MAS.S68, Spring 2023
3/8

Generative AI
for Constructive Communication

Evaluation and New Research Methods

center for
constructive
communication

Agenda

Mina Lee

Zoom talk

Q&A after her talk

center for
constructive
communication

Second half of class:

5 minute break

Feedback notes

Break after talks, assignments focused on project

Human Subjects Research

Lecture

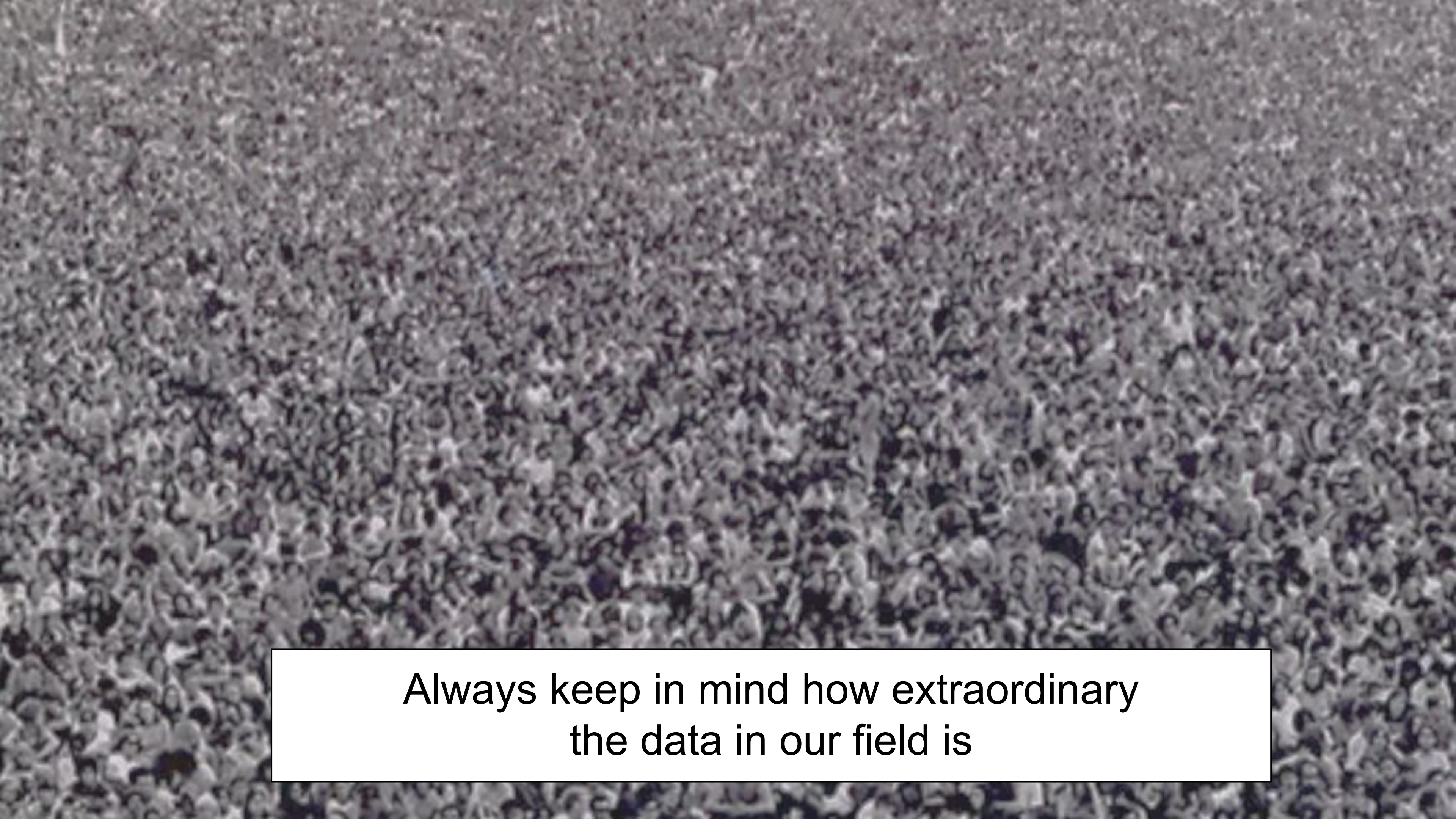
Human subject research for LM-based applications

Scope of this lecture

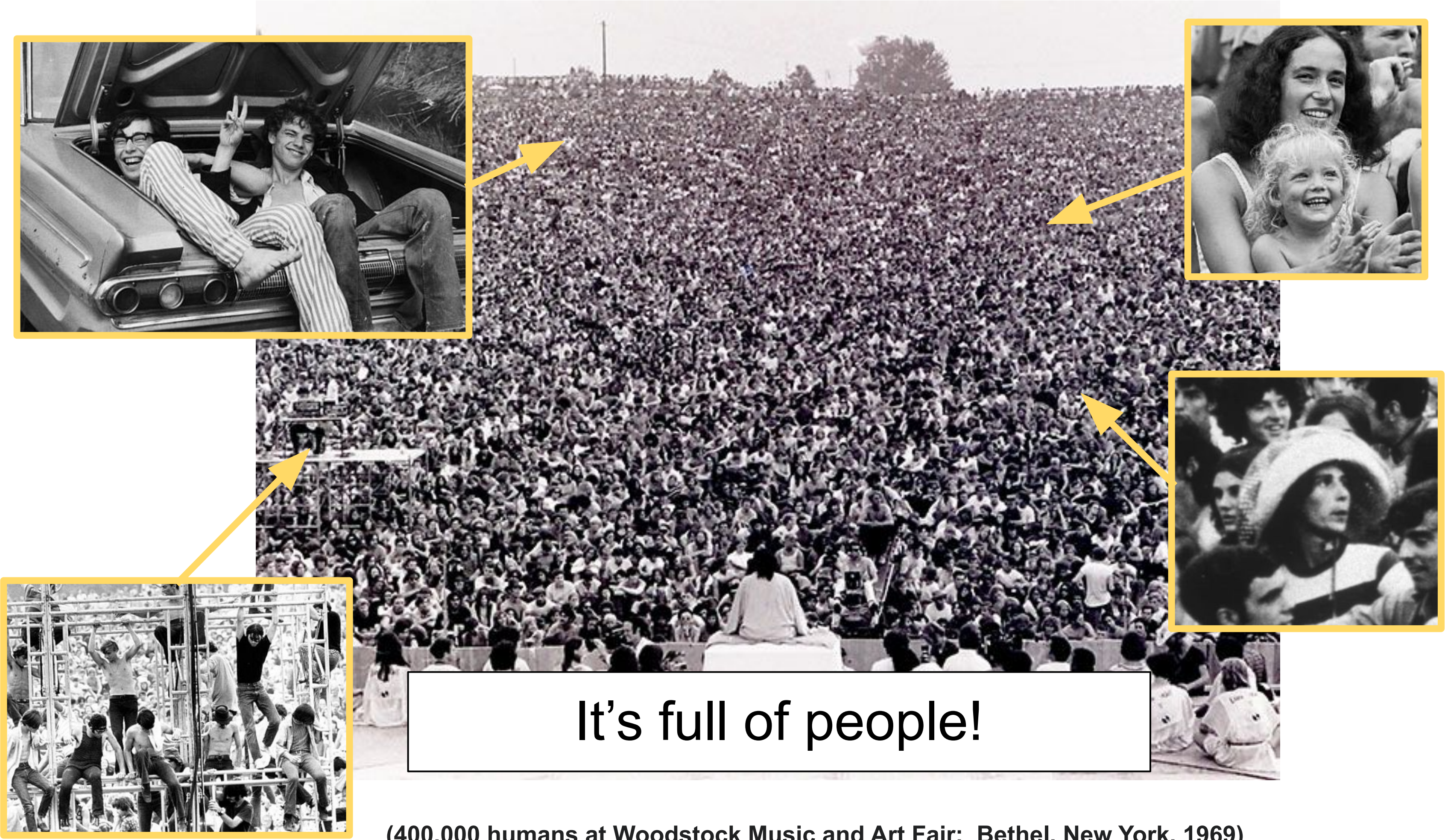
Our goal is to **summarize practical advice on measuring the impact of your work on people**. We'll cover many subjects, each of which can stand on its own as a whole course! There are links to more reading throughout this presentation.

Outline:

- [10 min] Motivation + examples
- [5 min] Ethics / IRB
- [5 min] Methods: Surveys
- [5 min] Methods: Interviews
- [10 min] Methods: Randomized control trials
- [5 min] Platforms and resources



Always keep in mind how extraordinary the data in our field is



It’s full of people!

(400,000 humans at Woodstock Music and Art Fair; Bethel, New York, 1969)

Every stage is full of people!

People write the words that LMs are trained on

People express the preferences that we use to fine-tune these LMs

People provide the labels that let us evaluate these LMs

People decide which applications to create

People use these applications

People are impacted by this usage

How do these people relate to each other, and how can we best serve them?

What is human subject research?

“Interventional”

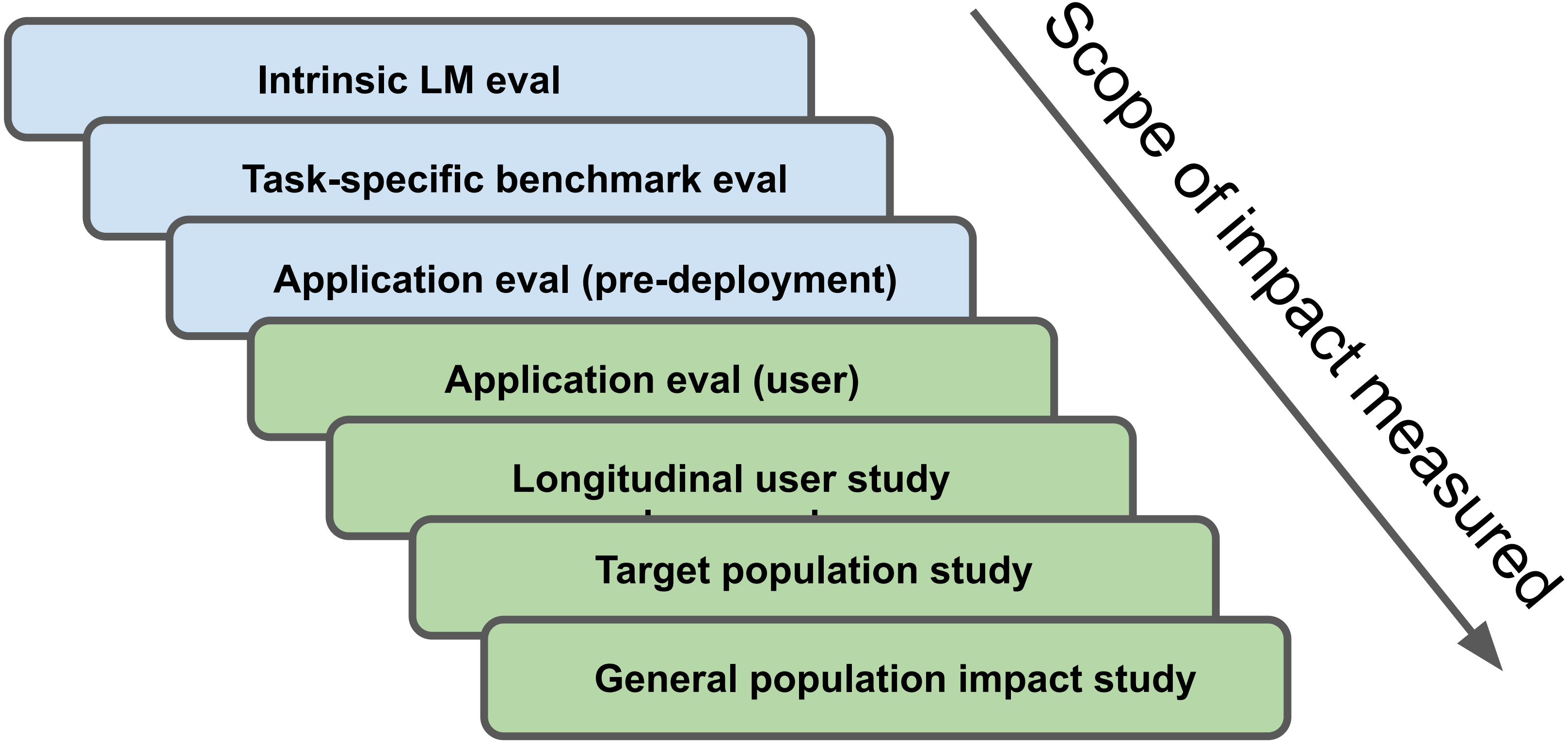
Studies in which real human beings are affected or observed.

