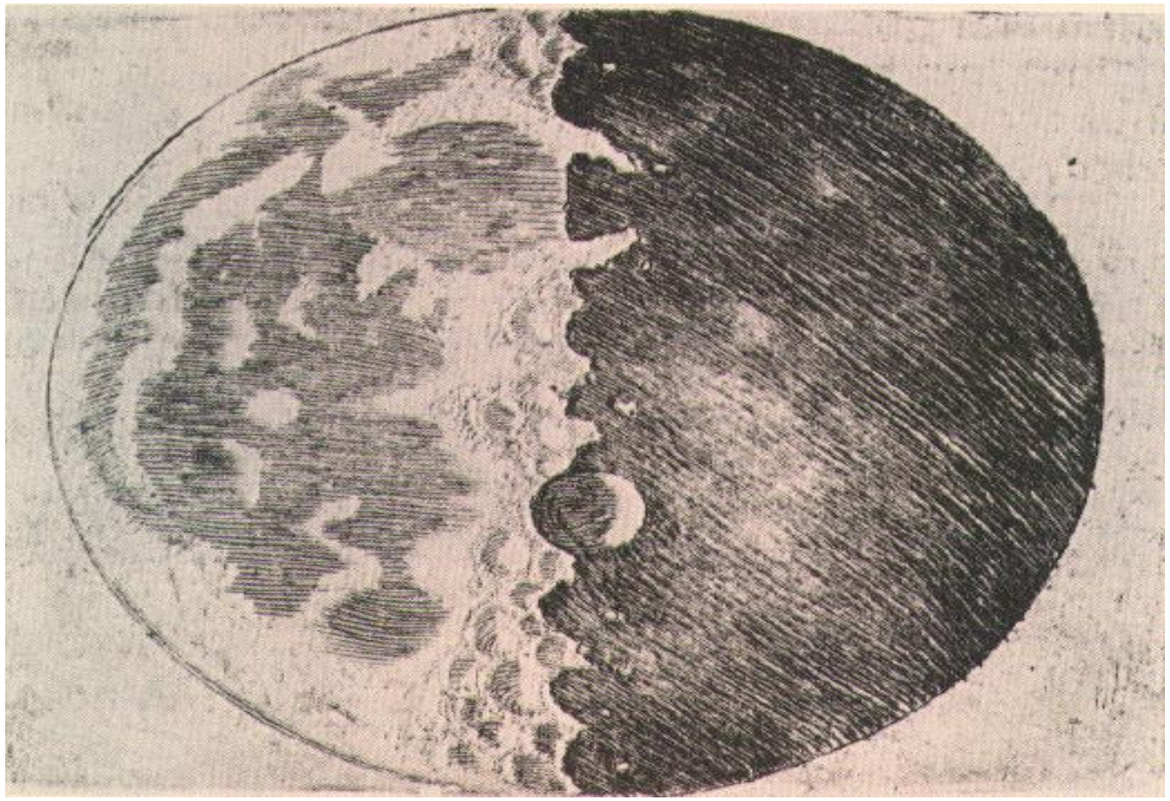
# External Validity



Galileo's hand-made drawing of the moon surface. Sidereus Nuncius (1610)

# External Validity

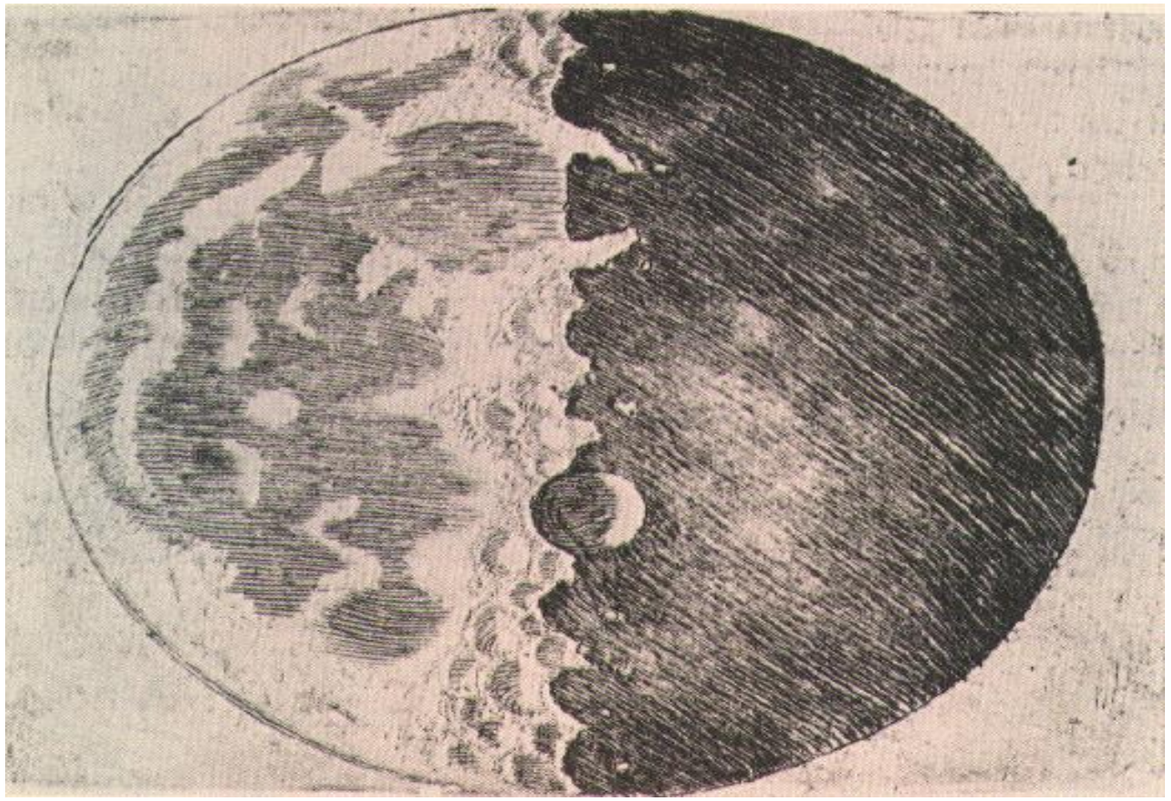
**Telescope:**
From Land to Sky

Galileo's hand-made drawing of the moon surface. Sidereus Nuncius (1610)

# External Validity

**Telescope:**
From Land to Sky

**Experiments:**
From Lab to Real Life

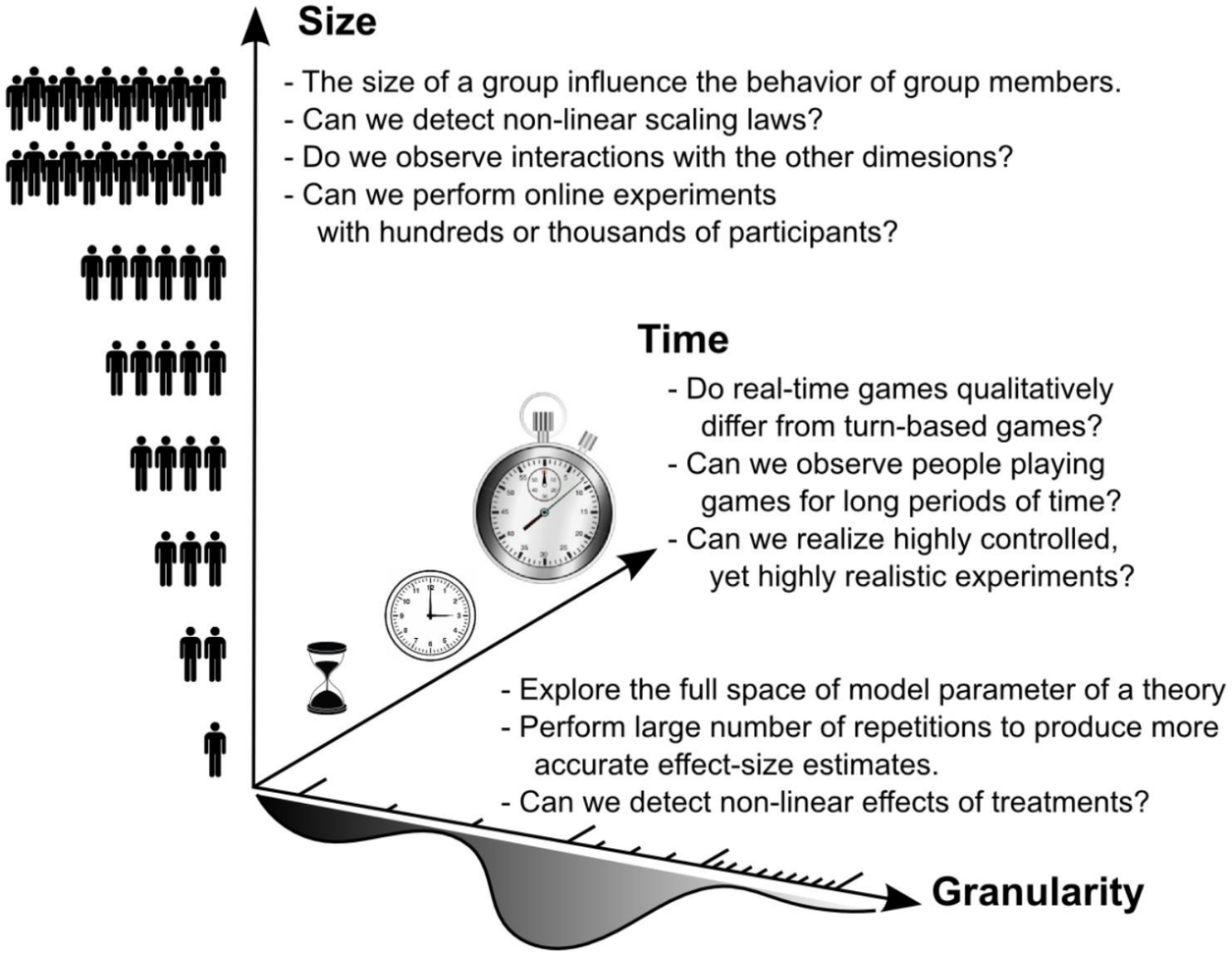
Galileo's hand-made drawing of the moon surface. Sidereus Nuncius (1610)

