Lecture 5: Experiments; Human Subjects

CMSC 25910

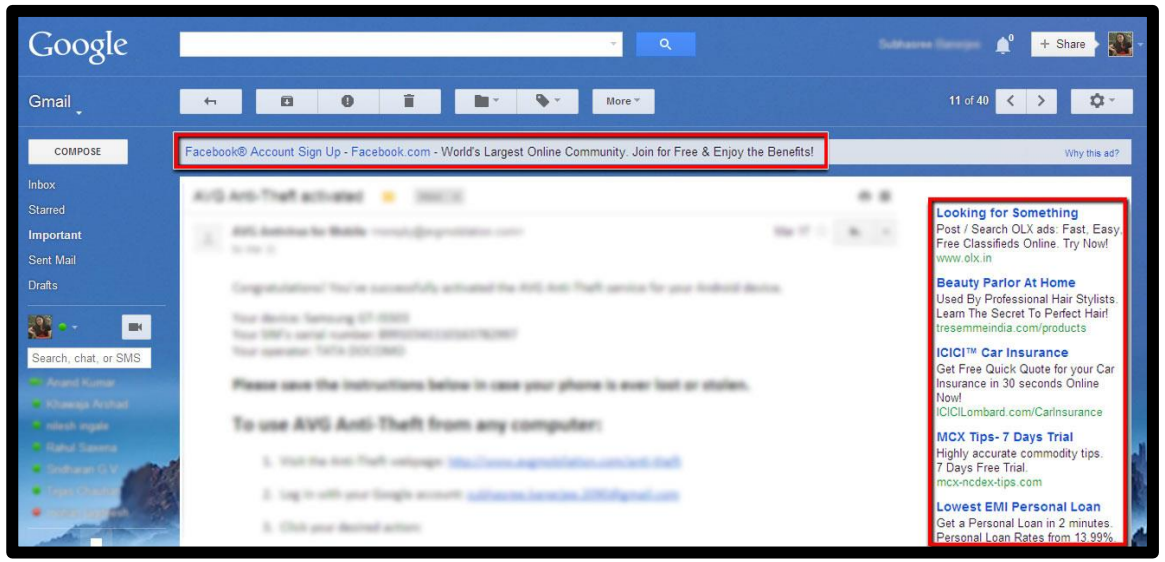
Spring 2022

The University of Chicago

THE UNIVERSITY OF CHICAGO

Running Experiments

A/B Testing & Data-Based **Experiments**



Google's commitment to data-driven decisions is well reported, and the company has been ridiculed for the "50 shades of blue" episode, when then Google executive Marissa Meyer led a project testing the impact of using different coloured links in ads.

But a new insight proves that the company significantly benefitted from the experiment, to the tune of $200m.

The figure comes from Google UK's managing director Dan Cobley, speaking on Tuesday at an event organised by law firm DLA Piper, who positioned the company's approach to data against the traditional route of the "highest paid person's opinion".

"About six or seven years ago, Google launched ads on Gmail," Cobley explained. "In our search we have ads on the side, little blue links that go to other websites: we had the same thing on gmail. But we recognised that the shades of blue in those two different products were slightly different when they linked to ads.

"In the world of the hippo, you ask the chief designer or the marketing director to pick a blue and that's the solution. In the world of data you can run experiments to find the right answer.

"We ran '1%' experiments, showing 1% of users one blue, and another experiment showing 1% another blue. And actually, to make sure we covered all our bases, we ran forty other experiments showing all the shades of blue you could possibly imagine.

"And we saw which shades of blue people liked the most, demonstrated by how much they clicked on them. As a result we learned that a slightly purpler shade of blue was more conducive to clicking than a slightly greener shade of blue, and gee whizz, we made a decision.

"But the implications of that for us, given the scale of our business, was that we made an extra $200m a year in ad revenue."

The form of testing Google undertook is known as A/B testing (offering users two different versions of a site and picking the most effective one); this particular battle was widely seen as a turning point for the company, the moment it sided with engineers against designers. In 2009, Doug Bowman, then the company's top designer, cited it as part of the reason for his departure.