“Observational”

What is human subject research?

Studies in which real human beings are affected or observed. They vary in **scope**, **method**, and **scale**.

Evaluation landscape for LM-based applications



Intrinsic LM eval

Task-specific benchmark eval

Application eval (pre-deployment)

Application eval (user)

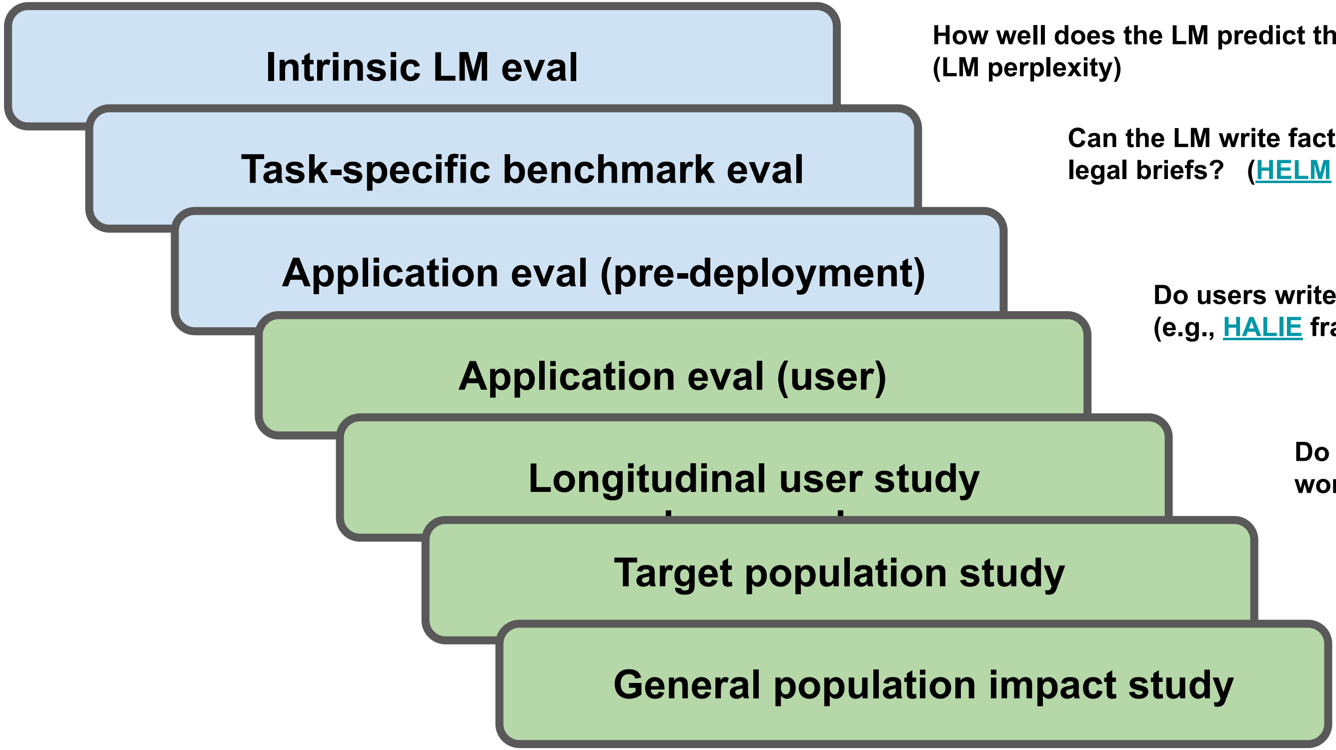
Longitudinal user study

Target population study

General population impact study

Scope of impact measured

Evaluation landscape: “Chatbot for lawyers” app example



Intrinsic LM eval

How well does the LM predict the next word?
(LM perplexity)

Task-specific benchmark eval

Can the LM write factual, fluent
legal briefs? (HELM metrics)

Application eval (pre-deployment)

Application eval (user)

Do users write good briefs with it?
(e.g., HALIE framework)

Longitudinal user study

Do lawyers adopt it into their
workflows?

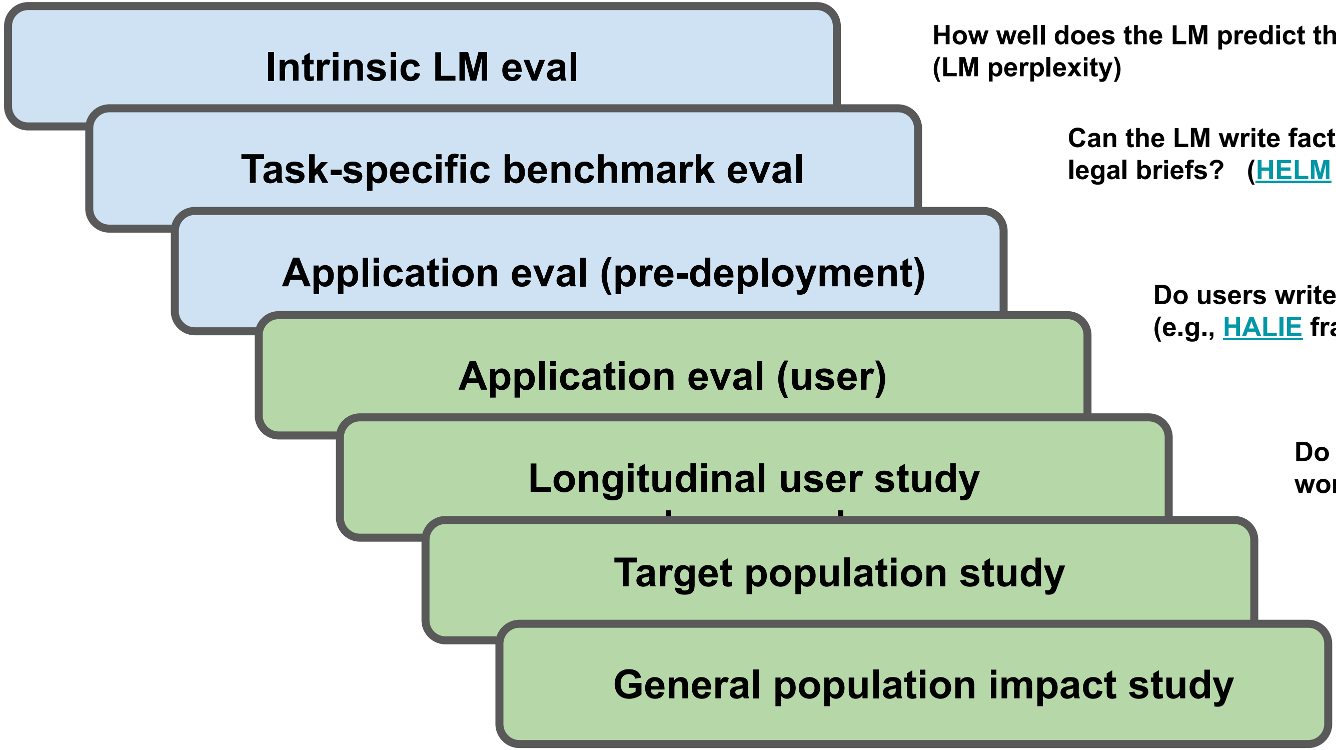
Target population study

Do law firms change their hiring?

General population impact study

Do people trust the justice system?

Evaluation landscape: “Chatbot for lawyers” app example



Intrinsic LM eval

How well does the LM predict the next word? (LM perplexity)

Task-specific benchmark eval

Can the LM write factual, fluent legal briefs? (HELM metrics)

Application eval (pre-deployment)

Application eval (user)

Do users write good briefs with it? (e.g., HALIE framework)

Longitudinal user study

Do lawyers adopt it into their workflows?

Target population study

Do law firms change their hiring?

General population impact study

Do people trust the justice system?

Recent examples of LM-oriented human subject research

- Interventional
  - **Randomized control trial** measuring how ChatGPT impacts productivity
    Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence (+ appendix)

  - **A/B experiments** in chatbots
    Understanding the user experience of customer service chatbots: An experimental study of chatbot interaction design

- Observational
  - **Surveys** of teachers about chatbots in the classroom
    Teachers and Students Embrace ChatGPT for Education

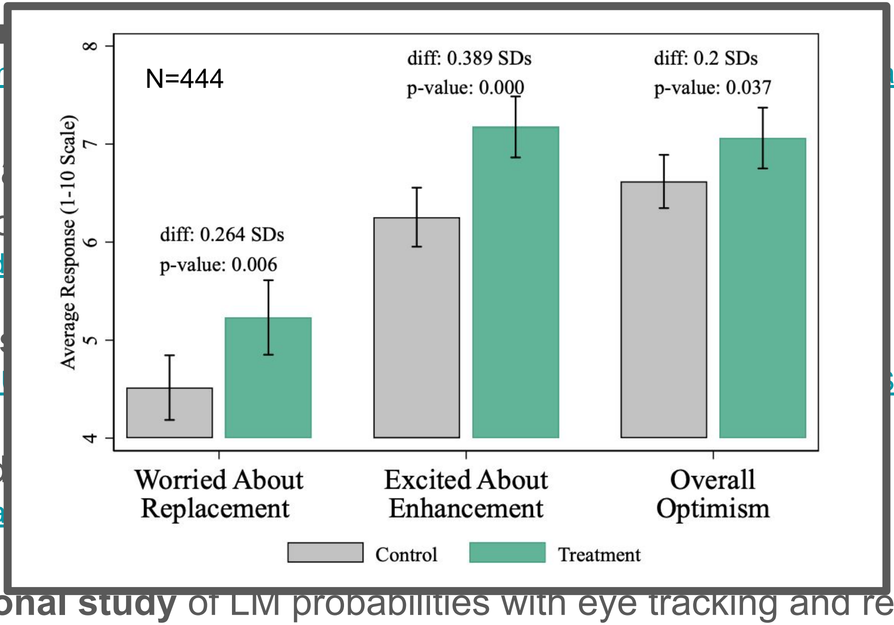
  - **Interviews** with users of customer service chatbots
    What Makes Users Trust a Chatbot for Customer Service? An Exploratory Interview Study

  - **User studies** of users of chatbots on different websites (2018)
    Evaluating and Informing the Design of Chatbots

  - **Correlational study** of LM probabilities with eye tracking and reading time data
    On the Predictive Power of Neural Language Models for Human Real-Time comprehension Behavior

Recent examples of LM-oriented human subject research

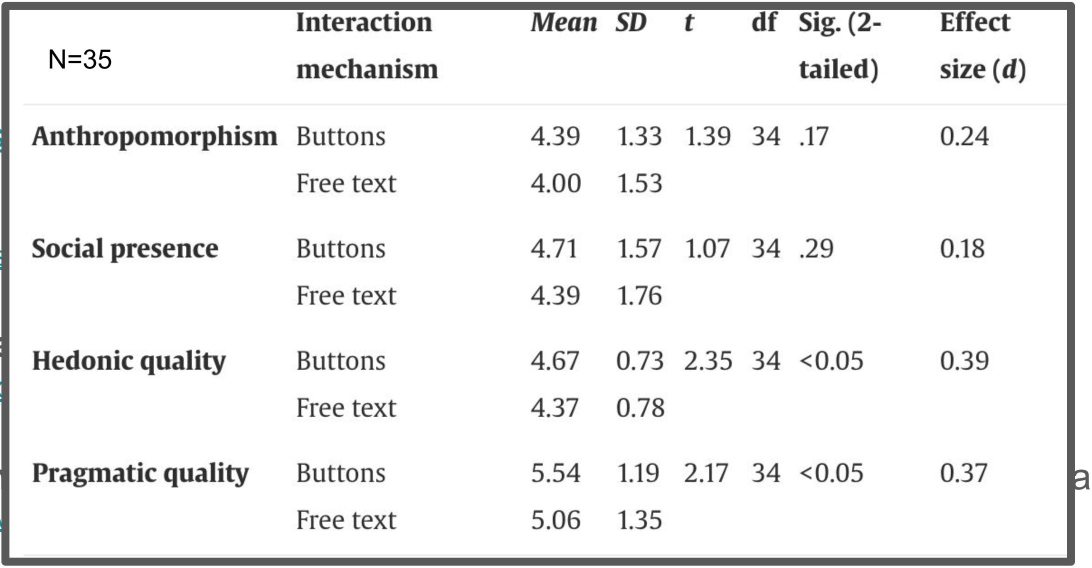
- Interventional
  - **Randomized control trial** measuring how ChatGPT impacts productivity
    [Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence]() (+ [appendix]())



