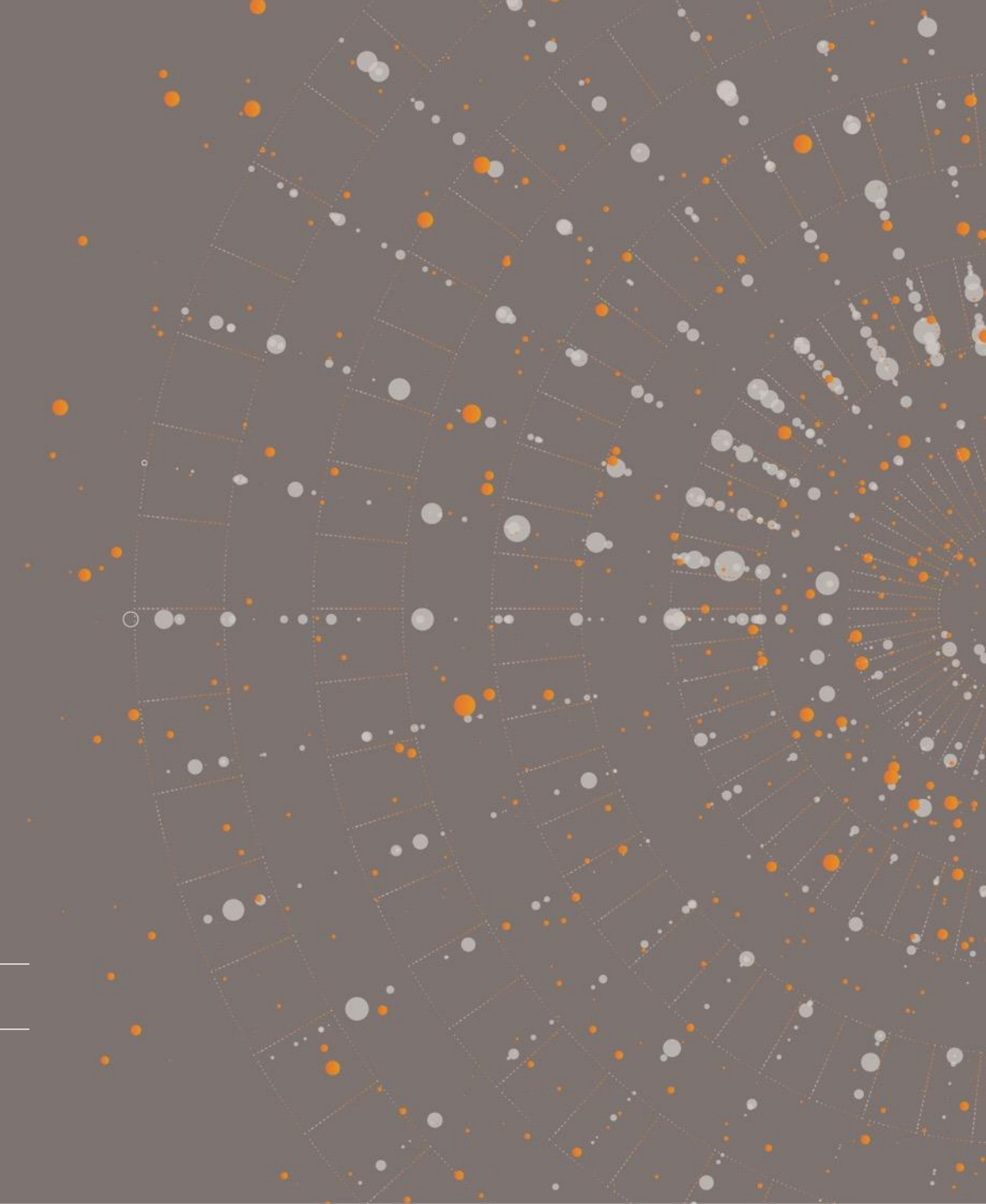


Generative AI for Surveys

Prioritizing User and Respondent Experience

04.22.2024

Elizabeth Dean, Brandon Sepulvado, Soubhik Barari, Zoe Slowinski, Ting Yan



Artificial Intelligence (AI)

An Overview

What do we mean by AI?

National AI Initiative Act of 2020

“A machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations or decisions influencing real or virtual environments”

AI Examples

- ChatGPT, Claude, Llama, Copilot
- Features in software, e.g., Grammarly, Nvivo

Examples that are not AI (but are closely related and might be part of AI systems)

- Topic model
- OLS regression

There are several types of AI work relating to surveys.