Alex Hern. Why Google has 200m reasons to put engineers over designers. The Guardian. https://www.theguardian.com/technology/2014/feb/05/why-google-engineers-designers

A/B Testing & Data-Based **Experiments**

- Randomly / systematically vary a quantity of interest
- Measure some sort of quantifiable outcome
- Make decisions based on data
  - *Is this always the right way?*

Google's commitment to data-driven decisions is well reported, and the company has been ridiculed for the "50 shades of blue" episode, when then Google executive Marissa Meyer led a project testing the impact of using different coloured links in ads.

But a new insight proves that the company significantly benefitted from the experiment, to the tune of $200m.

The figure comes from Google UK's managing director Dan Cobley, speaking on Tuesday at an event organised by law firm DLA Piper, who positioned the company's approach to data against the traditional route of the "highest paid person's opinion".

"About six or seven years ago, Google launched ads on Gmail," Cobley explained. "In our search we have ads on the side, little blue links that go to other websites: we had the same thing on gmail. But we recognised that the shades of blue in those two different products were slightly different when they linked to ads.

"In the world of the hippo, you ask the chief designer or the marketing director to pick a blue and that's the solution. In the world of data you can run experiments to find the right answer.

"We ran '1%' experiments, showing 1% of users one blue, and another experiment showing 1% another blue. And actually, to make sure we covered all our bases, we ran forty other experiments showing all the shades of blue you could possibly imagine.

"And we saw which shades of blue people liked the most, demonstrated by how much they clicked on them. As a result we learned that a slightly purpler shade of blue was more conducive to clicking than a slightly greener shade of blue, and gee whizz, we made a decision.

"But the implications of that for us, given the scale of our business, was that we made an extra $200m a year in ad revenue."

The form of testing Google undertook is known as A/B testing (offering users two different versions of a site and picking the most effective one); this particular battle was widely seen as a turning point for the company, the moment it sided with engineers against designers. In 2009, Doug Bowman, then the company's top designer, cited it as part of the reason for his departure.

Alex Hern. Why Google has 200m reasons to put engineers over designers. The Guardian. https://www.theguardian.com/technology/2014/feb/05/why-google-engineers-designers

Designing an Experiment

Defining the purpose and goals

- What are you hoping to learn?
  - That is, what are your research questions?
  - **Precisely stated research questions are crucial**
- What are your explicit hypotheses?
- What are your metrics?
  - What data might be directly or indirectly helpful?
- What, if anything, are you comparing to?
  - Control condition or baseline

Research questions (RQs)

- Succinct, **precisely stated**
  - Generally a **falsifiable statement** or **specific question**
  - Usually, but not always, encodes some sort of hypothesis
- Goals of the research can be broad, whereas RQs are usually more narrow
- Let your RQs guide the design of your experiment

Broad types of studies

- Formative (initial) vs. summative (validation)
- **Descriptive study**: describe a phenomenon
- **Relational study**: correlation between variables
- **Experimental study**: causation

- We can do studies with **humans**
  - How humans use a system
  - System-relevant characteristics of humans
- We can also study **systems themselves**

Quantitative vs. Qualitative

- **Quantitative**: numbers
  - Timing how long we awkwardly wait for you all to answer a question
  - Ratings of the course staff’s awesomeness on a numerical scale
  - How long it took a computational process to complete
- **Qualitative**: non-numerical data
  - Free-text thoughts, opinions, understanding, types of errors

What kind of data? (stats implications)

- Quantitative
  - Continuous
  - Discrete
- Categorical
  - Nominal (no order)
  - Ordinal (ordered)

Roles for data

- **Independent variables**: explanatory variables
  - Which variant/condition/treatment was assigned
  - **Covariates**: characteristics of participants (demographics, experiences, or other aspects) that could explain differences
- **Dependent variable(s)**: your main metric(s) of interest
  - The primary thing you're measuring that you expect to change based on the variant/condition/treatment

Point(s) of comparison

- Control condition / baseline:
  - May be a **placebo** (no actual intervention)…
  - …or it may be a state-of-the-art system
- You often create variants (**conditions, treatments**)
  - Think about your research questions and how comparing pairs or groups of conditions lets you answer your research questions
  - A common rookie mistake is not having sensibly matched sets of conditions, introducing **confounds** (other factors that might cause any differences observed)
- When studying a system, you might need “typical” (or intentionally atypical) workloads or traces

Study designs

- **Within-subjects**
  - Every participant tests everything
  - Crucial to randomize order! (learning effect)
  - Fewer participants needed, but longer study
- **Between-subjects**
  - Each participant tests 1 version of the system
  - You compare these groups
  - Groups should be similar (verify!)
  - Still randomize!