  - **A/B expe**
    Understandir … l study of chatbot interaction design
- Observationa
  - **Surveys c**
    Teachers and
  - **Interviews**
    What Makes … Study
  - **User stud**
    Evaluating a
  - **Correlational study** of LM probabilities with eye tracking and reading time data
    On the Predictive Power of Neural Language Models for Human Real-Time comprehension Behavior

Recent examples of LM-oriented human subject research

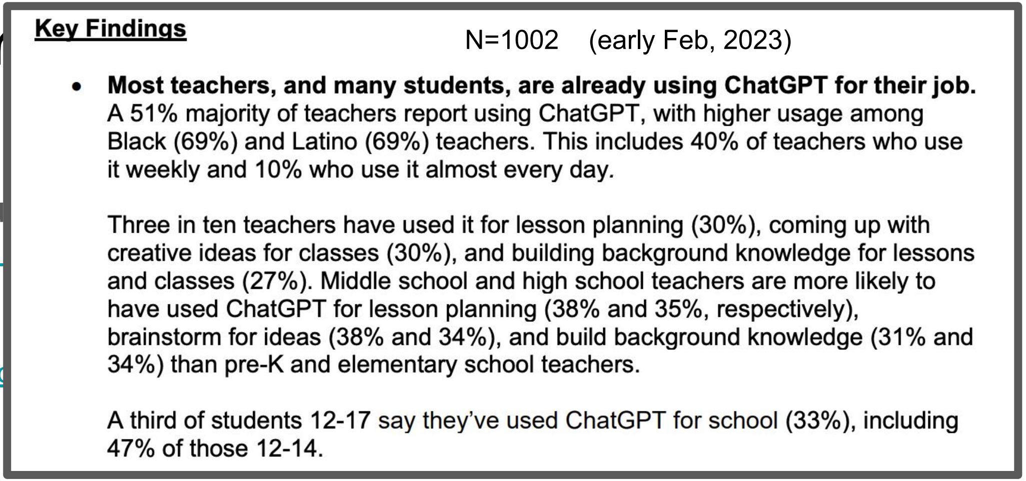
- Interventional
  - **Randomized control trial** measuring how ChatGPT impacts productivity
    Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence (+ appendix)
  - **A/B experiments** in chatbots
    Understanding the user experience of customer service chatbots: An experimental study of chatbot interaction design
- Observationa
  - **Surveys** of
    Teachers and S
  - **Interviews**
    What Makes Us
  - **User studie**
    Evaluating and
  - **Correlation** ata
    On the Predicti



N=35	Interaction mechanism	*Mean*	*SD*	*t*	df	Sig. (2-tailed)	Effect size (*d*)
Anthropomorphism	Buttons	4.39	1.33	1.39	34	.17	0.24
	Free text	4.00	1.53				
Social presence	Buttons	4.71	1.57	1.07	34	.29	0.18
	Free text	4.39	1.76				
Hedonic quality	Buttons	4.67	0.73	2.35	34	<0.05	0.39
	Free text	4.37	0.78				
Pragmatic quality	Buttons	5.54	1.19	2.17	34	<0.05	0.37
	Free text	5.06	1.35				

Recent examples of LM-or

- Interventional
  - **Randomized control trial** measu
    [Experimental Evidence on the Productivity](#)
  - **A/B experiments** in chatbots
    [Understanding the user experience of custo](#)
- Observational
  - **Surveys** of teachers about chatbots in the classroom
    [Teachers and Students Embrace ChatGPT for Education](#)



Key Findings N=1002 (early Feb, 2023)

- **Most teachers, and many students, are already using ChatGPT for their job.** A 51% majority of teachers report using ChatGPT, with higher usage among Black (69%) and Latino (69%) teachers. This includes 40% of teachers who use it weekly and 10% who use it almost every day.

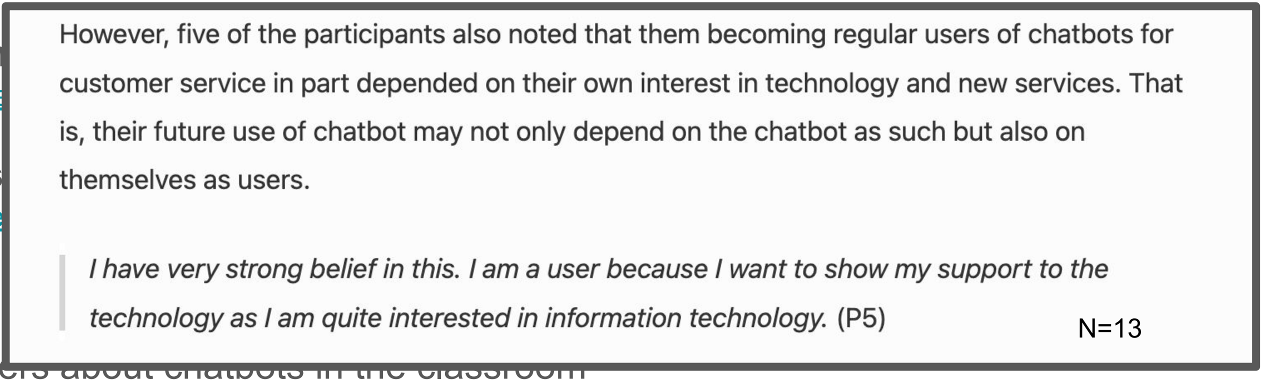
  Three in ten teachers have used it for lesson planning (30%), coming up with creative ideas for classes (30%), and building background knowledge for lessons and classes (27%). Middle school and high school teachers are more likely to have used ChatGPT for lesson planning (38% and 35%, respectively), brainstorm for ideas (38% and 34%), and build background knowledge (31% and 34%) than pre-K and elementary school teachers.

  A third of students 12-17 say they've used ChatGPT for school (33%), including 47% of those 12-14.

  - **Interviews** with users of customer service chatbots
    [What Makes Users Trust a Chatbot for Customer Service? An Exploratory Interview Study](#)
  - **User studies** of users of chatbots on different websites (2018)
    [Evaluating and Informing the Design of Chatbots](#)
  - **Correlational study** of LM probabilities with eye tracking and reading time data
    [On the Predictive Power of Neural Language Models for Human Real-Time comprehension Behavior](#)

Recent examples of LM-oriented human subject research

- Interventional
  - **Randomized con**
    Experimental Evidence
  - **A/B experiments**
    Understanding the use



However, five of the participants also noted that them becoming regular users of chatbots for customer service in part depended on their own interest in technology and new services. That is, their future use of chatbot may not only depend on the chatbot as such but also on themselves as users.

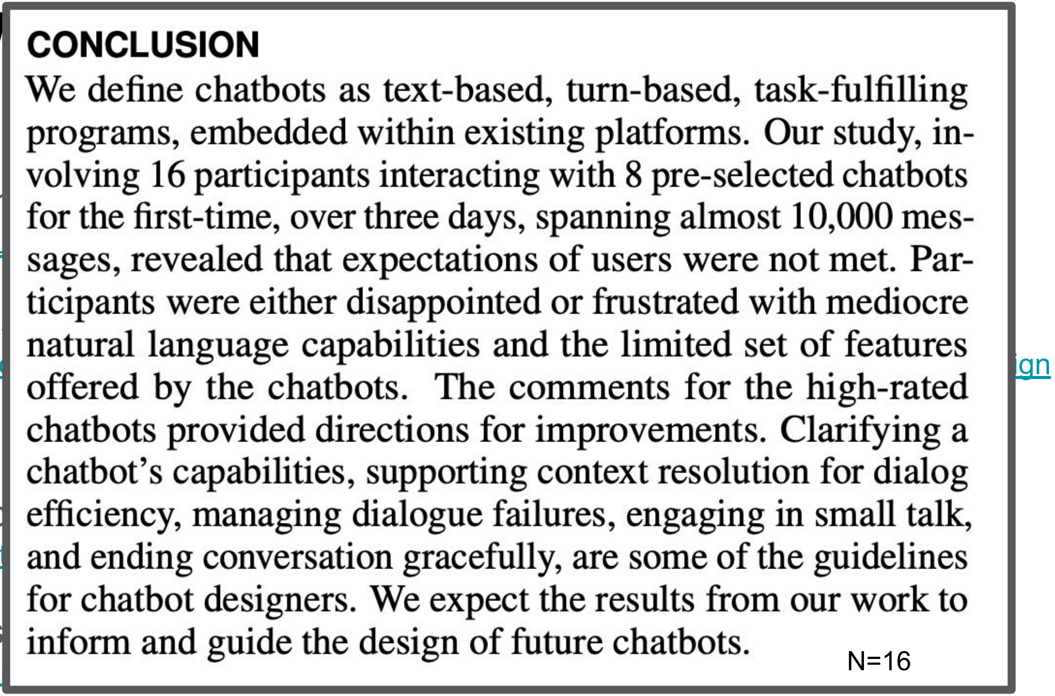
> *I have very strong belief in this. I am a user because I want to show my support to the technology as I am quite interested in information technology.* (P5)

N=13

- Observational
  - **Surveys** of teachers about chatbots in the classroom
    Teachers and Students Embrace ChatGPT for Education
  - **Interviews** with users of customer service chatbots
    What Makes Users Trust a Chatbot for Customer Service? An Exploratory Interview Study
  - **User studies** of users of chatbots on different websites (2018)
    Evaluating and Informing the Design of Chatbots
  - **Correlational study** of LM probabilities with eye tracking and reading time data
    On the Predictive Power of Neural Language Models for Human Real-Time comprehension Behavior

Recent examples of LM

- Interventional
  - **Randomized control trial** m
    Experimental Evidence on the Produ
  - **A/B experiments** in chatbot
    Understanding the user experience c ign
- Observational
  - **Surveys** of teachers about c
    Teachers and Students Embrace Chat
  - **Interviews** with users of cus
    What Makes Users Trust a Chatbot fo
  - **User studies** of users of chatbots on different websites (2018)
    Evaluating and Informing the Design of Chatbots
  - **Correlational study** of LM probabilities with eye tracking and reading time data
    On the Predictive Power of Neural Language Models for Human Real-Time comprehension Behavior

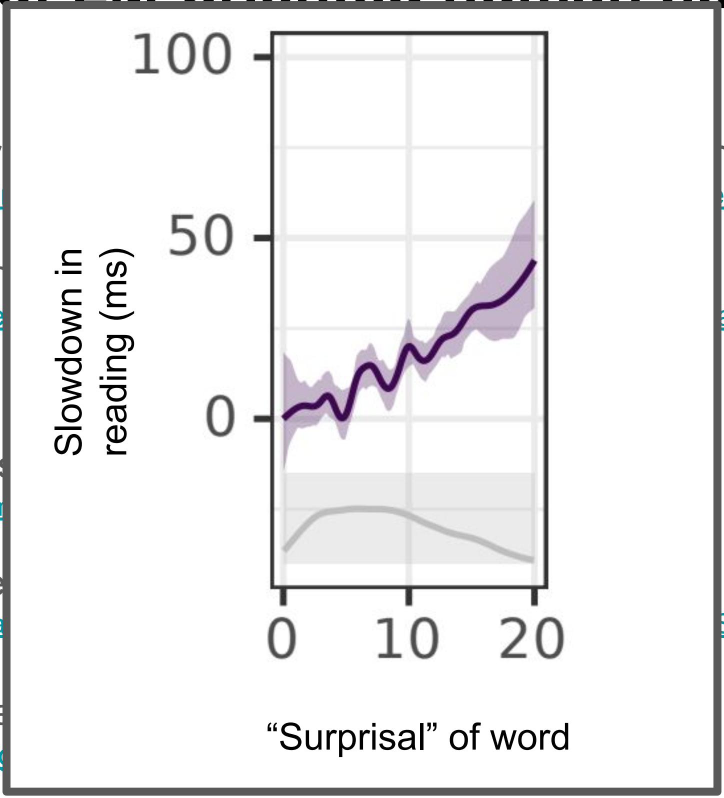


CONCLUSION

We define chatbots as text-based, turn-based, task-fulfilling programs, embedded within existing platforms. Our study, involving 16 participants interacting with 8 pre-selected chatbots for the first-time, over three days, spanning almost 10,000 messages, revealed that expectations of users were not met. Participants were either disappointed or frustrated with mediocre natural language capabilities and the limited set of features offered by the chatbots. The comments for the high-rated chatbots provided directions for improvements. Clarifying a chatbot's capabilities, supporting context resolution for dialog efficiency, managing dialogue failures, engaging in small talk, and ending conversation gracefully, are some of the guidelines for chatbot designers. We expect the results from our work to inform and guide the design of future chatbots.