# Some Limitations of Traditional Experimental Research

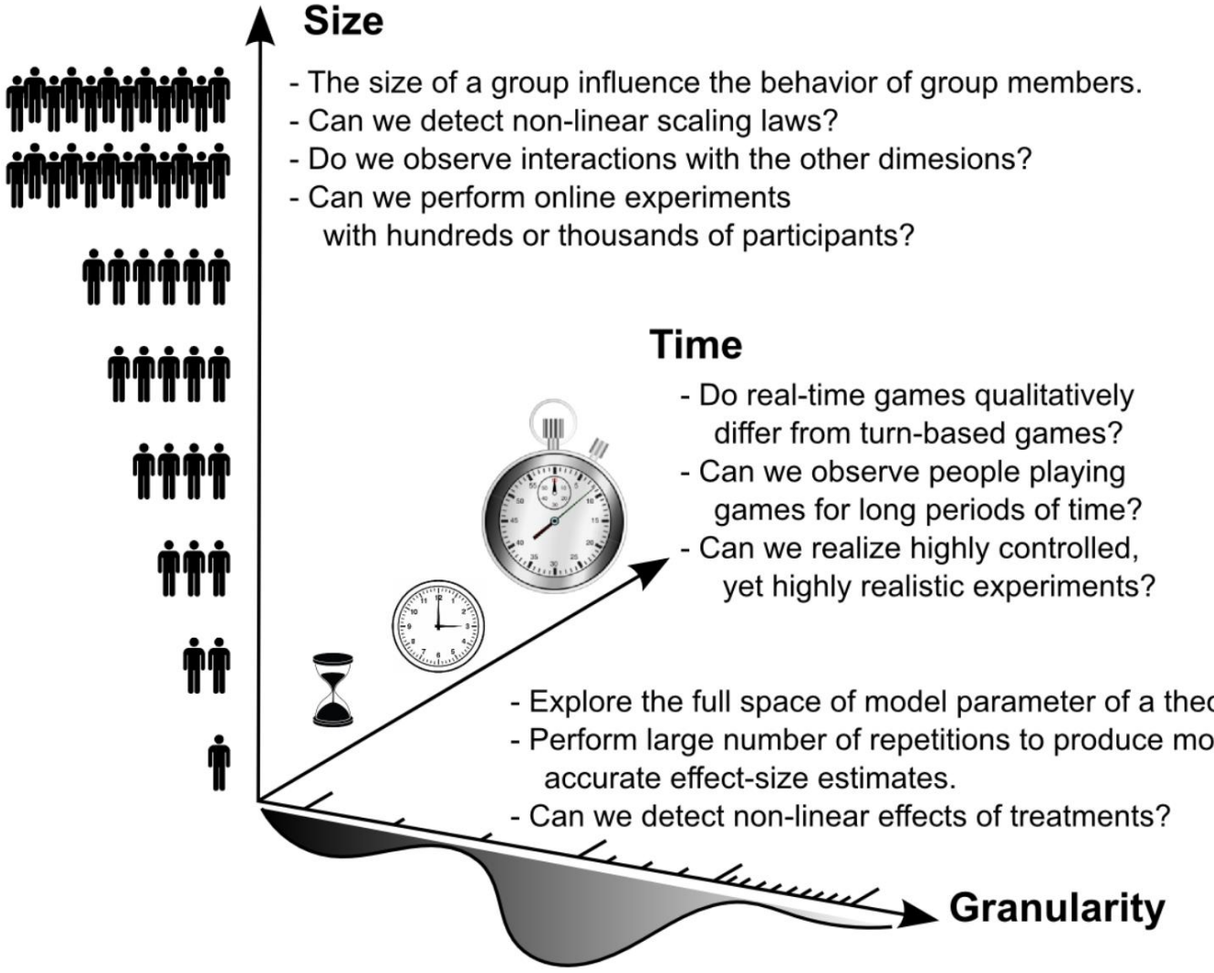
# Some Limitations of Traditional Experimental Research

- *Small groups* of individuals, generally 4-12 (bounded by size uni. Lab)
- Behavior of individuals in *social networks* under-researched
- Limited in the number of *hypotheses* that can be tested
- Observations are limited in *time*
- Generally *high costs* for an academic budget
- Limited *external validity:*
  - Sample WEIRD student
  - Artificial conditions

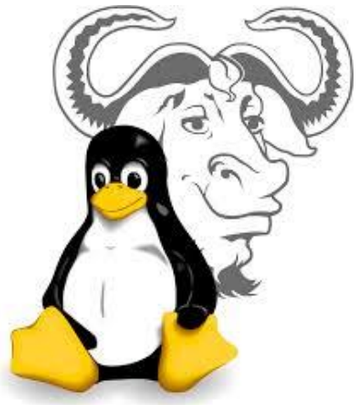
# Goals for Present and Future Behavioral Research



# Goals for Present and Future Behavioral Research



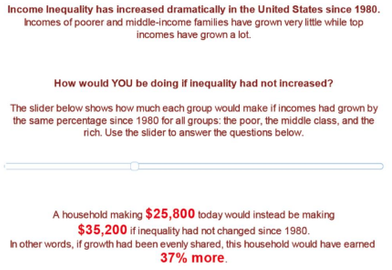


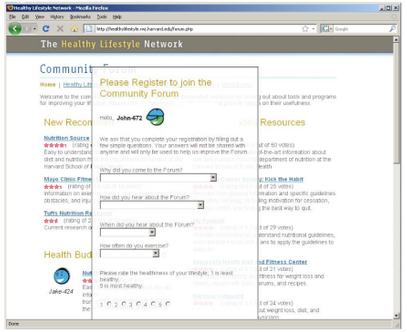



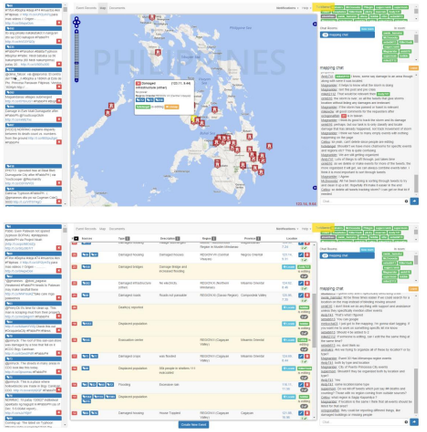
express

# Examples of Online Experiments

- Survey Experiments
- Game-Based Asynchronous Experiments
- Game-Based Synchronous Experiments