Conclusions From Studies

Validity

- To what degree are we confident that X causes Y (**internally valid**)?
- To what degree can we generalize about our results (**externally valid**)?
  - What biases does our sample introduce?
- Is this study **ecologically valid**?
  - Does it mirror real-life conditions and context?
- Balancing all of these is hard!

What we conclude from studies

- It’s very rare that we conclude something like “for all humans there is an X% effect of Y” or “Z% of people care about ethics”
  - Be clear about what population you have sampled
- We often use proxies in measurement

What we conclude long-term

- **Repeatability**: findings consistent with same researchers and same infrastructure
- **Reproducibility**: findings consistent with different researchers and different (comparable) infrastructure
- Sadly, few studies are replicated
  - Bias against successful replication in peer review
  - (Also) bias against publishing negative results

Some potential confounds (1/3)

- Measurement accuracy / resolution
- Differences caused by different experimental platforms
- Order of recruiting matters
  - Round-robin (123123123, etc.), Latin squares
- Time of day for recruiting matters
- Failing to account for study dropout or non-participation (very subtle!)
- Changing multiple aspects across conditions that are compared

Some potential confounds (2/3)

- Learning effect
  - Randomize order of tasks
  - Consider learning effect as a covariate
- Different instructions for different participants
- Biases of recruitment / representativeness
- Self-report biases
  - Don’t ask people to rate expertise

Some potential confounds (3/3)

- Different demographics in conditions
- **Placebo effect**
  - Why you need a control condition
- **Hawthorne effect** (changing behavior in response to being observed)
- Habituation / novelty
  - People pay more attention to new things
- Participants try to please experimenter
  - I like yours better!
  - Minimize knowledge of what's being tested

Process

An example study

- (Bad) research question: “Is UChicago the place where fun comes to die?”
- Recruiting participants: what can go wrong?
- Independent variable: assign a university?
- Dependent variable: some proxy for fun
  - Hours not studying?
  - Hours not in the Reg?
  - Agreement with statement “We are having fun”

Another example study

- What if you have a computationally expensive ML-based intrusion detection system you have created to detect network-based attacks
- What are your research questions?
- What are your variables of interest?

Describing your methods

- Be clear and honest about what you did
  - Be honest, earnest, and upfront about limitations
- Give enough detail for someone to replicate
  - Study materials as appendix if possible
  - Correctly report stats (e.g., APA guidelines)
- Release code if possible
- Release data if possible
  - Requires approval from IRB **and** participants

Pilot studies

- Conduct pilot studies!!!
- Check wording
- Cognitive interviews
  - Encourage pilot participants to say when there is ambiguity or uncertainty
  - Are there answers missing?
  - Pick terminology and elicit pilot participants' understanding
- Verify that you're getting the measurements you thought and that your software works
- Have people "think aloud"

Data to collect during experiments

- Actions and decisions
- Performance (time, success rate, errors)
- Opinions and attitudes (self-reported)
- Audio recording, screen capture, video, mouse movements, keystrokes
- Information about participants' backgrounds/demographics

The (Non-)Protection of Human Subjects

The Monster Study (1939)

- University of Iowa, Wendell Johnson and Mary Tudor
  - RQ: Impact of affirmation/criticism in speech therapy
- Participants: 22 orphan children
  - One group received positive speech therapy and praise for fluency
  - Another group received criticism for every imperfection
- Children in the 2nd group developed permanent speech issues

The Milgram Study (1961)

- Yale University, Stanley Milgram
  - RQ: Obedience of authority figures
- Participants shocked another “participant” (actually an actor who was a confederate of the experimenters) for wrong answers
- Most subjects expressed a desire to stop, but continued when told they would not be held responsible

Stanford Prison Experiment (1971)

- Stanford psychology professor Philip Zimbardo
  - RQ: Is brutality a personality trait of guards or situational?
- Participants: 24 men recruited for 2-week experiment
  - 12 took role of *prisoners,* assigned numbers and uniforms
  - 12 took role of *guards* with wooden batons and uniforms
- Prisoners were arrested at home, stripped naked at the “prison”
- Rebellion, degradation, breakdowns occurred
- Physical and psychological trauma
  - Experiment terminated after 6 days