N=16

Recent examples of LM-oriented human subject research



- Interventional
  - **Randomized contr** roductivity
    Experimental Evidence o ence (+ appendix)
  - **A/B experiments** i
    Understanding the user e al study of chatbot interaction design
- Observational
  - **Surveys** of teacher
    Teachers and Students Em
  - **Interviews** with use
    What Makes Users Trust a Study
  - **User studies** of use
    Evaluating and Informin
  - **Correlational study** of LM probabilities with eye tracking and reading time data
    On the Predictive Power of Neural Language Models for Human Real-Time comprehension Behavior

Human subject research: ethics

Three principles of ethical human subject research (The Belmont Report)

- **Respect for Persons**
  - Requirement for autonomy
  - Requirement to protect those with diminished autonomy

- **Beneficence**
  - Do no harm
  - Maximize possible benefits

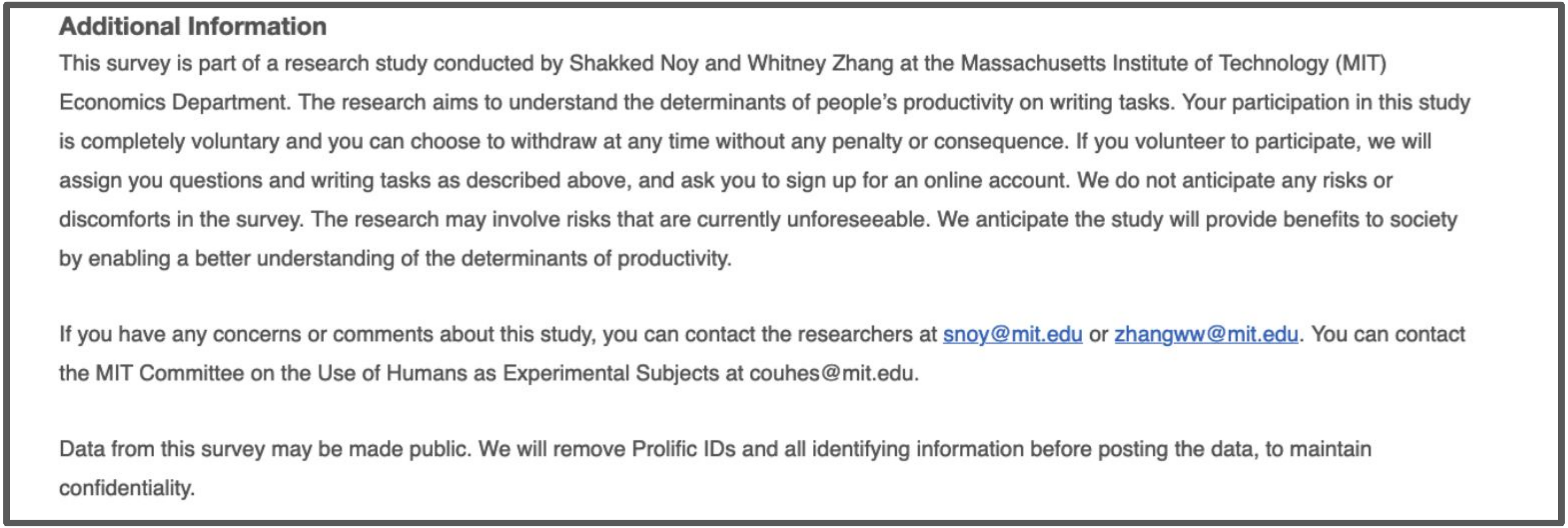
- **Justice**
  - Fairness in selection of research subjects
  - Fairness in distribution of research benefits

Institutional Review Board (IRB)

“The purpose of IRB review is to assure, both in advance and by periodic review, that appropriate steps are taken to protect the rights and welfare of humans participating as subjects in the research”

At MIT, our IRB is called *Committee on the Use of Humans as Experimental Subjects* (https://couhes.mit.edu/).

Research protocols that involve human subjects must be approved by COUHES prior to the start of research. Submit online via https://couhes-connect.mit.edu/



Additional Information

This survey is part of a research study conducted by Shakked Noy and Whitney Zhang at the Massachusetts Institute of Technology (MIT) Economics Department. The research aims to understand the determinants of people's productivity on writing tasks. Your participation in this study is completely voluntary and you can choose to withdraw at any time without any penalty or consequence. If you volunteer to participate, we will assign you questions and writing tasks as described above, and ask you to sign up for an online account. We do not anticipate any risks or discomforts in the survey. The research may involve risks that are currently unforeseeable. We anticipate the study will provide benefits to society by enabling a better understanding of the determinants of productivity.

If you have any concerns or comments about this study, you can contact the researchers at snoy@mit.edu or zhangww@mit.edu. You can contact the MIT Committee on the Use of Humans as Experimental Subjects at couhes@mit.edu.

Data from this survey may be made public. We will remove Prolific IDs and all identifying information before posting the data, to maintain confidentiality.

Consent language from Noy and Zhang's study

COUHES review categories

- **Exempt review**
  - “Benign intervention” or analysis of existing data; minimal risk; no PII
  - *Most* surveys or educational tests qualify for exempt review. Those involving deception, embarrassment, or children may not. Those in which subjects are prisoners never do.
  - Still **must submit a form to COUHES**; PI approval required.

- **Expedited review**
  - Minimal risk intervention
  - Reviewed on rolling basis

- **Full committee review**
  - Reviewed at monthly board meetings

(See https://couhes.mit.edu/definitions for more)

Human subject research: Observational Methods

1) **Surveys**
2) Interviews

Surveys

- Useful for assessing people's **opinions**, **behaviors**, and **experiences**

- **Pros**: Easy to deploy online; low/benign impact on subject

- **Cons**: Many sources of bias to be aware of! Less versatile than interviews.

- Three good resources on how to design surveys:
    - Writing Survey Questions (Pew Research)
    - Harvard's tip sheet
    - Methods of Study Designs- Observational Studies & Surveys

Case study: ChatGPT in the workplace (Noy, Zhang)

Their research question: What kinds of professionals benefit the most and the least from exposure to ChatGPT in terms of satisfaction, efficacy, and productivity?

Their surveys asked:

- Demographics (employment status, income)
- Objective measures of experience, job tenure
- Subjective skill assessment
- Familiarity with other software

- How much did you enjoy doing the task?
- How skilled/effective did you feel while doing the task?
- How long did it take you to do the task?

Case study: ChatGPT in the workplace (Noy, Zhang)

Their research question: What kinds of professionals benefit the most and the least from exposure to ChatGPT in terms of satisfaction, efficacy, and productivity?

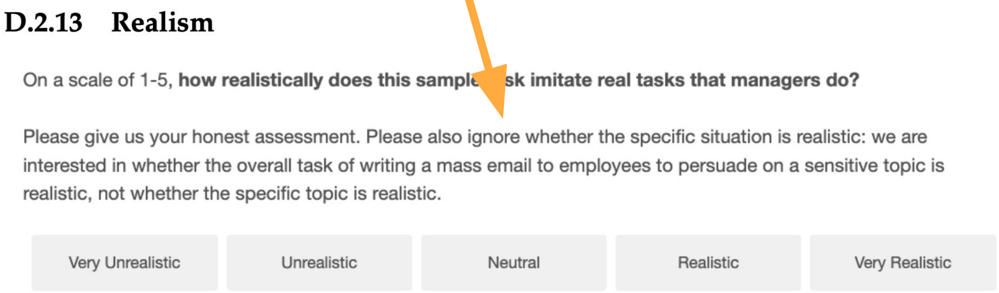
Their surveys asked:

- Demographics (employment status, income)
- Objective measures of experience, job tenure
- Subjective skill assessment
- Familiarity with other software

- How much did you enjoy doing the task?
- How skilled/effective did you feel while doing the task?
- How long did it take you to do the task?

Case study: ChatGPT in the workplace (Noy, Zhang)

Scale question



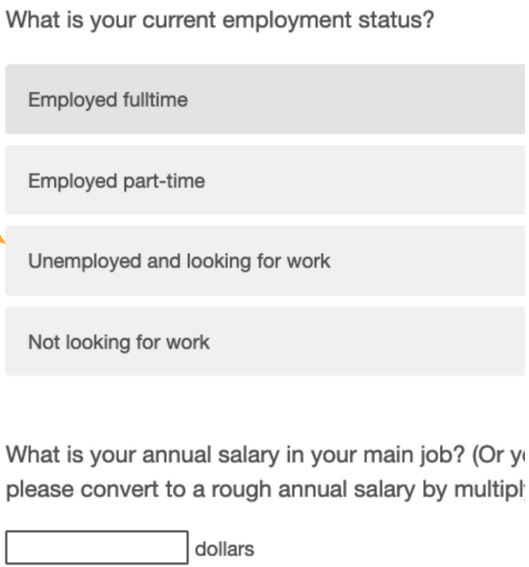
D.2.13 Realism

On a scale of 1-5, **how realistically does this sample task imitate real tasks that managers do?**

Please give us your honest assessment. Please also ignore whether the specific situation is realistic: we are interested in whether the overall task of writing a mass email to employees to persuade on a sensitive topic is realistic, not whether the specific topic is realistic.

Very Unrealistic | Unrealistic | Neutral | Realistic | Very Realistic

Closed question



What is your current employment status?

- Employed fulltime
- Employed part-time
- Unemployed and looking for work
- Not looking for work

What is your annual salary in your main job? (Or y
please convert to a rough annual salary by multipl

□ dollars

Open question

Total Survey Error framework

Goal of TSE is to maximize data quality given a fixed budget.

“Total survey error = representation errors + measurement errors”