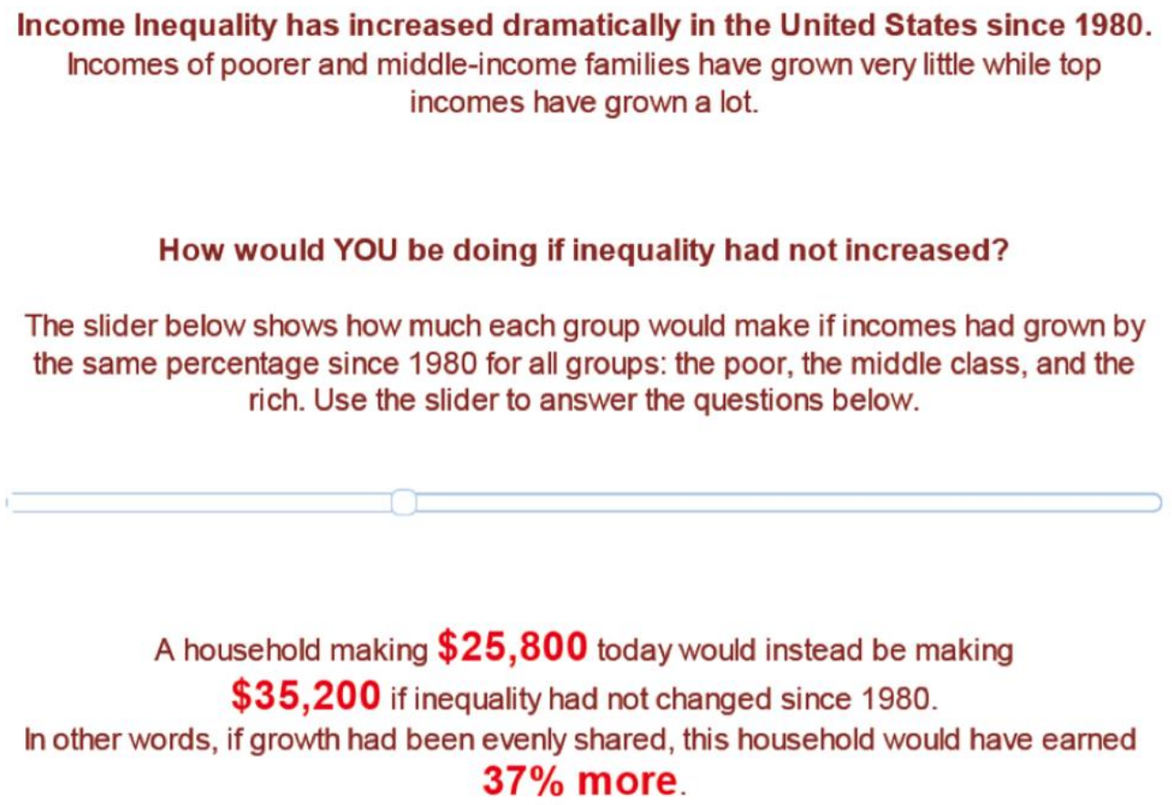
# Examples of Online Experiments

## Survey Experiments

Insert
- a new block of questions, or
- provide additional info

Measure preferences for outcome variable at the *individual* level

Relatively simple to implement




Kuziemko et al. (2015) How Elastic Are Preferences for Redistribution

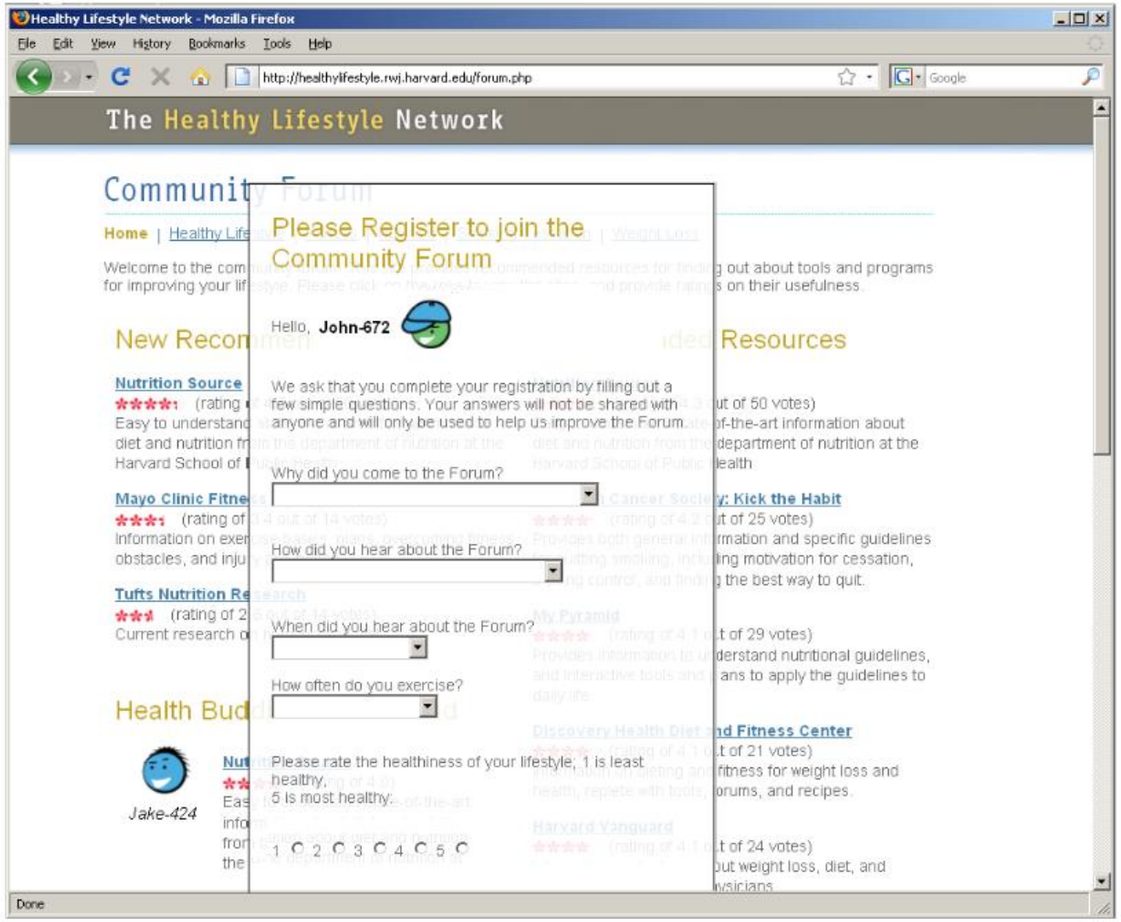
# Examples of Online Experiments

## Game-Based Asynchronous Experiments

Create a common environment
where a single variable is changed (e.g network structure)