Tuskegee Syphilis Experiment (1972)

- US Public Health Service and Tuskegee University *"Tuskegee Study of Untreated Syphilis in the Negro Male"* 1932-1972
- Participants: 600 impoverished, African-American sharecroppers
  - 399 with syphilis, 201 without (control group)
- Not told they had syphilis
- Not treated with penicillin
  - Notably, syphilis was entirely treatable by the end of the experiment

Facebook Social Contagion (2014)

- “We show, via a massive (*N* = 689,003) experiment on Facebook, that emotional states can be transferred to others via emotional contagion, leading people to experience the same emotions without their awareness. We provide experimental evidence that emotional contagion occurs without direct interaction between people (exposure to a friend expressing an emotion is sufficient), and in the complete absence of nonverbal cues.” https://www.pnas.org/content/111/24/8788
- Your thoughts?

The Protection of Human Subjects

Belmont Report

- *“Ethical Principles and Guidelines for the Protection of Human Subjects of Research, Report of the National Commission for the Protection of Human Subjects of Biomedical and Behavioral Research”*
- September 30, 1978
- Three key principles (following slides)

Principle 1: Respect for Persons

- Protect the autonomy of persons
- Informed consent
- No deception* (will revisit in a few slides)

Principle 2: Beneficence

- “Do no harm”
- Maximize benefits relative to risks

Principle 3: Justice

- Do not exploit participants
- Fairly distribute costs/benefits to prospective participants

Menlo Report

- The Menlo Report (2012) focused on the intersection between human-subjects experiments and cybersecurity research
- It added a fourth principle: respect for the law and public interest

Study Methods

Types of human-subjects studies (1)

- What people want/think/do overall:
  - Surveys
  - Interviews
  - Focus groups
- What people want/think in context:
  - Contextual inquiry (interviews)
  - Diary study (prompt people)
  - Observations in the field

Types of human-subjects studies (2)

- Usability test:
  - Laboratory (“think aloud”)
  - Online study
  - Log analysis

Types of human-subjects studies (3)

- Controlled experiments to test causation
- Varying different conditions
  - Full-factorial design or not
  - Independent and dependent variables
- Many methods apply (e.g., surveys can be used to test causation)
  - Role-playing studies
  - Field studies

Non-human-subjects studies

- Analyze existing data
  - Is the data public or private?
- Analyze computational performance
- Scrape data from the web
  - Does a scrape violate a website's terms of service?
- Expert evaluation of usability (*not really human-subjects studies*):
  - Cognitive walkthrough
  - Heuristic evaluation

Logistics for a study

- How many participants?
  - Statistical power
  - Time, budget, participants' time
- What kind of participants?
  - Skills, background, interests
  - Their motivations
  - Often not a representative sample
- What do you need to build, if anything?
  - Prototype fidelity

Participants, ethics, and deception

Participants (1)

- Recruit people to do something remotely (e.g., online)
- Recruit people to come to your lab
- Recruit people to let you into their “context”
- Observe people (if possible, get consent! If not possible, consider necessity of design)

Participants (2)

- What recruitment mechanisms?
  - Craigslist, flyers, participant pools, representative sample, standing on street
- How do you compensate them?
  - Ethics of paying $0.00 vs. $10.00 vs. $100,000
- How do you get informed consent?
- What happens to their data?
- Prior knowledge / “what” are they?

Ethics

- How do we protect participants?
  - What risks do we introduce?
- Is there a less invasive method that would give equivalent insight?
- How do we make sure participation is voluntary throughout the experiment?

Deception

- Do we mind if participants know precisely what is being studied?
  - Sometimes, it's crucial that we observe their organic responses in context
- What "deception" or "distraction" task can we introduce?
- How do we **debrief** people at the end?