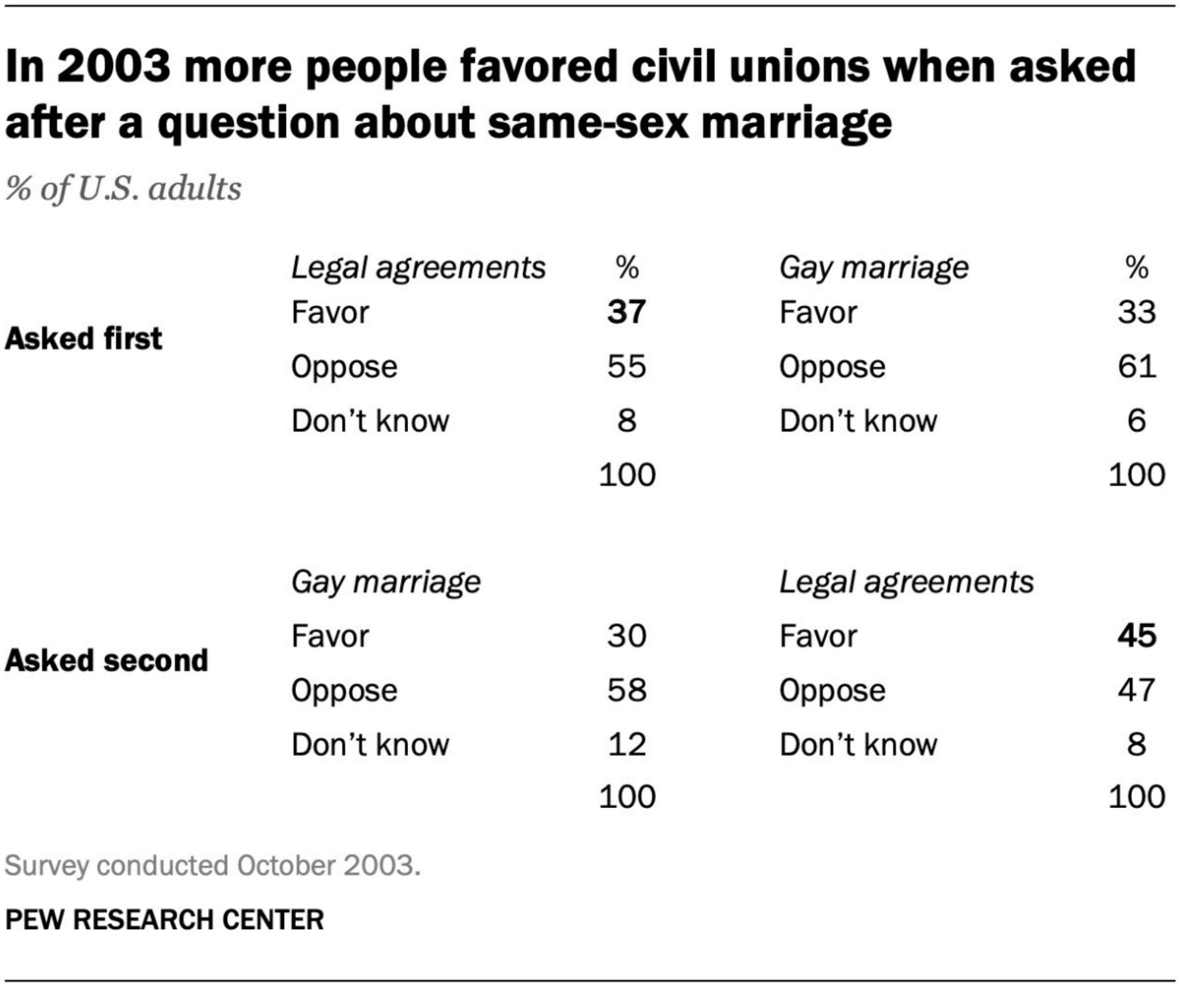
Errors come from who we ask: (“Representation”)

- Sampling bias
- Non-response bias

…and how we ask them: (“Measurement”)

- Question wording biases
- Question ordering biases
- Social desirability bias
- Acquiescence bias

See: Bit by Bit: Social Research in the Digital Aget (Salganik)
Total Survey Error: Design, Implementation, and Evaluation (Biemer)
Nonresponse rates on open-ended survey questions vary by demographic group, other factors (Pew)



In 2003 more people favored civil unions when asked after a question about same-sex marriage

% of U.S. adults

	Legal agreements	%	*Gay marriage*	%
Asked first	Favor	**37**	Favor	33
	Oppose	55	Oppose	61
	Don't know	8	Don't know	6
		100		100
	Gay marriage		*Legal agreements*	
Asked second	Favor	30	Favor	**45**
	Oppose	58	Oppose	47
	Don't know	12	Don't know	8
		100		100

Survey conducted October 2003.

PEW RESEARCH CENTER

Human subject research: Observational Methods

1) Surveys
2) **Interviews**

Qualitative methods

You can learn a lot from… talking to people.

Want to know what educators think of ChatGPT?

    You could ask them!

Anonymized interviews are often exempt sources of research data.

Interview research

In brief:

- Decide who you’re interviewing
- Plan interview guide - what will you ask about?
- Collect notes (according to IRB approval, might need to be anonymous)
- Do *qualitative coding:*
    - Inductive: look at the data, derive themes from the data, label themes in the data
    - Deductive: start with a list of themes, label the data with list of themes
- Analyze!
    - Describe the data, perform statistical analysis, etc.

[If you want more details on this, come talk to us!]

How do text to image models affect design?

Qualitative and quantitative methods can support each other, and you can come to the same conclusions using different methods.

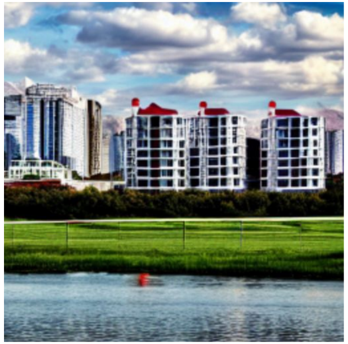
A true story:

"Green red and white buildings in front of a shiny river with a white fence on the side"

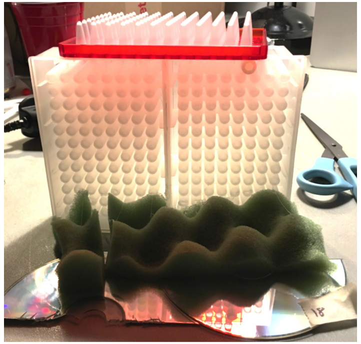
User

Input prompt
to Stable Diffusion

Raw materials



Model generates image
(up to 3)



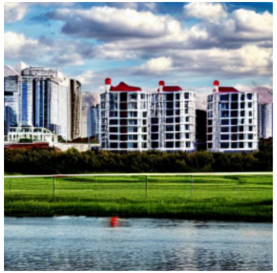
User creates
sculpture

Trash to Treasure: Using text-to-image models to inform the design of physical artefacts
A Smith, H Schroeder, Z Epstein, M Cook, S Colton, A Lippman
The AAAI-23 Workshop on Creative AI Across Modalities

Mixed methods research:

“Green red and white buildings in front of a shiny river with a white fence on the side”

User



Model generates image

Interviewer asks the user:

Did this image inform your design?

Interviewer takes notes

Researcher synthesizes answers, pulls out and analyzes themes

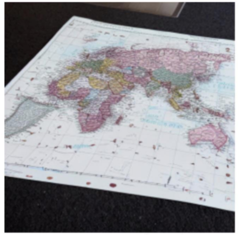
Participants are using prompting for different purposes!

Creating and applying qualitative codes to interviews

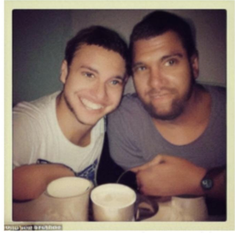
Identified 3+ “styles”

Labeled 30 participants by hand based on their 3 prompts

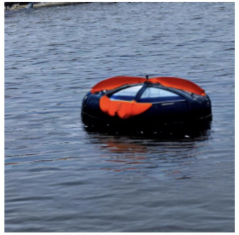
EXPLORER



Prompt 1: ‘My printed map required too much plastic’

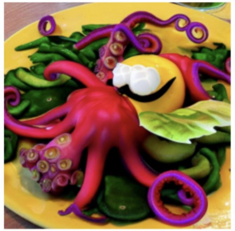


Prompt 2: ‘I’m exhausted but I’m still having fun’

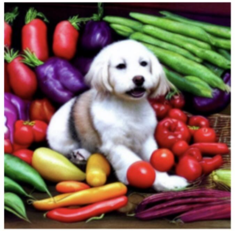


Prompt 3: ‘My hovercraft is full of eels’

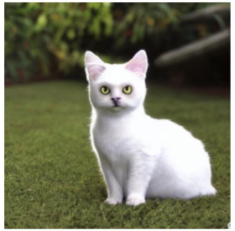
REPHRASER



Prompt 1: ‘Angel vegetables octopus’

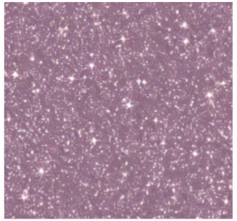


Prompt 2: ‘Angel vegetables dog’

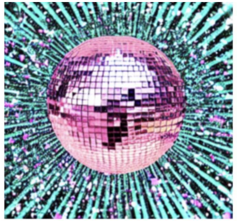


Prompt 3: ‘Cat dog angel’

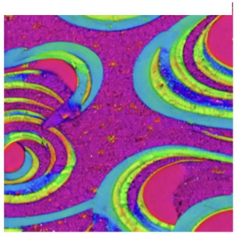
REFINER



Prompt 1: ‘Orb pink glitter utopian soundscape’



Prompt 2: ‘Disco ball pink glitter utopian soundscape waves’



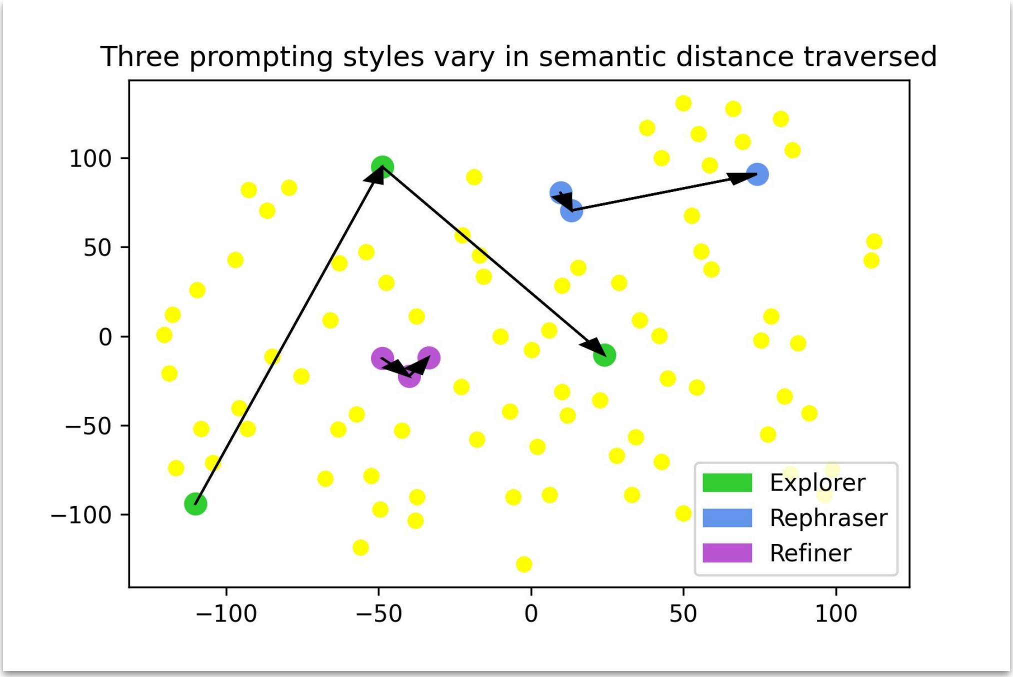
Prompt 3: ‘Disco ball pink glitter utopian soundscape waves diatom ripples’

Trash to Treasure: Using text-to-image models to inform the design of physical artefacts
A Smith, H Schroeder, Z Epstein, M Cook, S Colton, A Lippman
The AAAI-23 Workshop on Creative AI Across Modalities

Mixed methods research

Quantitative version:

Quantified degree of “conceptual exploration” using average cosine distance over 3 prompts



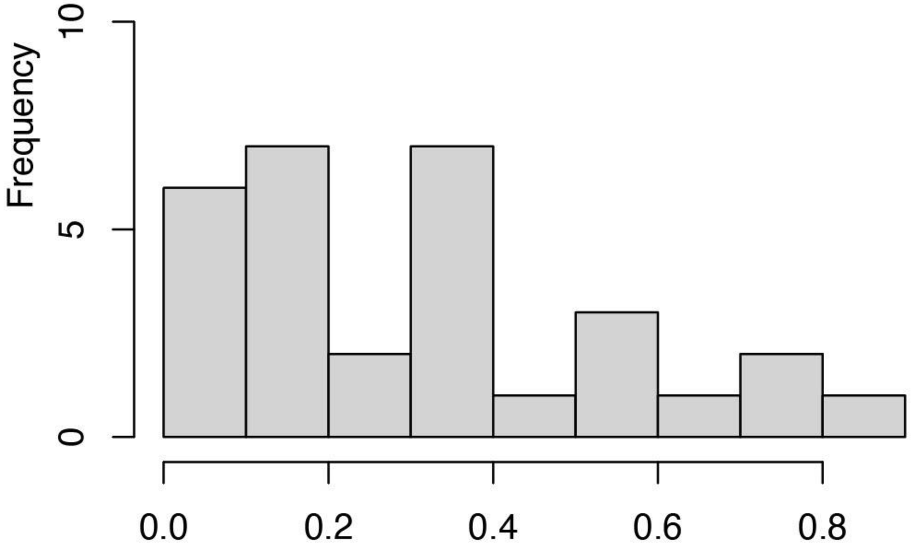
Mixed methods research

Findings reinforced each other!