Observe variation on outcome variable at the *group* level

Generally complex design
Can extend for long periods of time




Centola (2010) The Spread of Behavior in Online Social Networks

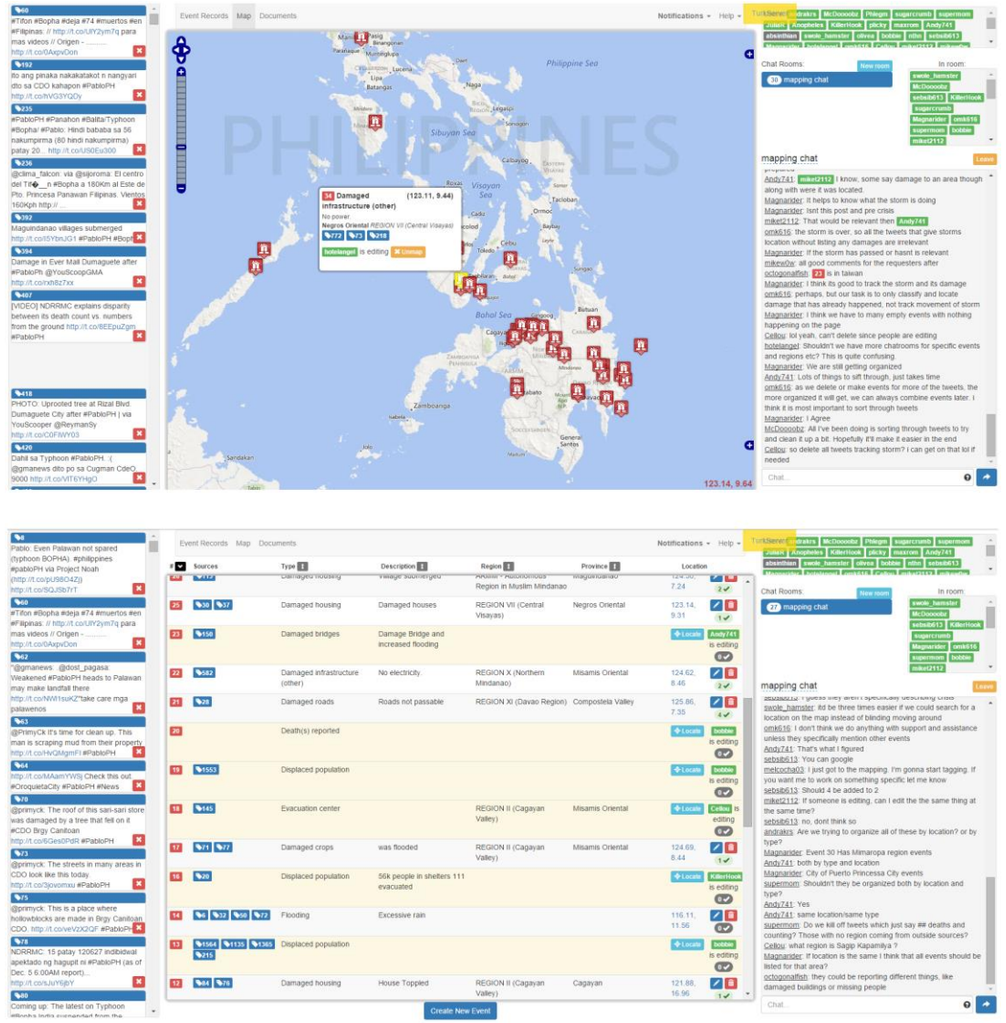
# Examples of Online Experiments

## Game-Based Synchronous Experiments

Create a common environment
where a single variable is changed
(e.g., group size or payoff)

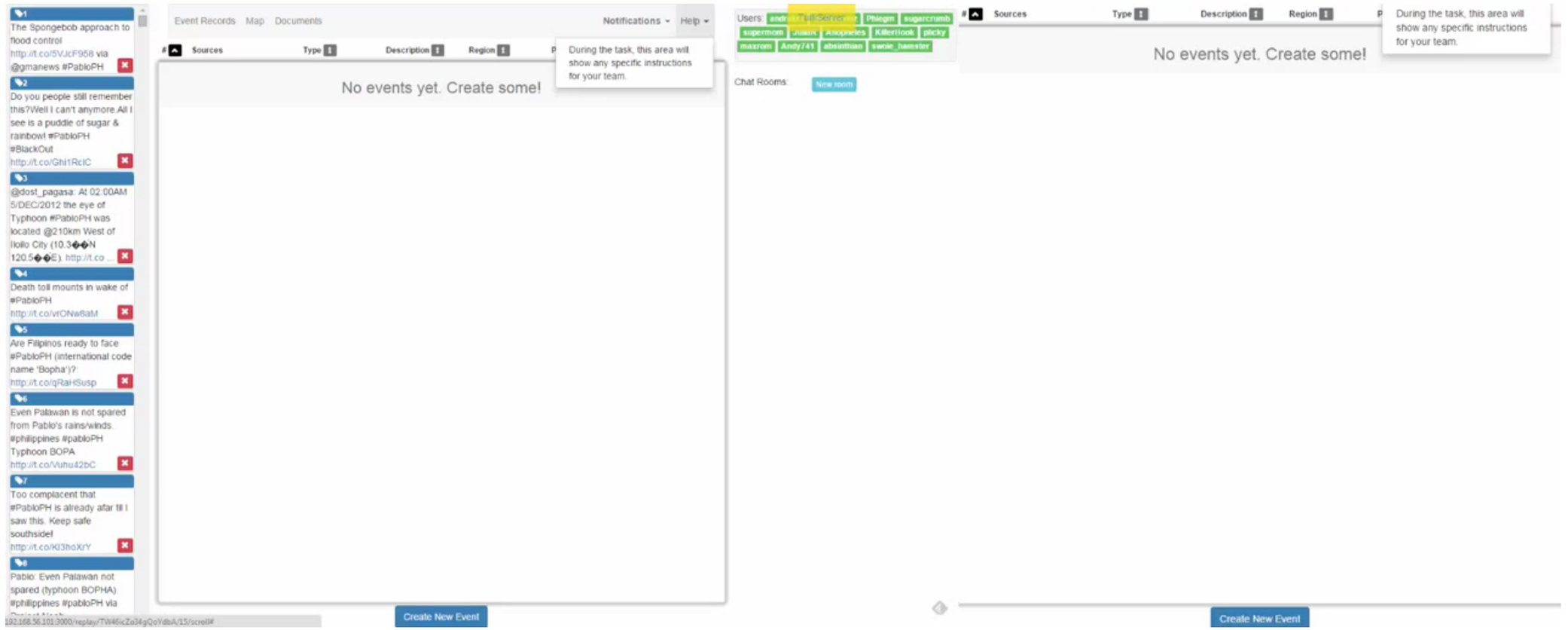
Observe variation on outcome variable at the *group* level

Can be really complex or stylized



Mao et al. (2016) An Experimental Study of Team Size and Performance on a Complex Task

# Examples of Online Experiments




Mao et al. (2016) An Experimental Study of Team Size and Performance on a Complex Task

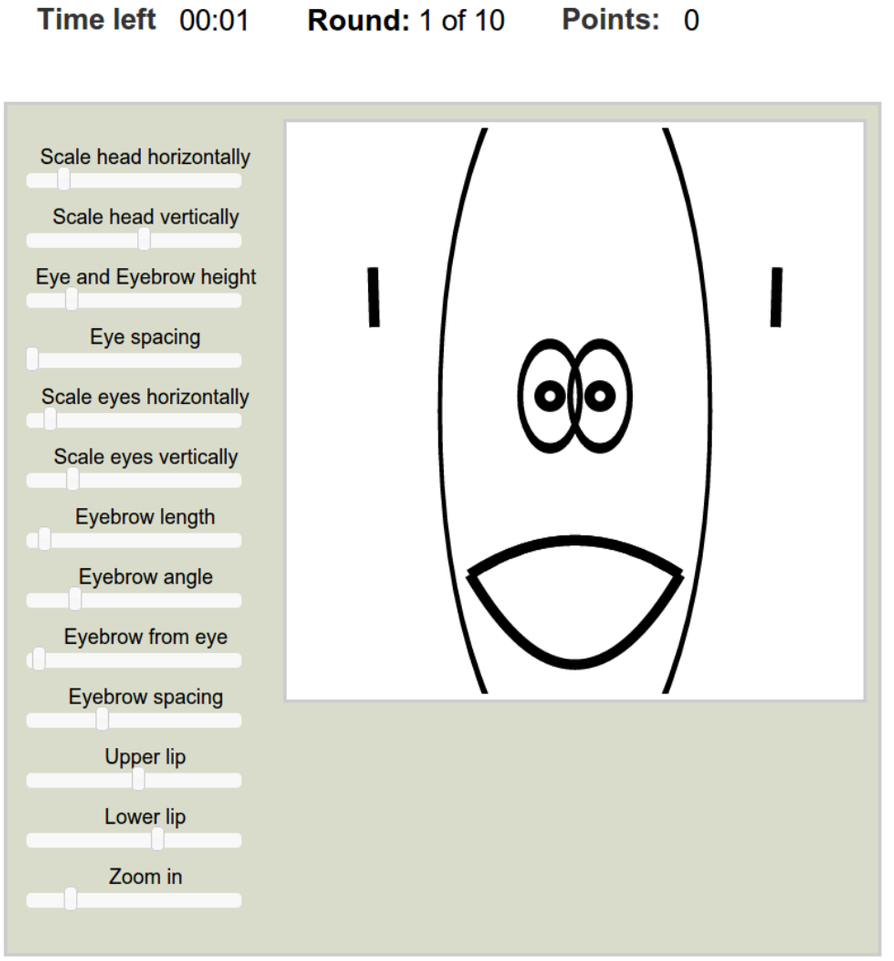
# Examples of Online Experiments

## Game-Based Synchronous Experiments

Create a common environment
where a single variable is changed
(e.g., group size or payoff)