Institutional Review Board (IRB)

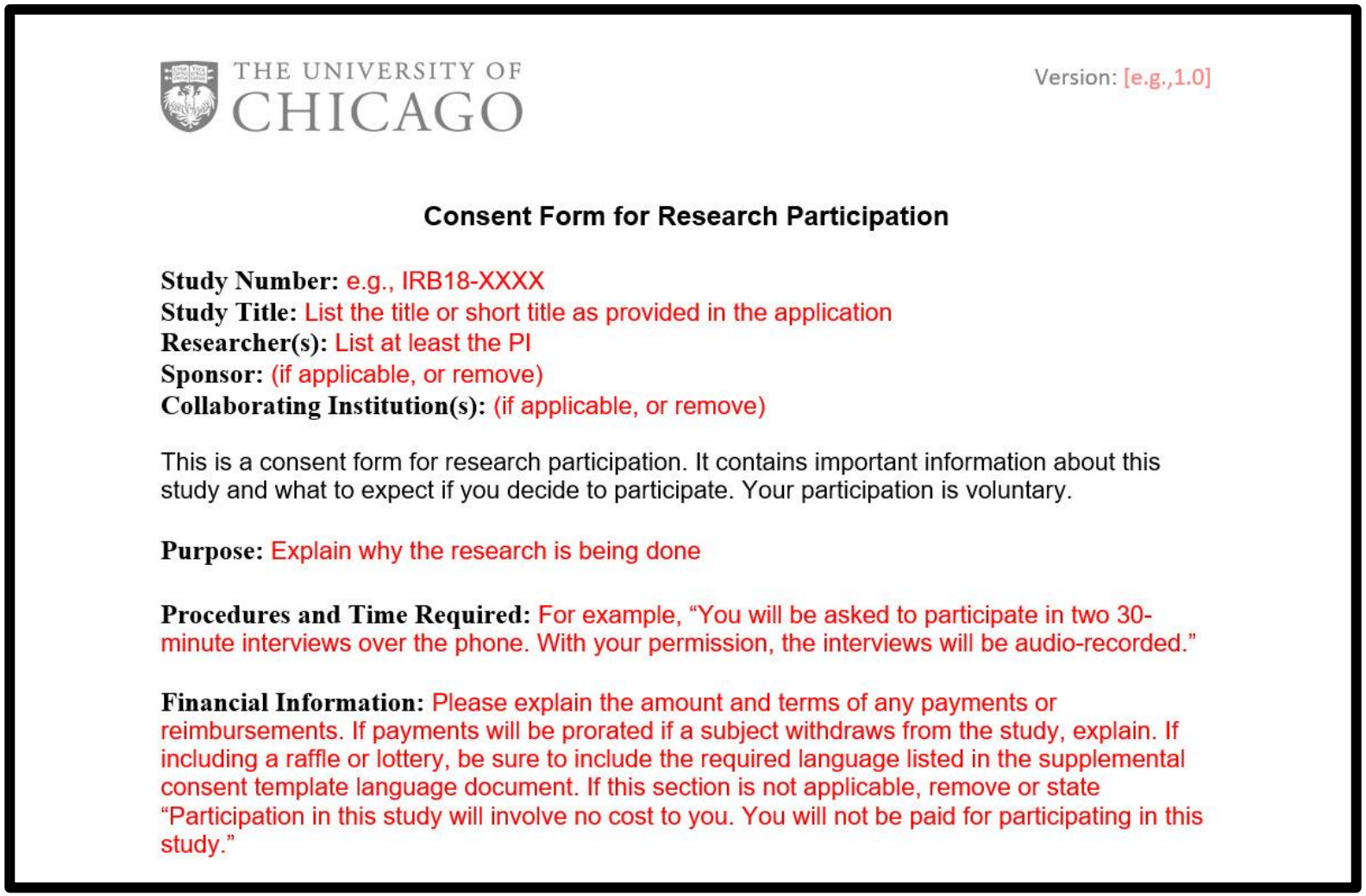
- Is it research? Are there human subjects?
- IRB is one arbiter of ethics; experimenters themselves are another crucial arbiter
- Full review vs. expedited vs. exempt
- Fill out and submit protocol
  - Include all study materials (e.g., surveys)
  - Include recruitment text and/or poster
  - Leave plenty of time

What to submit to an IRB

- Full consent form (use UChicago model)
- All scripts, survey questions, instructions
- Recruitment plan
- Recruitment materials
  - Don’t emphasize compensation
- Information about how data will be handled
  - Password protection, encryption, etc.
  - Meetings to discuss

Informed Consent Templates

https://sbsirb.uchicago.edu/templates/



THE UNIVERSITY OF CHICAGO

Version: [e.g.,1.0]

Consent Form for Research Participation

Study Number: e.g., IRB18-XXXX
Study Title: List the title or short title as provided in the application
Researcher(s): List at least the PI
Sponsor: (if applicable, or remove)
Collaborating Institution(s): (if applicable, or remove)

This is a consent form for research participation. It contains important information about this study and what to expect if you decide to participate. Your participation is voluntary.