Point is:

-There are multiple versions of every project

- Choose methods that match your background and goals!



Average cosine distance
(Amount of conceptual exploration)

Human subject research: Interventional methods

1) Randomized Controlled Trials (RCTs)

A Case Study

Pre-ordained outcomes

Research Question: What are the productivity effects of ChatGPT in the context of mid-level professional writing tasks? (Noy and Zhang 2023)

Specific context

Three Elements of an RCT

Randomization

Pre-ordained outcome measures

Blinding

RCTs: Randomization

Why can't we just observe people that use ChatGPT and those who don't?

- Send out a survey to MIT students that asks them to partake in the study
- Measure "writing productivity" for the students that do/don't use ChatGPT and compare the two groups

Selection bias!

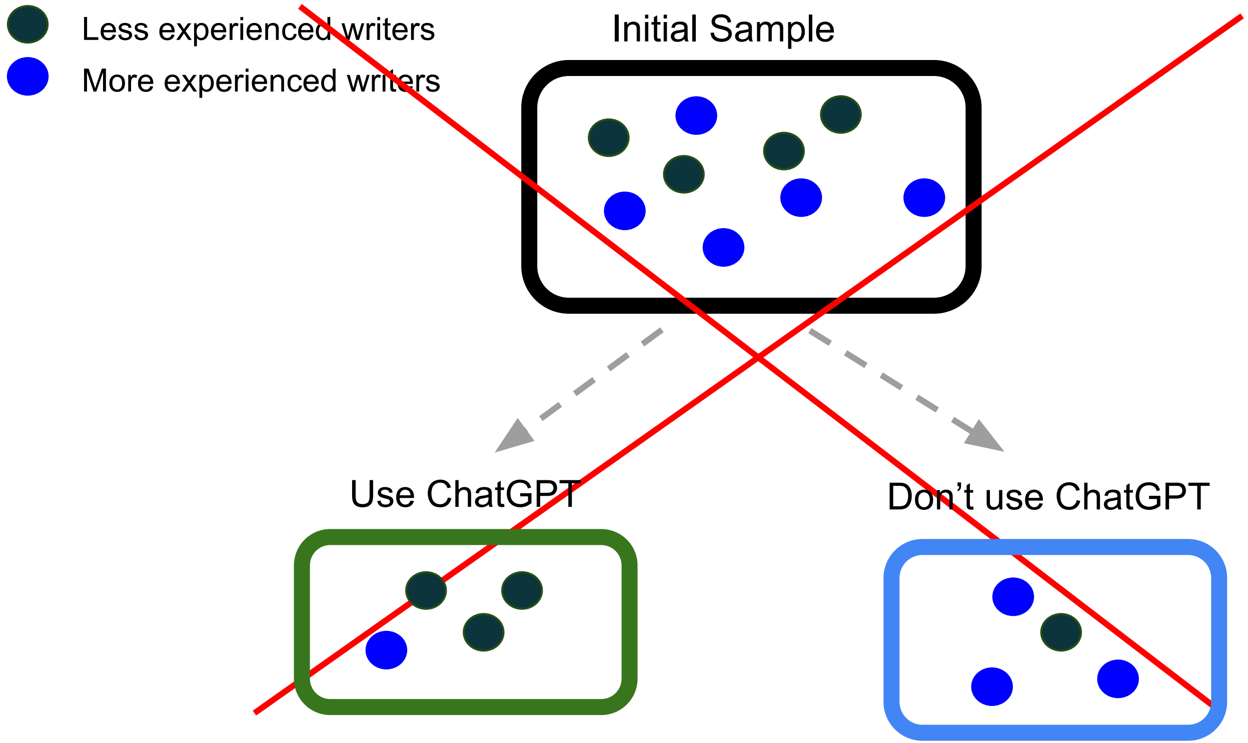
*"occurs when individuals or groups in a study **differ systematically from the population of interest** leading to a systematic error in an association or outcome." (**Catalog of Bias, Oxford University**)*

Confounding!

*"Confounding variables are those that **affect other variables in a way that produces spurious or distorted associations between two variables**. They confound the "true" relationship between two variables." (ICPSR, University of Michigan)*

RCTs: Randomization

What can we do instead?



This is NOT what we want!

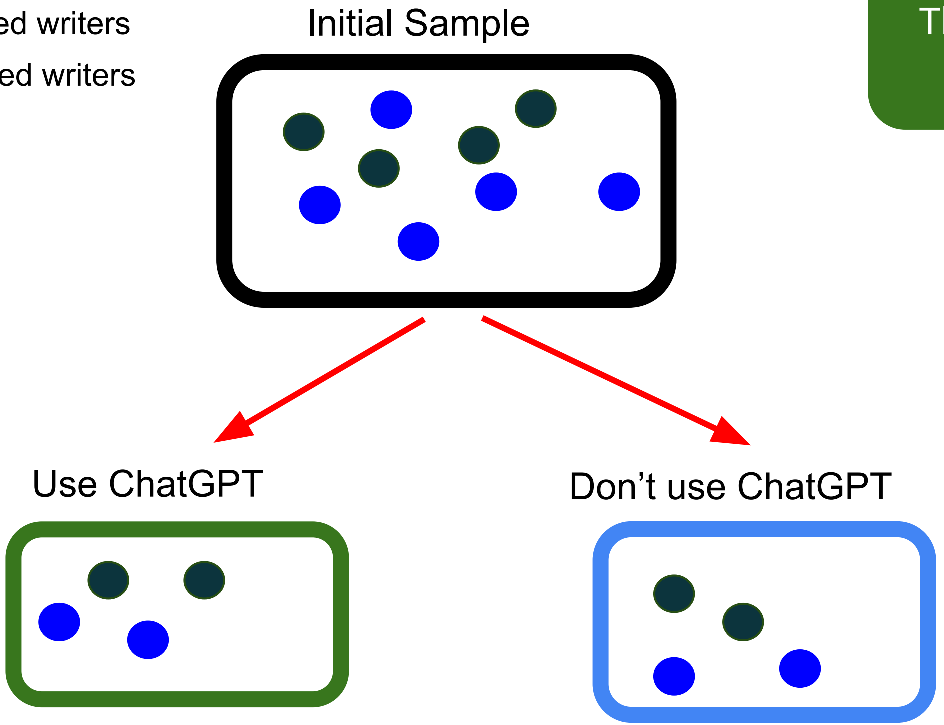
RCTs: Randomization

What can we do instead?

● Less experienced writers

● More experienced writers

This is what we want!



RCTs: Preordained outcome measures

Decide what/how you want to measure **before** you run your RCT

# of tasks completed in a certain time

Quality of writing as determined by human raters

Factual correctness of answers

- Be **precise** and **exhaustive** on your data collection and analysis **before** you run your experiment
- Most of the thinking should be done before running your experiment
- What other data would you want to collect?
  - Demographics
  - Writing experience
- Negative results are positive results!
  - Most experiments don’t work, which is totally fine

RCTs: Blinding

Who knows what about the experiment?

- **Single blind**
  - Only the researcher knows who got what treatment (e.g. surgery)
- **Double blind**
  - Neither the researcher nor the subject knows who got what treatment
    - Ideal case
    - Least chance of bias

RCTs: Practical Analysis Tips

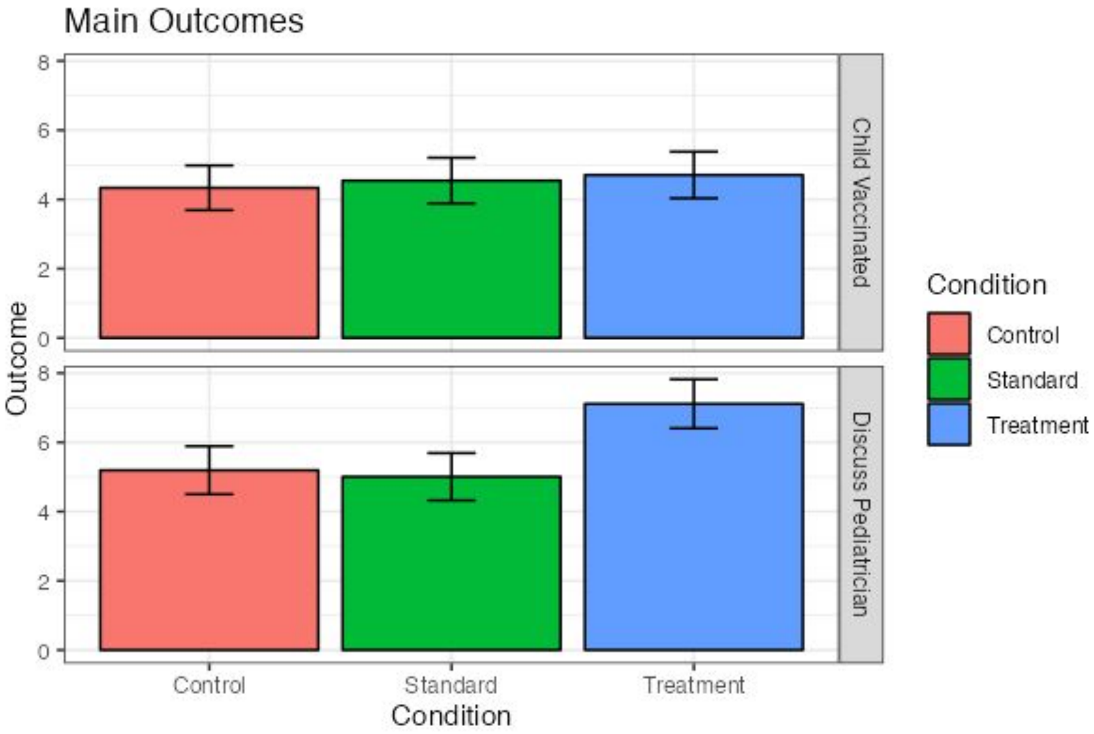
Sanity Checks

- Include **attention** checks!
  - Were your participants actually doing what you asked?
- Ensure **proper randomization**
  - E.g. are the groups balanced on demographics?
- Did **all subjects comply** with the treatment?

What do I do first?

- **Plot the difference** between treatment and control outcomes, and have a sense for what is a “big deal”
  - Don’t rely only on p-values
- **Plot the distribution** of outcomes
  - Is it very skewed?
- **Use regressions** to control for any confounders not balanced by randomization
- **Qualitative analysis** is great! Can collect comments from survey participants
  - Hypothesis for mechanism; why did the intervention work?

RCTs


Main Outcomes
Outcome
8
6
4
2
0
Child Vaccinated
Discuss Pediatrician
Control
Standard
Treatment
Condition
Condition
🟥 Control
🟩 Standard
🟦 Treatment

Human subject research: platforms

1) **Crowdworking platforms**
2) Survey design platforms
3) Other resources

Platform: Amazon Mechanical Turk

- Oldest and most popular “crowdsourcing marketplace”.

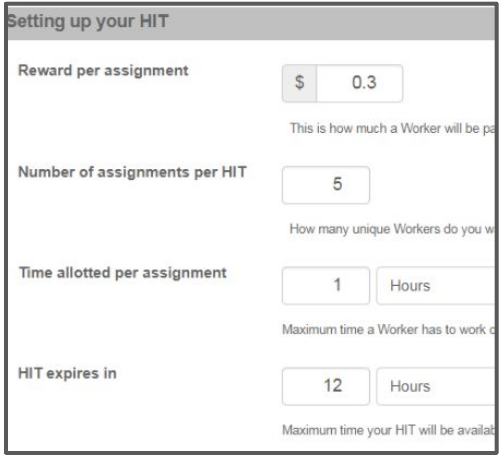
- **Requestors** post “**High Intelligence Tasks**” such as surveys that **Workers** complete these for pay

- ~250k workers, 90% in U.S.

- Workers can be selected based on self-reported **qualifications** such as age, location, political affiliation, education, and many more.

- Good for reaching a large group quickly; bad for having them do hard tasks.