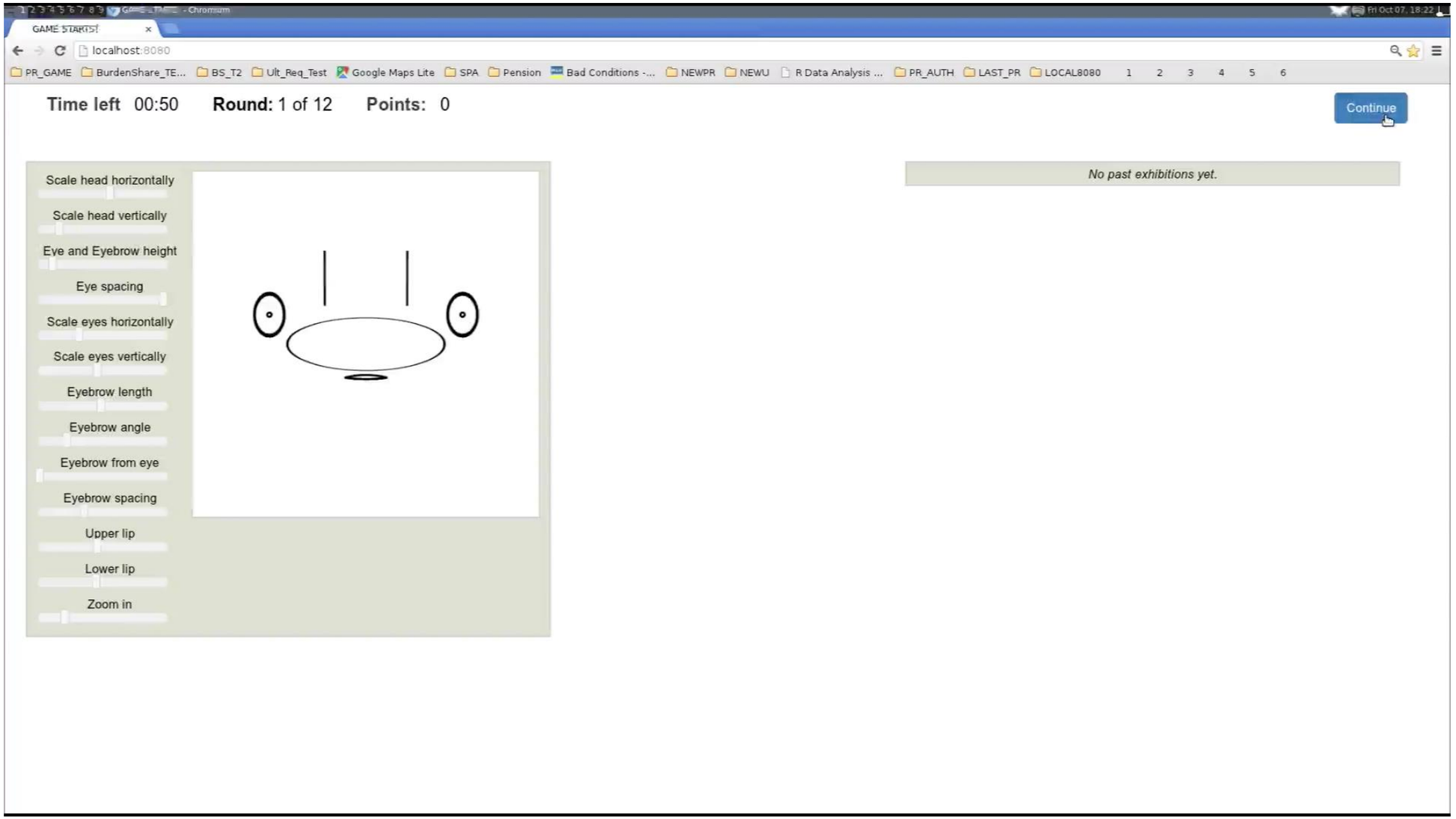
Observe variation on outcome variable at the *group* level

Can be really complex or stylized



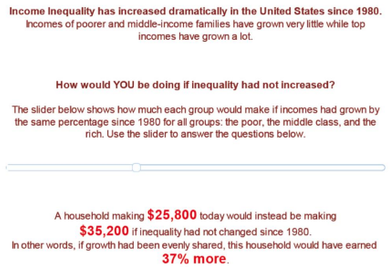

Balietti, Goldstone and Helbing (2016) “Peer Review and Competition in the Art Exhibition Game” PNAS 113(30) pp. 8414–8419

# Art Exhibition Game

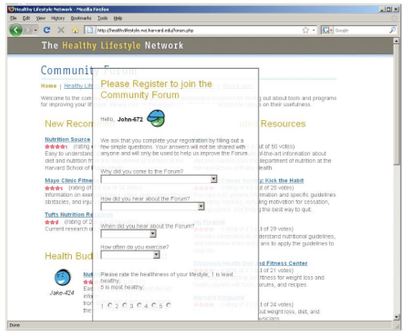



Balietti, Goldstone and Helbing (2016) “Peer Review and Competition in the Art Exhibition Game” PNAS 113(30) pp. 8414–8419

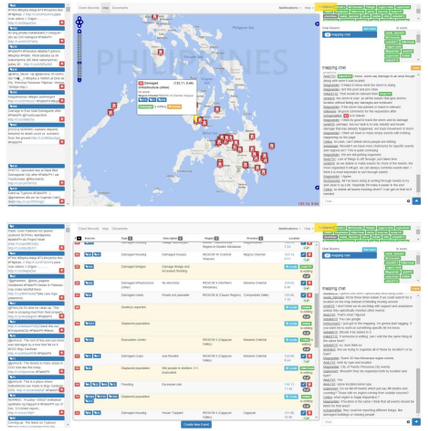
# Examples of Online Experiments




Survey Experiments




Game-Based Asynchronous Experiments



Game-Based Synchronous Experiments

COMPLEXITY OF EXECUTION

# Challenges in Online Experiments

- Recruitment
- Payment

- Anonymity

## Dropouts

- Fairness

- Interference

  - Data Quality
  - Farming
  - Cheating

- Technical Barriers for implementation
- Cross-Device Compatibility
- Accessibility (impaired users)

# Challenges in Online Experiments

- Recruitment
- Payment

- Anonymity

## Dropouts

- Fairness

- Interference
  - Data Quality
  - Farming
  - Cheating

- Technical Barriers for implementation
- Cross-Device Compatibility
- Accessibility (impaired users)

# Recruiting Platforms

- Amazon Mechanical Turk
  https://www.mturk.com
- TurkPrime
  https://www.turkprime.com
- Figure Eight (ex Crowd Flower, ML)
  https://www.figure-eight.com
- Prolific.ac
  https://www.prolific.ac
- Reddit (for bursty access)
  https://www.reddit.com/r/WebGames
- YouGov (country-specific)
  https://yougov.co.uk
- PollFish (also built-in surveys)
  https://www.pollfish.com
- Psychological Research on the Net
  http://psych.hanover.edu/research/exponnet.html
- The Web Experiment List
  http://www.wexlist.net
- Online Social Psychology Studies
  http://www.socialpsychology.org/expts.htm

# Citizen Science Initiatives

- Volunteer Science
  https://volunteerscience.com
- Science@Home
  https://www.scienceathome.org
- Zoouniverse
  https://www.zooniverse.org
- Lab in the Wild
  https://labinthewild.org
- CitizenLab
  https://www.citizenlab.co

# Challenges in Online Experiments

- Recruitment
- Payment

- Anonymity

## Dropouts

- Fairness

- Interference

- Data Quality
- Farming
- Cheating

- Technical Barriers for implementation
- Cross-Device Compatibility
- Accessibility (impaired users)