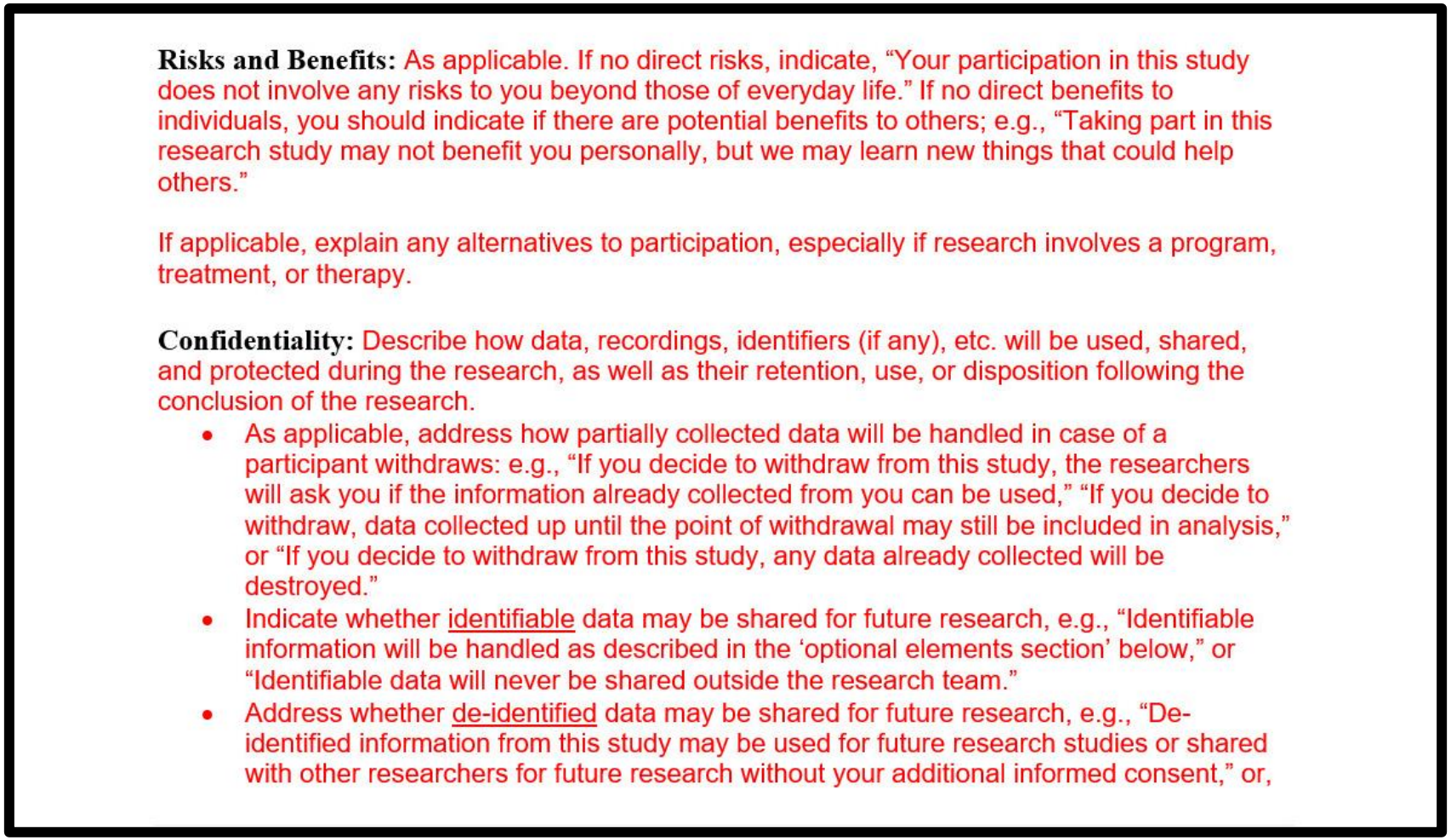
Purpose: Explain why the research is being done

Procedures and Time Required: For example, "You will be asked to participate in two 30-minute interviews over the phone. With your permission, the interviews will be audio-recorded."