(Here’s a good presentation from UMich about MTurk)



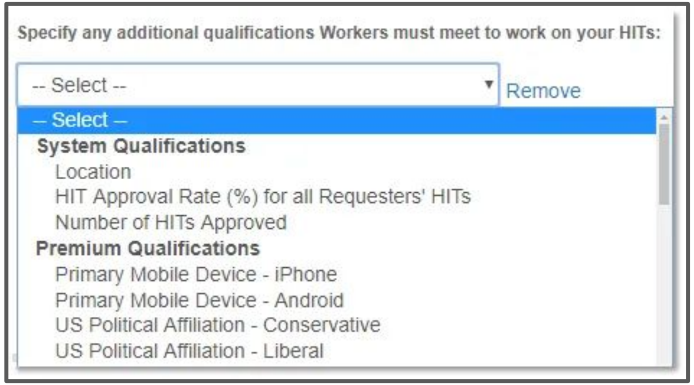
Setting up your HIT

Reward per assignment: $ 0.3
This is how much a Worker will be pa

Number of assignments per HIT: 5
How many unique Workers do you w

Time allotted per assignment: 1 Hours
Maximum time a Worker has to work o

HIT expires in: 12 Hours
Maximum time your HIT will be availab



Specify any additional qualifications Workers must meet to work on your HITs:

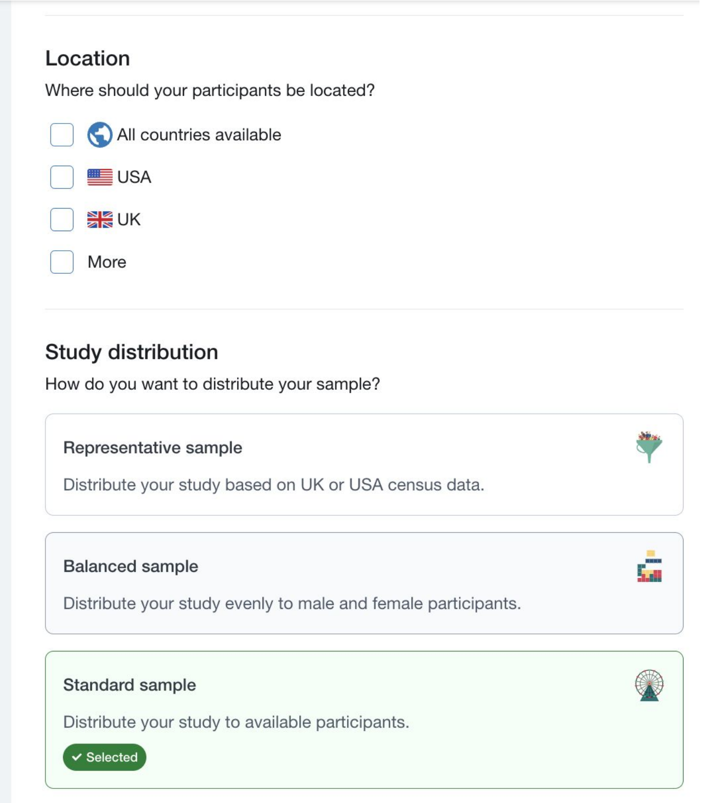
-- Select --
Remove
-- Select --
System Qualifications
Location
HIT Approval Rate (%) for all Requesters' HITs
Number of HITs Approved
Premium Qualifications
Primary Mobile Device - iPhone
Primary Mobile Device - Android
US Political Affiliation - Conservative
US Political Affiliation - Liberal

Platform: Prolific

- More targeted at academic research in behavioral sciences. Better when you need survey responses or have a longer-duration task

- More modern UI than MTurk

- Generally higher data quality. See: Data quality of platforms and panels for online behavioral research



Location

Where should your participants be located?

□ All countries available

□ USA

□ UK

□ More

Study distribution

How do you want to distribute your sample?

Representative sample

Distribute your study based on UK or USA census data.

Balanced sample

Distribute your study evenly to male and female participants.

Standard sample

Distribute your study to available participants.

✓ Selected

Tips about crowd-work platforms in general

- Guard against fraud:
  - Some workers lie about qualifications in order to answer survey
  - Some workers rush through surveys in order to get rewards more quickly

- Tips for improving data quality:
  - Increase qualifications required (e.g., MTurk “Master” status)
  - Attention checks
  - Golden data
  - For a good guide on these and other data annotation practices, see section 3 in Human-in-the-Loop Machine Learning. (Copy available here)

- Always measure inter-annotator agreement

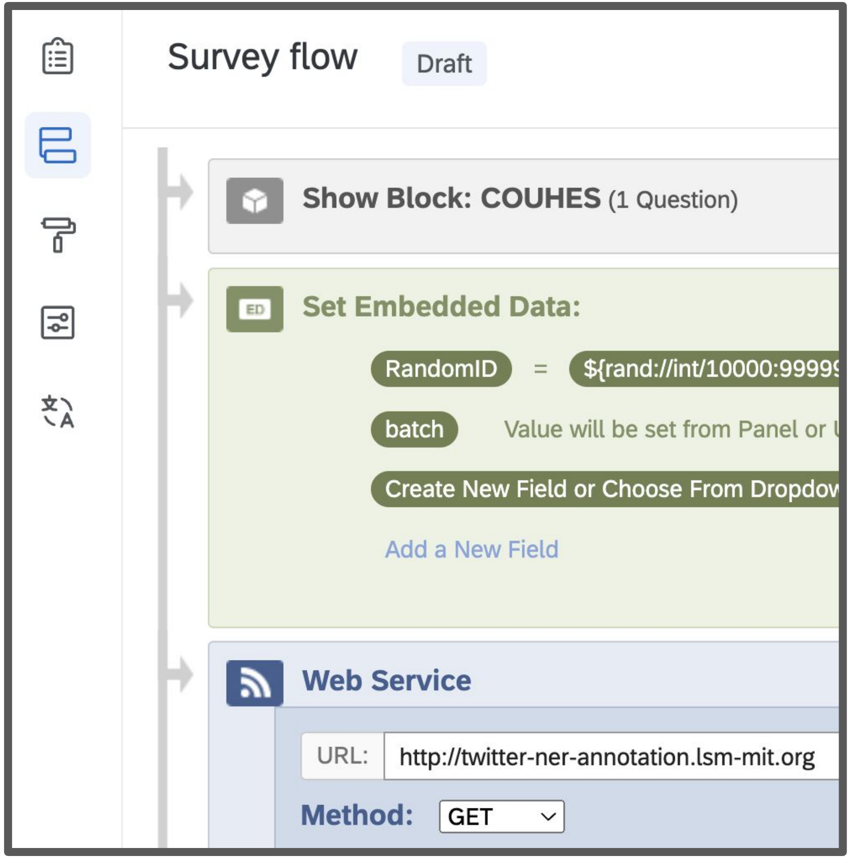
- Pay fairly! At least the prevailing minimum wage in the location of the worker.

Also see:

- Online panels in social science research: Expanding sampling methods beyond Mechanical Turk
- The Perils of Using Mechanical Turk to Evaluate Open-Ended Text Generation

Platform: Qualtrics

- Survey design / deployment tool
- Nice GUI supporting common randomization / looping logic
- Can pair with MTurk/Prolific to recruit respondents
- Incudes library of time-tested demographic questions
- MIT has a license. Set up your account at https://qualtrics.mit.edu/



Survey flow Draft

Show Block: COUHES (1 Question)

Set Embedded Data:

RandomID = ${rand://int/10000:9999

batch Value will be set from Panel or

Create New Field or Choose From Dropdo

Add a New Field

Web Service

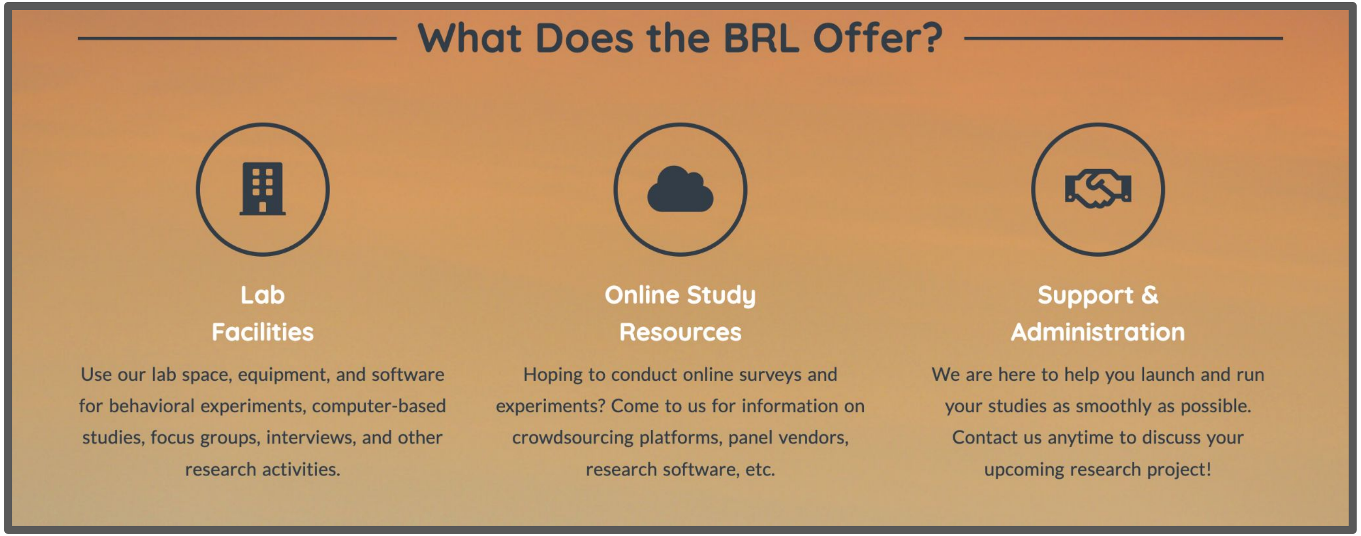
URL: http://twitter-ner-annotation.lsm-mit.org

Method: GET

MIT’s Behavioral Research Lab

Offers MIT researchers assistance with recruitment of research study participants, both on online platforms and in their own “participant pool”

See: https://brl.mit.edu/



What Does the BRL Offer?

Lab Facilities

Use our lab space, equipment, and software for behavioral experiments, computer-based studies, focus groups, interviews, and other research activities.

Online Study Resources

Hoping to conduct online surveys and experiments? Come to us for information on crowdsourcing platforms, panel vendors, research software, etc.

Support & Administration

We are here to help you launch and run your studies as smoothly as possible. Contact us anytime to discuss your upcoming research project!

Logistics

Next week:

Media Lab Research Panel

In person!

Reminders:

Project 1 pager due Friday!

Sign up for office hours

Homework is light this week and due Monday

center for constructive communication

Out-takes

(Topics we didn’t get to cover)

Human subject research: A/B experiments in industry

A/B experiments in industry

“A/B experiment”: industry jargon for an RCT in which the test hypothesis is whether **a change to an application leads to a change in some business criterion.**

Typically:

- **Control** = status quo; **Treatment** = proposed change (e.g., use a new LM to score search results)
- **Experimental unit** is some unit of activity of the application (e.g. users, sessions, search queries)
- **Key metrics** are defined based on this activity (e.g. click-through rate, active days per user, “like” rate)
- Some “**overall evaluation criterion**” (OEC) is defined in terms of the key metrics
- Change is launched if OEC(treatment) > OEC(control) at some level of confidence

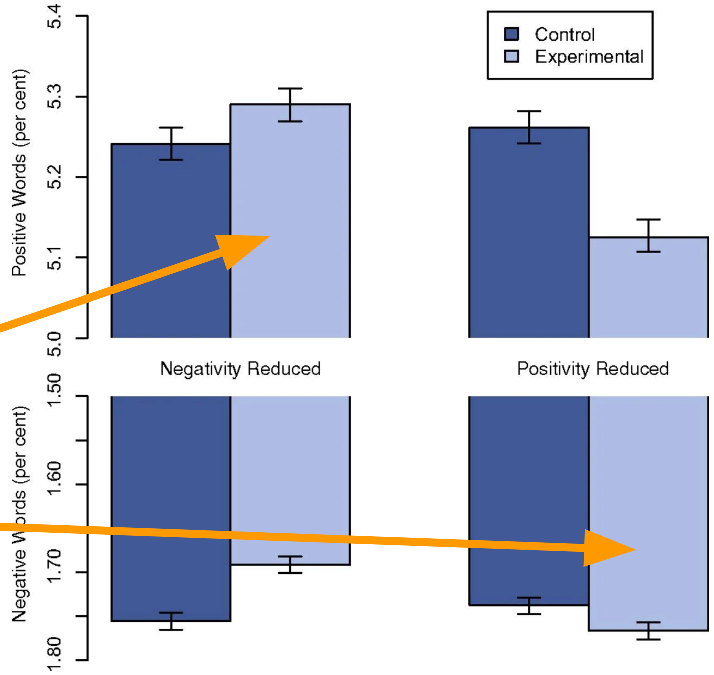
- Building a Culture of Experimentation (HBR)
- Trustworthy Online Controlled Experiments : A Practical Guide to A/B Testing