# Challenges in Online Experiments

- Recruitment
- Payment

- Anonymity

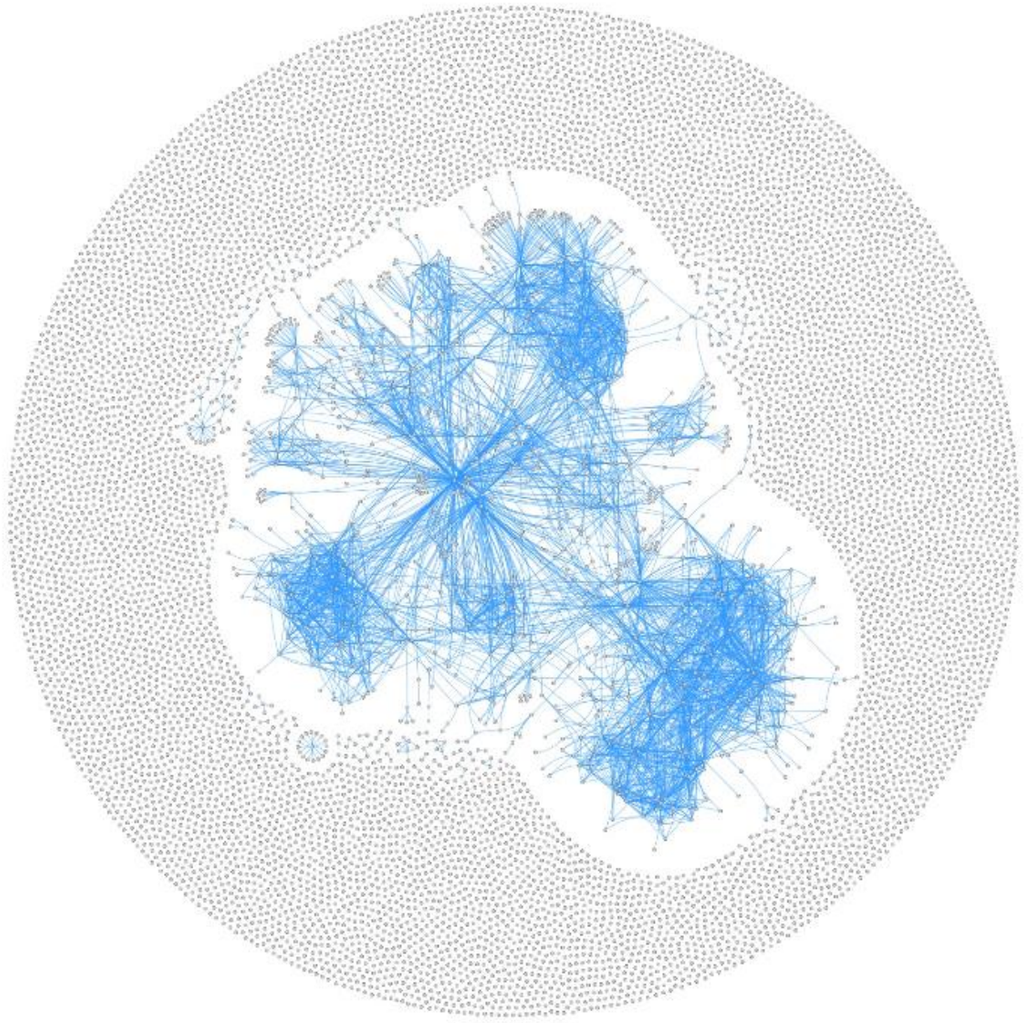
## Dropouts

- Fairness

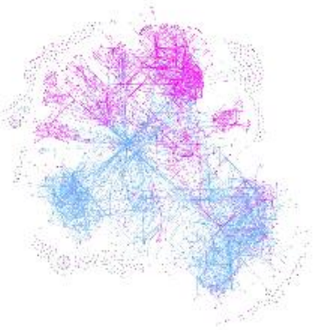
- <u>Interference</u>

- Data Quality
- Farming
- Cheating

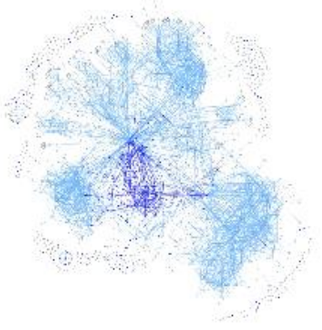
- Technical Barriers for implementation
- Cross-Device Compatibility
- Accessibility (impaired users)



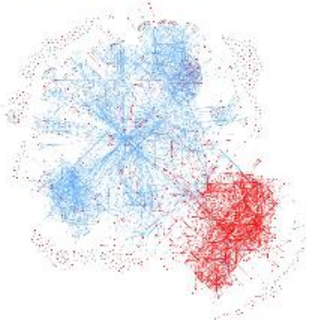
(a) The communication network



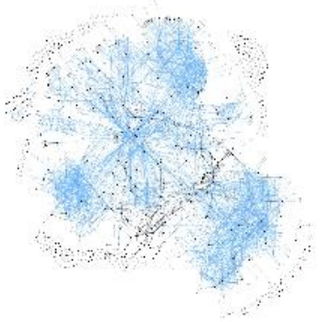
(b) Reddit HWTF



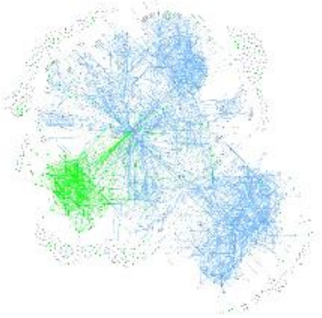
(e) Facebook



(c) MTurkGrind



(f) MTurkForum



(d) TurkerNation

Figure 1: 1a: The communication network among Amazon Mechanical Turk workers. 1b-1f: Subnetworks for Reddit HWTF (magenta; 660 workers, 1837 edges), MTurkGrind (red; 392 workers, 1331 edges), TurkerNation (green; 200 workers, 740 edges), Facebook (blue; 133 workers, 357 edges), and MTurkForum (black; 312 workers, 244 edges).

Ming et al. (2016) “The Communication Network Within the Crowd”

# Challenges in Online Experiments

- Recruitment
- Payment

- Anonymity

## Dropouts

- Fairness

- Interference

  - Data Quality
  - Farming
  - Cheating

- Technical Barriers for implementation
- Cross-Device Compatibility
- Accessibility (impaired users)

# Challenges in Online Experiments

- Recruitment
- Payment

- Anonymity

## Dropouts

- Fairness

- Interference
  - Data Quality
  - Farming
  - Cheating

- Technical Barriers for implementation
- Cross-Device Compatibility
- Accessibility (impaired users)

# Challenges in Online Experiments

- Anonymity might increase deviant behavior
- (but get access to specialized populations)

- The same person uses the same computer to participate repeatedly.
- The same person uses different computers to participate repeatedly.
- Different persons use the same computer to participate.

# Challenges in Online Experiments

- Recruitment
- Payment

- Anonymity

## Dropouts

- Fairness

- Interference

  - Data Quality
  - Farming
  - Cheating

- Technical Barriers for implementation
- Cross-Device Compatibility
- Accessibility (impaired users)

# Challenges in Online Experiments

- Recruitment
- Payment

- Anonymity

## Dropouts

- Fairness

- Interference

  - Data Quality
  - Farming
  - Cheating

- Technical Barriers for implementation
- Cross-Device Compatibility
- Accessibility (impaired users)

# Improve Data Quality and Attention

***Content responsive faking***:

(a) responses influenced by the item content but (b) are not completely accurate.

purposeful faking (e.g., malingering in medical surveys)
socially desirable response of both intentional and nonintentional varieties

**Content nonresponsivity:**

responding without regard to item content.

Large variance estimation (2.5 – 50%; 10.6 for students)
Inattentive only on some questions

Meade, A. W. and S. B. Craig (2012). Identifying Careless Responses in Survey Data. Psychological methods 17(3), 437.

# Improve Data Quality and Attention

- **Engagement is key**
  ui design, interactive elements, colors, length, feedback…
- Nonsensical or “bogus” items
  “30th February”
- Instructed response items
  “To monitor quality, please respond with a two for this item”
- Unlikely sets
  “Select all the good qualities you possess”
- Consistency indexes
  Evaluating  differences among responses to items highly similar in content
- Too quick response times and other obvious patterns in responses

Meade, A. W. and S. B. Craig (2012). Identifying Careless Responses in Survey Data. Psychological methods 17(3), 437.