Financial Information: Please explain the amount and terms of any payments or reimbursements. If payments will be prorated if a subject withdraws from the study, explain. If including a raffle or lottery, be sure to include the required language listed in the supplemental consent template language document. If this section is not applicable, remove or state "Participation in this study will involve no cost to you. You will not be paid for participating in this study."

Informed Consent Templates

https://sbsirb.uchicago.edu/templates/



Risks and Benefits: As applicable. If no direct risks, indicate, "Your participation in this study does not involve any risks to you beyond those of everyday life." If no direct benefits to individuals, you should indicate if there are potential benefits to others; e.g., "Taking part in this research study may not benefit you personally, but we may learn new things that could help others."

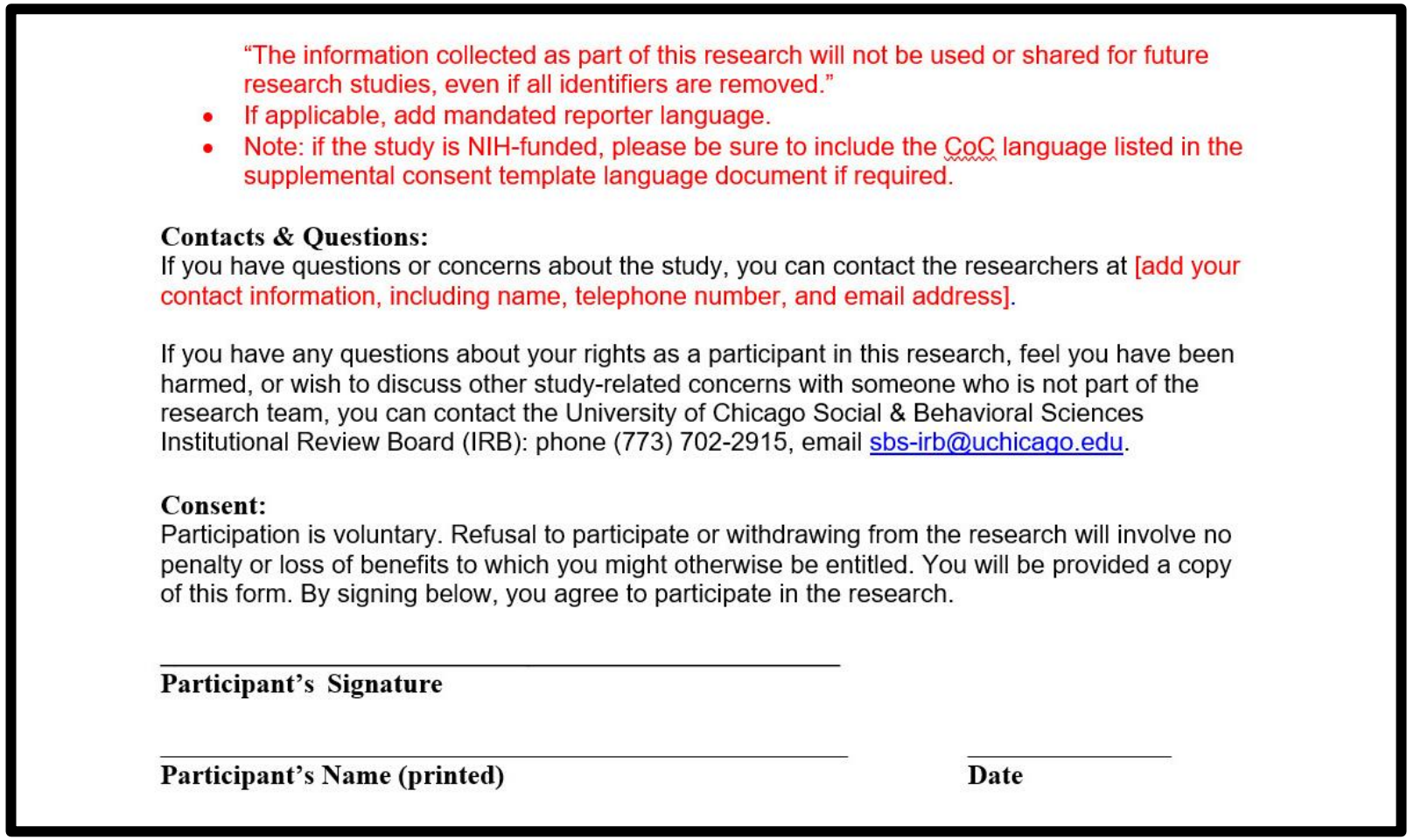
If applicable, explain any alternatives to participation, especially if research involves a program, treatment, or therapy.