A/B experiments in industry: Facebook case study

“Emotional contagion experiments” in 2012:

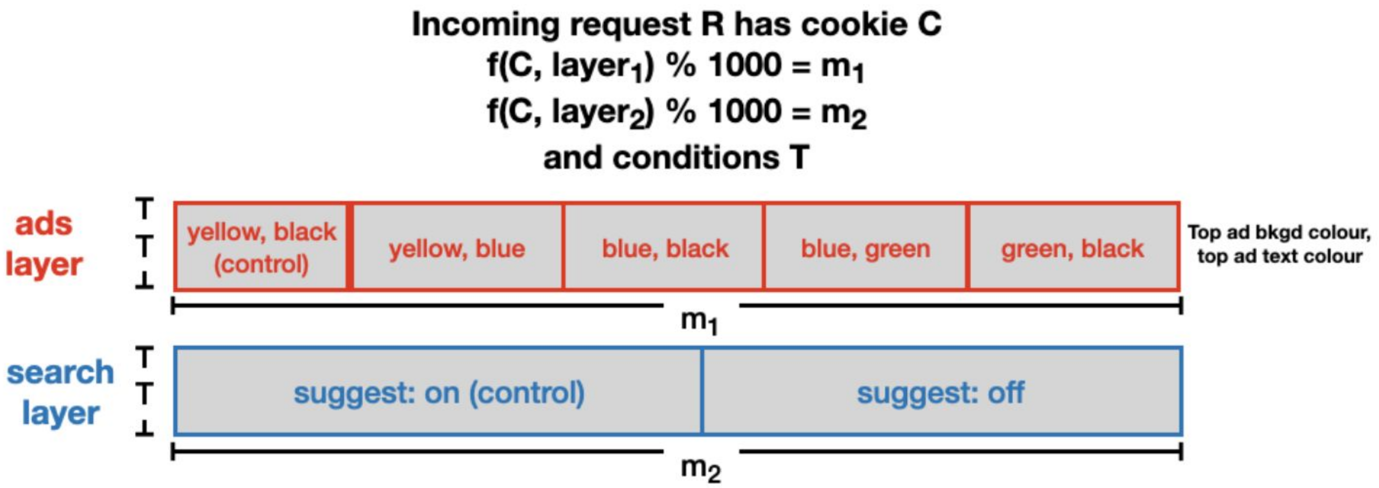
- Control → User gets normal news feed
- Treatment → Some fraction of “positive” or “negative” posts omitted
- Users who had negative posts reduced made more positive status updates
- …and vice versa

(2014 PNAS paper)



Also: Everything We Know About Facebook’s Secret Mood-Manipulation Experiment

A/B experiments in industry: Google example



Incoming request R has cookie C

$f(C, \text{layer}_1) \ \% \ 1000 = m_1$

$f(C, \text{layer}_2) \ \% \ 1000 = m_2$

and conditions T

ads layer

yellow, black (control)	yellow, blue	blue, black	blue, green	green, black

Top ad bkgd colour, top ad text colour

m_1

search layer

suggest: on (control)	suggest: off

m_2

Diagram source: Experimentation platforms at scale (Ambiata)

Implementation Science

How can an evidence-based intervention reach a larger population?

Efficacy trials vs. Effectiveness trials [more]

Concepts:

- Fidelity: How much does the implemented EBI looks like the original plan?
- Sustainability: Will the value of EBI sustain over time?

Work from the developing world

See District outreach paper

Implementation Science at a Glance (NIH)

Preregistration

The Preregistration revolution

Also see https://docs.google.com/presentation/d/11pmZ5jPFdZOGPyTrzmr1eBSJvtVgDgCYs8VNMg89vlw/edit#slide=id.p

Degrees of impact

(Inspired by Stewardship of Global Collective Behavior.)

Scenario	Typical 1st order concern (developer-focused)	2nd order (immediate impact on user)	3rd order (long-term impact on user)	4th order (collective behavior)
Making a new social network	Does it get a lot of usage?	Do users report that they enjoy the time they spend on it?	Do users strengthen their bonds with their local community through it?	Is affective polarization reduced by the existence of this social network?
Adding “autocomplete” to a search input box	Do people make more searches when the auto-complete suggestions are enabled?	How often do the users find the suggestions useful?	Are users seeing diverse perspectives over time?	Does misinformation spread less rapidly?
Adding a “read receipts” feature on a messaging app	Do I get more installs when this feature is offered?	Do users keep the feature turned on?	Is the user’s anxiety reduced?	Are relationships strengthened with the knowledge from this feature?
Putting ads on a web page	Does it make more money than before?	Are people buying things from the ads?	Are people buying things they actually need?	Do we avoid incentivizing “clickbait” content?

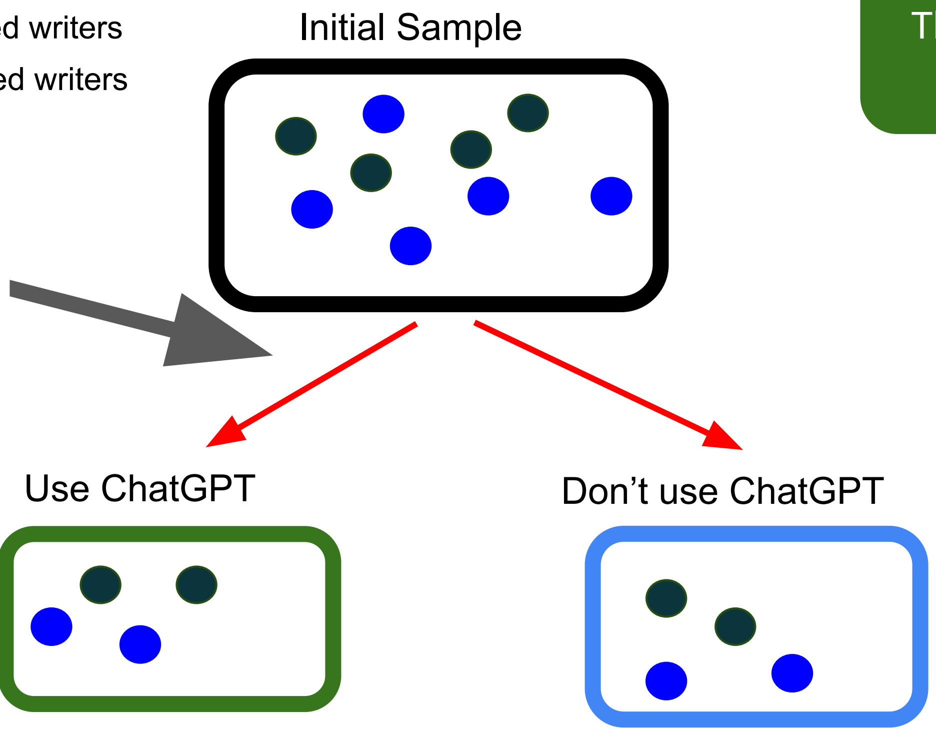
Natural Experiments

What can we do instead?

● Less experienced writers

🔵 More experienced writers

What if "nature" gives us this split?



This is what we want!

Natural/Quasi Experiments:

“The prefix quasi means “resembling.” Thus quasi-experimental research is ***research that resembles experimental research but is not true experimental research****. Although the independent variable is manipulated,* ***participants are not randomly assigned*** *to conditions or orders of conditions (Cook & Campbell, 1979, Research Methods in Psychology)”*

Natural/Quasi Experiments: Examples

- Social Media and Mental Health (Braghieri, Levy, Makarin)
  - Leveraged the **differential rollout** of Facebook across college campuses to estimate the effect of its introduction on student mental health
- The Persuasive Effect of Fox News: Non-Compliance with Social Distancing During the COVID-19 Pandemic (Simonov, Sacher, Dubé, Biswas)
  - Used the fact that **channel numbers** are **randomly assigned (?)** to estimate the effect of Fox News viewership on non-compliance with social distancing
- Large-Scale Psychological Differences Within China Explained by Rice Versus Wheat Agriculture (Talhelm et al.)
  - Geographic quasi-experiment
  - Used villages on **"rice-wheat" border** to test the effect of the type of farming on different psychological traits

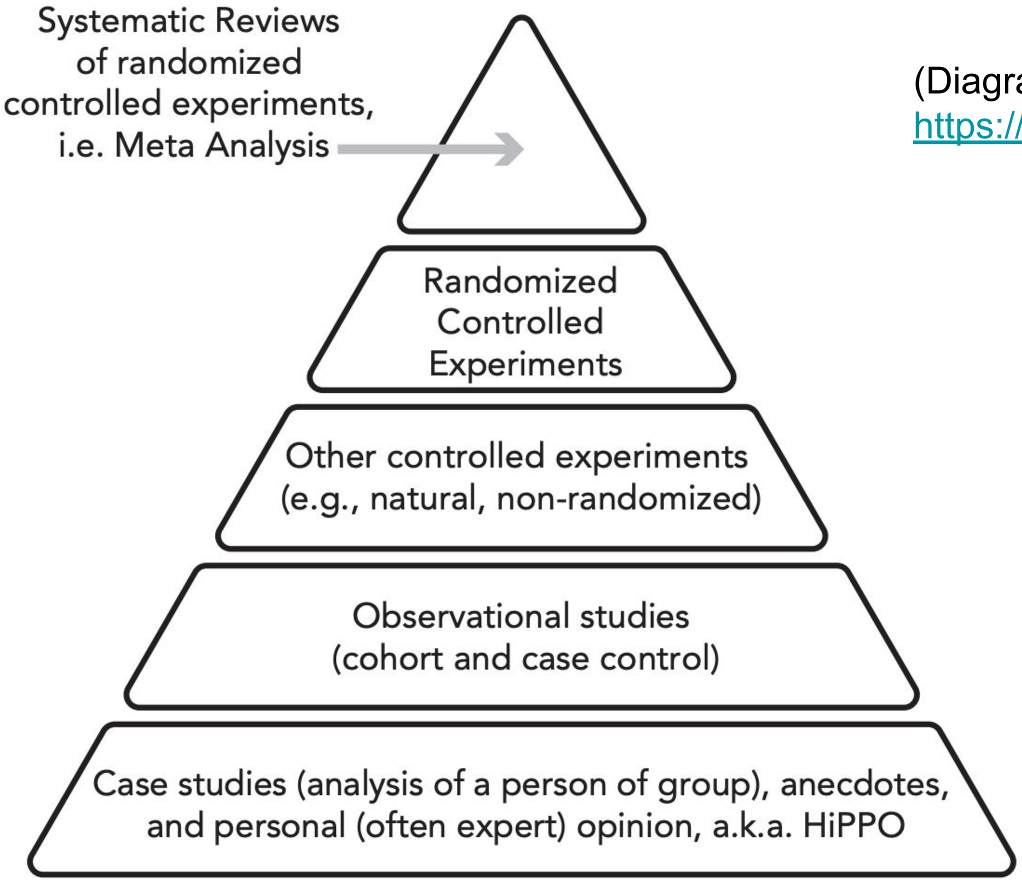
Natural/Quasi Experiments: Pros and Cons

Pros

- No need to recruit participants
- Often larger sample size
- Test hypothesis that would be unethical to test with an RCT (e.g. smoking)

Cons

- “Plausibly random” is not random
- Often missing key confounding variables
- Very little flexibility to collect more data



Systematic Reviews of randomized controlled experiments, i.e. Meta Analysis

Randomized Controlled Experiments

Other controlled experiments (e.g., natural, non-randomized)

Observational studies (cohort and case control)

Case studies (analysis of a person of group), anecdotes, and personal (often expert) opinion, a.k.a. HiPPO

Figure 1.3 A simple hierarchy of evidence for assessing the quality of trial design (Greenhalgh 2014)

(Diagram source: https://experimentguide.com/)