# Commitment Device

- Ask participants if they paid attention to the *previous* set of questions
- It prompts participants to pay attention to the *subsequent set of* questions
- Its purpose is fulfilled regardless of whether the respondents answer honestly

# Commitment Device Example

Alesina, Miano, and Stancheva (2018) “Immigration and Redistribution” NBER 24733

Before proceeding to the next set of questions, we want to ask for your feedback about the responses you provided so far. It is vital to our study that we only include responses from people who devoted their full attention to this study. This will not affect in any way the payment you will receive for taking this survey. **In your honest opinion, should we use your responses, or should we discard your responses since you did not devote your full attention to the questions so far?**

- Yes, I have devoted full attention to the questions so far and I think you should use my responses for your study
- No, I have not devoted full attention to the questions so far and I think you should not use my responses for your study

# AMT Bot Scare

**Dennis et al. (2018)**
“MTurk Workers' Use of Low-Cost 'Virtual Private Servers' to Circumvent Screening Methods: A Research Note”
https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3233954

**Ahler et al. (2018)**
The Micro-Task Market for “Lemons”: Collecting Data on Amazon’s Mechanical Turk
http://gsood.com/research/papers/turk.pdf

**TurkPrime Blog**
https://blog.turkprime.com/after-the-bot-scare-understanding-whats-been-happening-with-data-collection-on-mturk-and-how-to-stop-it
https://blog.turkprime.com/moving-beyond-bots-mturk-as-a-source-of-high-quality-data

## Tools to filter Farmers’ IP addresses

https://github.com/MAHDLab/rIP

https://itaysisso.shinyapps.io/Bots

# Challenges in Online Experiments

- Recruitment
- Payment

- Anonymity

## Dropouts

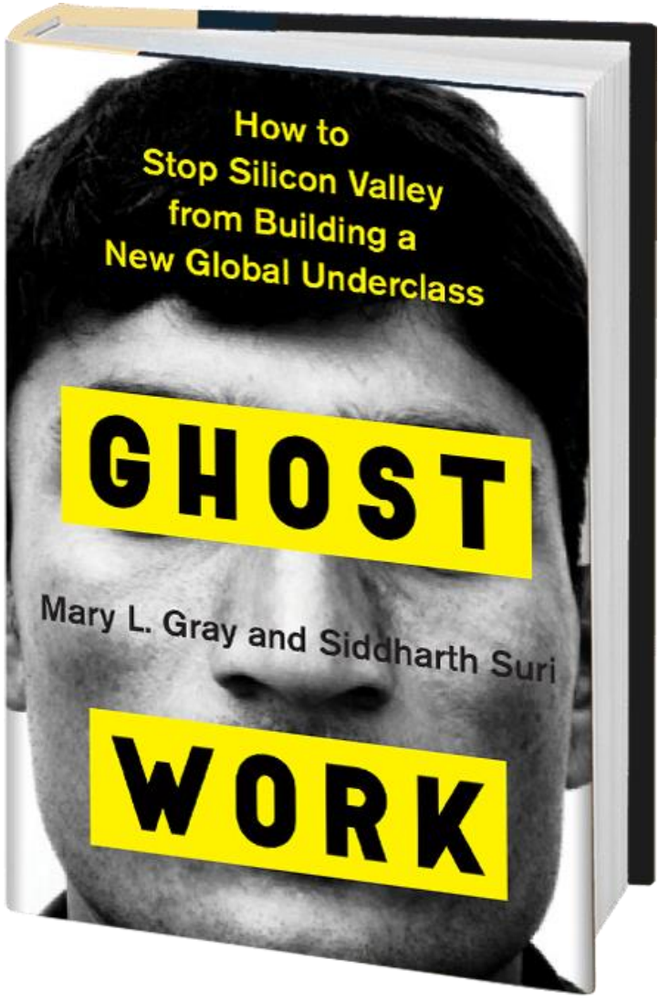
- Interference
  - Data Quality
  - Farming
  - Cheating

- Fairness

- Technical Barriers for implementation
- Cross-Device Compatibility
- Accessibility (impaired users)

# Fairness for Crowd Workers

- IRB, IRB, IRB (Institutional Review Board)
- Crowd worker have flexibility, but almost no rights
- Low pay, menial tasks, no accrual of sick days, or contributions to pension, no unions, etc.
- In the US, 20 million people (Pew Research)
- Primary income for:
  - 1.25-2.5 million (Economic Policy Institute, 2018)
  - 2-10 million (JPMorgan Chase Institute, 2017)




Mary Gray & Sid Suri

http://www.ghostwork.info
http://www.inthecrowd.org

# Challenges in Online Experiments

- Recruitment
- Payment

- Anonymity

## Dropouts

- Fairness

- Interference

  - Data Quality
  - Farming
  - Cheating

- Technical Barriers for implementation
- Cross-Device Compatibility
- Accessibility (impaired users)

# Challenges in Online Experiments

- Recruitment
- Payment

- Anonymity

## Dropouts

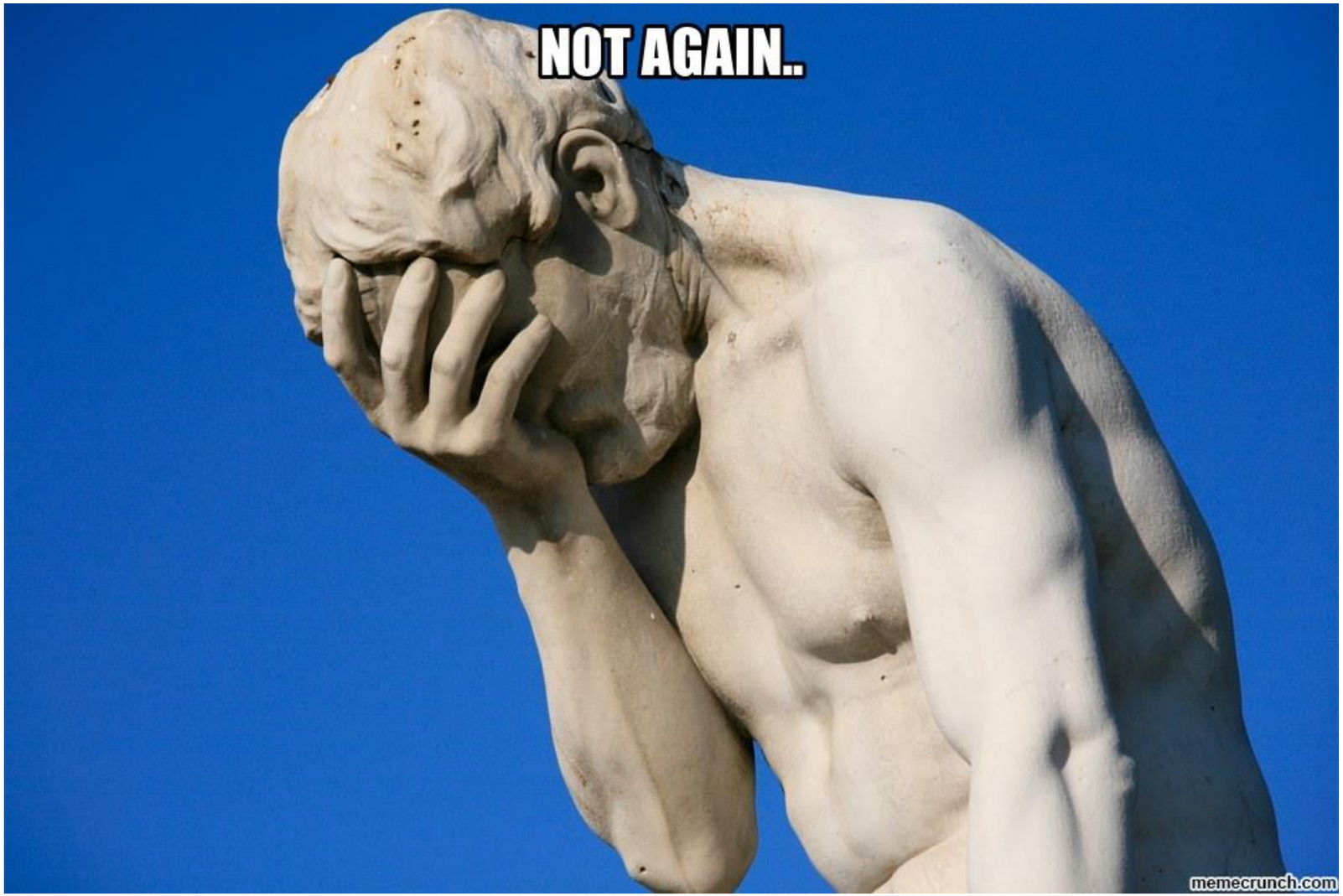
- Fairness

- Interference

- Data Quality
- Farming
- Cheating

- Technical Barriers for implementation
- Cross-Device Compatibility
- Accessibility (impaired users)