Confidentiality: Describe how data, recordings, identifiers (if any), etc. will be used, shared, and protected during the research, as well as their retention, use, or disposition following the conclusion of the research.

- As applicable, address how partially collected data will be handled in case of a participant withdraws: e.g., "If you decide to withdraw from this study, the researchers will ask you if the information already collected from you can be used," "If you decide to withdraw, data collected up until the point of withdrawal may still be included in analysis," or "If you decide to withdraw from this study, any data already collected will be destroyed."
- Indicate whether identifiable data may be shared for future research, e.g., "Identifiable information will be handled as described in the 'optional elements section' below," or "Identifiable data will never be shared outside the research team."
- Address whether de-identified data may be shared for future research, e.g., "De-identified information from this study may be used for future research studies or shared with other researchers for future research without your additional informed consent," or,

Informed Consent Templates

https://sbsirb.uchicago.edu/templates/



"The information collected as part of this research will not be used or shared for future research studies, even if all identifiers are removed."

- If applicable, add mandated reporter language.
- Note: if the study is NIH-funded, please be sure to include the CoC language listed in the supplemental consent template language document if required.

Contacts & Questions:
If you have questions or concerns about the study, you can contact the researchers at [add your contact information, including name, telephone number, and email address].

If you have any questions about your rights as a participant in this research, feel you have been harmed, or wish to discuss other study-related concerns with someone who is not part of the research team, you can contact the University of Chicago Social & Behavioral Sciences Institutional Review Board (IRB): phone (773) 702-2915, email sbs-irb@uchicago.edu.

Consent:
Participation is voluntary. Refusal to participate or withdrawing from the research will involve no penalty or loss of benefits to which you might otherwise be entitled. You will be provided a copy of this form. By signing below, you agree to participate in the research.

_____
Participant's Signature

_____ _____
Participant's Name (printed) **Date**

Survey Design

Overall survey considerations

- How do we distribute it?
- How long should it be?
- One-time survey? Longitudinal survey?
- Will you use personalized data?
- What will participants learn?
  - What can we randomize to minimize this?
- Can we **randomize** the questions / answer choices?

Are all answer options covered?

- With whom do you regularly share Facebook posts?
  - Family
  - Friends
- Allow multiple answers?
- Include “other” option (write-in)?
- Do we care about previous use?

Are all answer options covered?

- I connect to Facebook over HTTPS
  - True
  - False
- What about “I don’t know”?

Are we biasing the answer?

- Strangers seeing your Facebook posts would cause you grave privacy concern.
  - Strongly agree
  - Agree
  - Neither agree nor disagree
  - Disagree
  - Strongly disagree

How will responses be distributed?

- For how long have you had Facebook?
  - Less than one day
  - Between one day and one week
  - More than one week

Should we force an answer?

- What gender are you? (* required)
  -Female -Male
- What gender are you?
  -Female -Male -I prefer not to answer
- With what gender do you identify?
  -Female
  -Male
  -Non-binary
  -I prefer to self-describe___
  -I prefer not to answer

Likert-scale data?

- Respond to the following statement: Companies collect too much private data.
  - 7: Strongly agree
  - 6: Agree
  - 5: Somewhat agree
  - 4: Neutral
  - 3: Somewhat disagree
  - 2: Disagree
  - 1: Strongly disagree

Likert-scale data?

- I feel that companies collect too much private data.
  - 7: Strongly agree
  - 6: Agree
  - 5: Somewhat agree
  - 4: Neutral
  - 3: Somewhat disagree
  - 2: Disagree
  - 1: Strongly disagree

What demographics do we collect?

- Tech expertise, age, domain knowledge, gender, location, employment, etc.
- Don’t ask people to self-rate expertise
  - Ask questions with concrete answers
  - e.g., Have you earned a degree in, or held a job in, computer science?
  - Include a knowledge test if you want to know about expertise
- Consider why you are collecting this info