# Dropouts



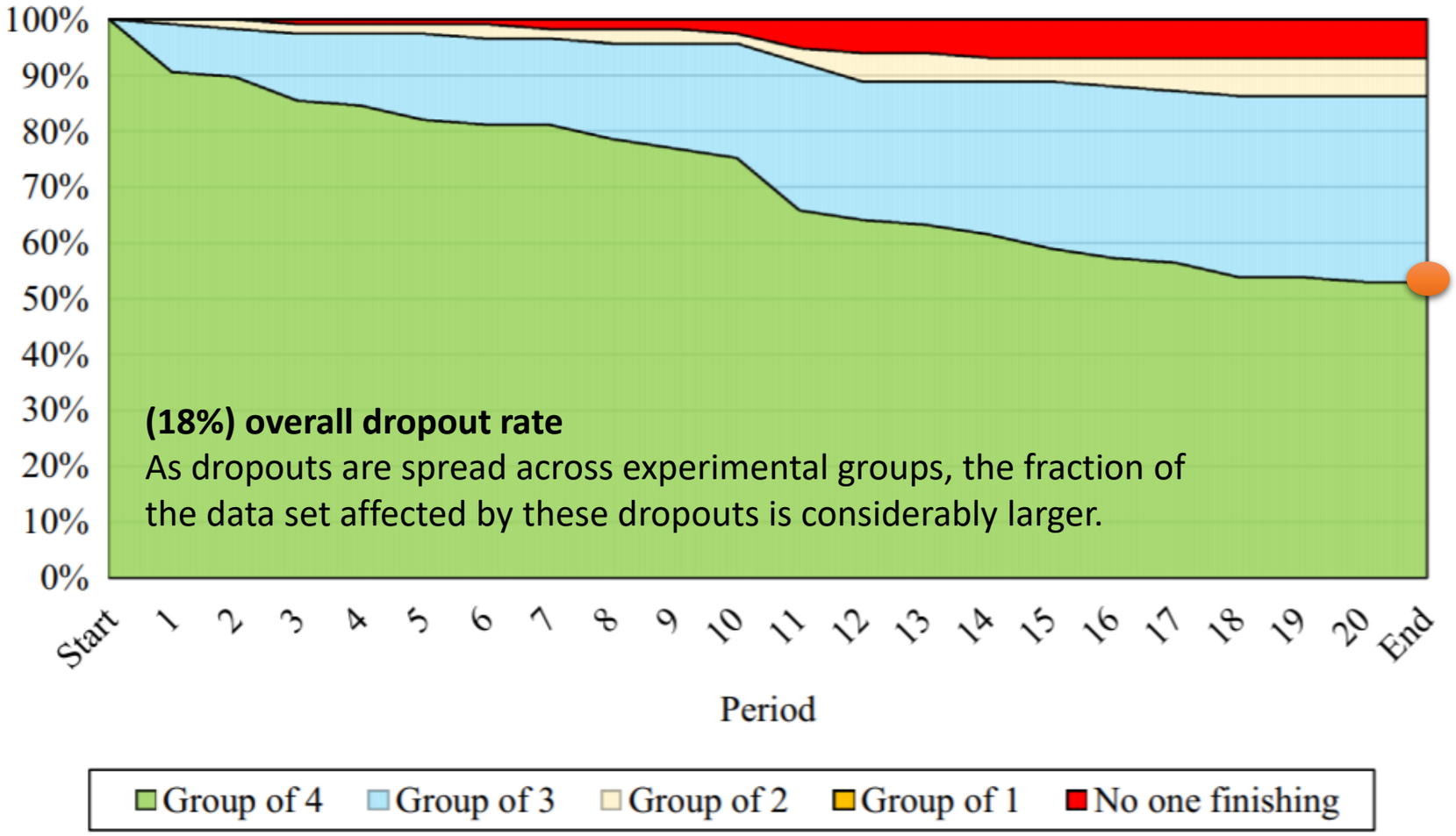
# Dropouts (Attrition)

Arechar et al. (2018) “Conducting interactive experiments online”

**“The most severe problem for online interactive studies, and the largest discrepancy with laboratory experiments, is attrition (participant dropout).”**

Major challenge to internal validity, if dropout rates vary with treatment, selection bias may arise

# Dropouts (Attrition)




Just a little bit more than 50% of the groups that started finished without dropouts

Arechar et al. (2018) “Conducting interactive experiments online”

# What Keeps Participants In?

**Intrinsic Motivation**
Volunteer Science, Galaxy Zoo, Science@Home (Deci, 2005)

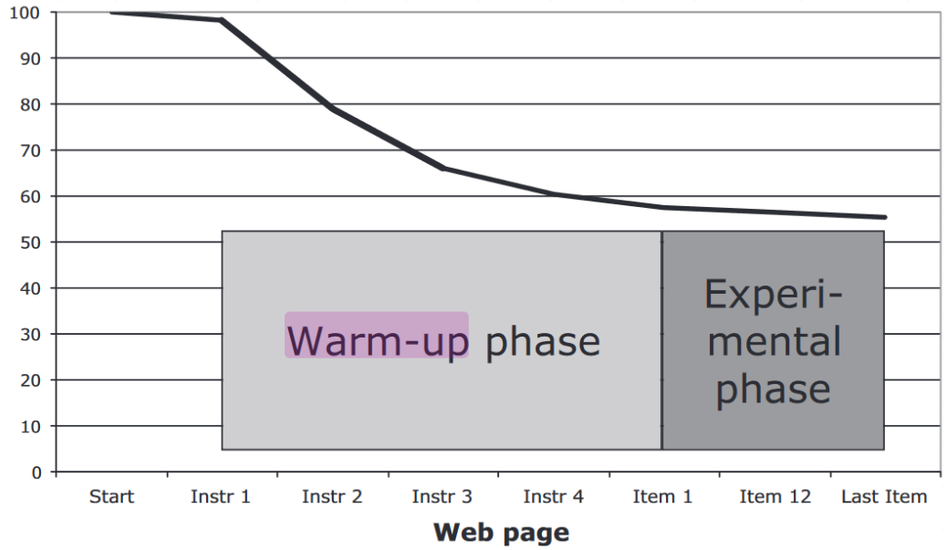
**Monetary incentives**
about 75% of US Turkers reported that MTurk was their primary or secondary source of income (Paolacci et al, 2009)

**Curiosity**
Solving and unrelated, but intriguing puzzle (Law et al., 2016)

**Personalized Feedback**
Lab-in-the-Wild (Ye et al., 2017)

# Two Simple Techniques To Limit Dropouts

- Implement warm-up phase (nodeGame levels)




- Implement seriousness checks

Reips (2002) Standards for Internet-Based Experimenting

# What to Do When Dropouts Happen?

# What to Do When Dropouts Happen?

- Terminate experiment immediately
- Continue with less participants (notify? Does behavior change?)
- Replace with bots (notify? Behavior is different)
- Replace with other humans (difficult to implement technically and humanly)

# What to Do When Dropouts Happen?

- Terminate experiment immediately
- Continue with less participants (notify? Does behavior change?)
- Replace with bots (notify? Behavior is different)
- Replace with other humans (difficult to implement technically and humanly)

→ Try to design games that are **dropout-robust**, Make Leaving Harder, Inform about total time for experiment, Inform about waiting times during the experiment, Wait for reconnections, Notify

# Challenges in Online Experiments

- Recruitment
- Payment

- Anonymity

## Dropouts

- Fairness

- Interference

  - Data Quality
  - Farming
  - Cheating

- Technical Barriers for implementation
- Cross-Device Compatibility
- Accessibility (impaired users)

# List of Platforms for Online experiments

## Group-Behavior

- nodeGame ☺
  https://nodeGame.org
- Wextor (pioneer)
  https://www.wextor.eu
- Otree (large base)
  https://www.otree.org
- Lioness (new)
  https://lioness-lab.org
- Breadboard (networks)
  http://breadboard.yale.edu
- Empirica (very new)
  https://empirica.ly
- TurkServer (discontinued)
  https://turkserver.readthedocs.io

## Individual

- JSPsych (many plugins)
  https://www.jspsych.org
- PsiTurk (groups possible)
  https://psiturk.org

## Other Resources

- Browser Stack
  https://www.browserstack.com
- CSS Frameworks
  https://getbootstrap.com

# nodeGame: the Good Parts

- Powerful API to customize experiments
- Access to low-level details
- Integrated JS database
- Fast and highly scalable
- Game Levels
- Modular design (games, widgets, window)
- Well-documented (and active Forum)
- Integration with Amazon Mechanical Turk