Productivity and Efficiency

**Making everyday
tasks easier and
quicker**

Examples

- Literature reviews
- Drafting documents and presentations

Research Using AI

**Using AI to
conduct research
and build tools**

Examples

- Leveraging text and audio data
- Monitoring data & question quality

Research About AI

**Understanding
how AI impacts
domains**

Examples

- Use of AI in STEM graduate programs
- How teens use AI

Evaluating AI Systems

**Uncovering LLM
properties and
limitations**

Examples

- Bias
- Feasibility assessments

User Experience

**Investigating how
AI impacts users
& participants**

Examples

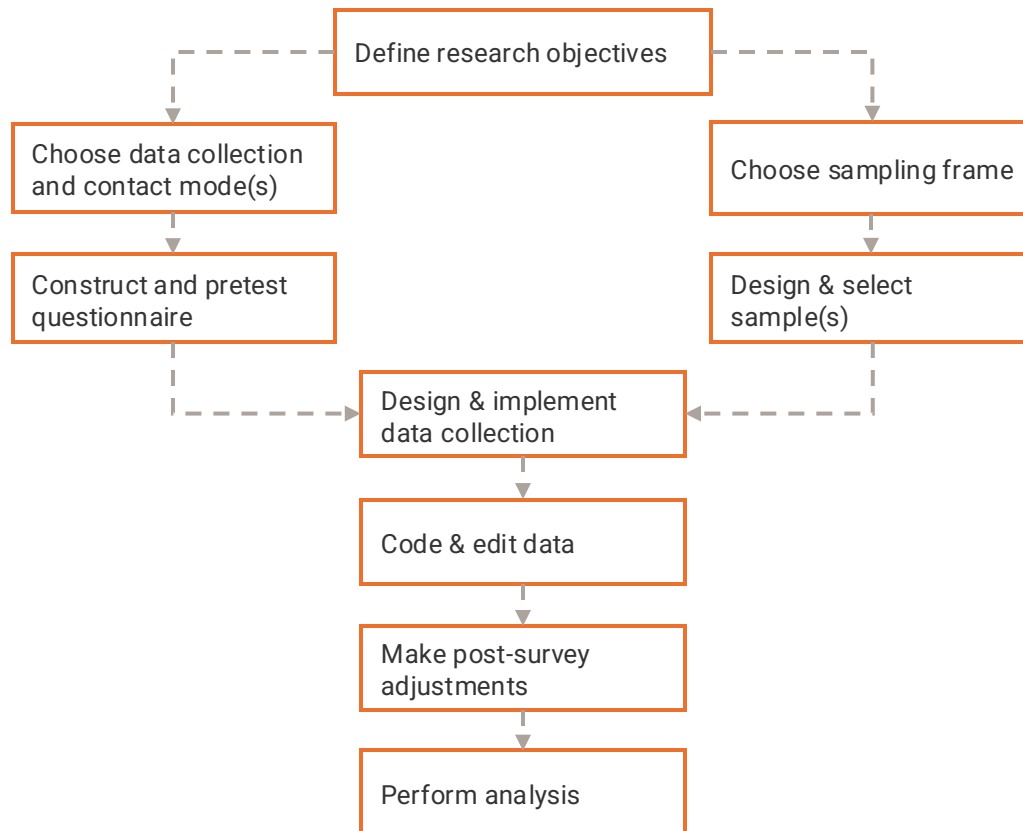
- How chatbots affect the survey experience
- AI agents for participant support

AI across the Survey Lifecycle

Overview and Examples

AI fits every stage of the survey process

An Overview of the Survey Lifecycle



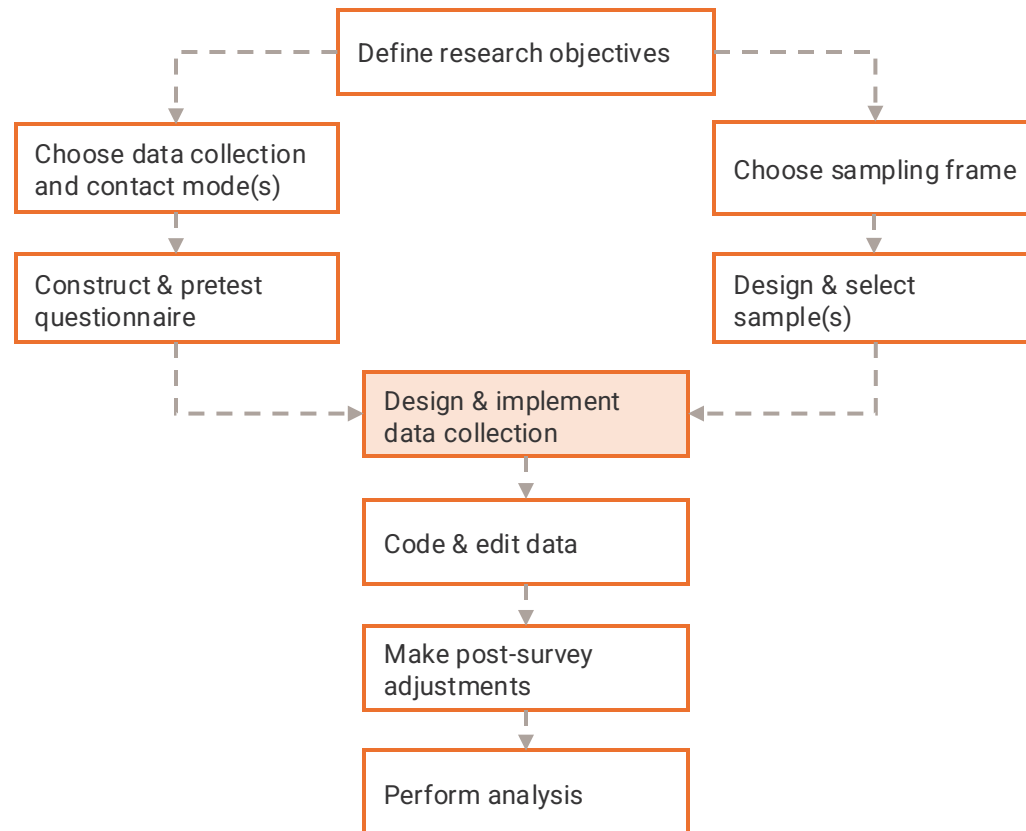
A Methodologist's Approach

- Use the survey lifecycle framework to guide how we think about implementing AI
- Helps identify gaps
- Helps determine scope of AI tools

Chatbots for Data Collection

Can AI agents improve data quality and improve the respondent experience?

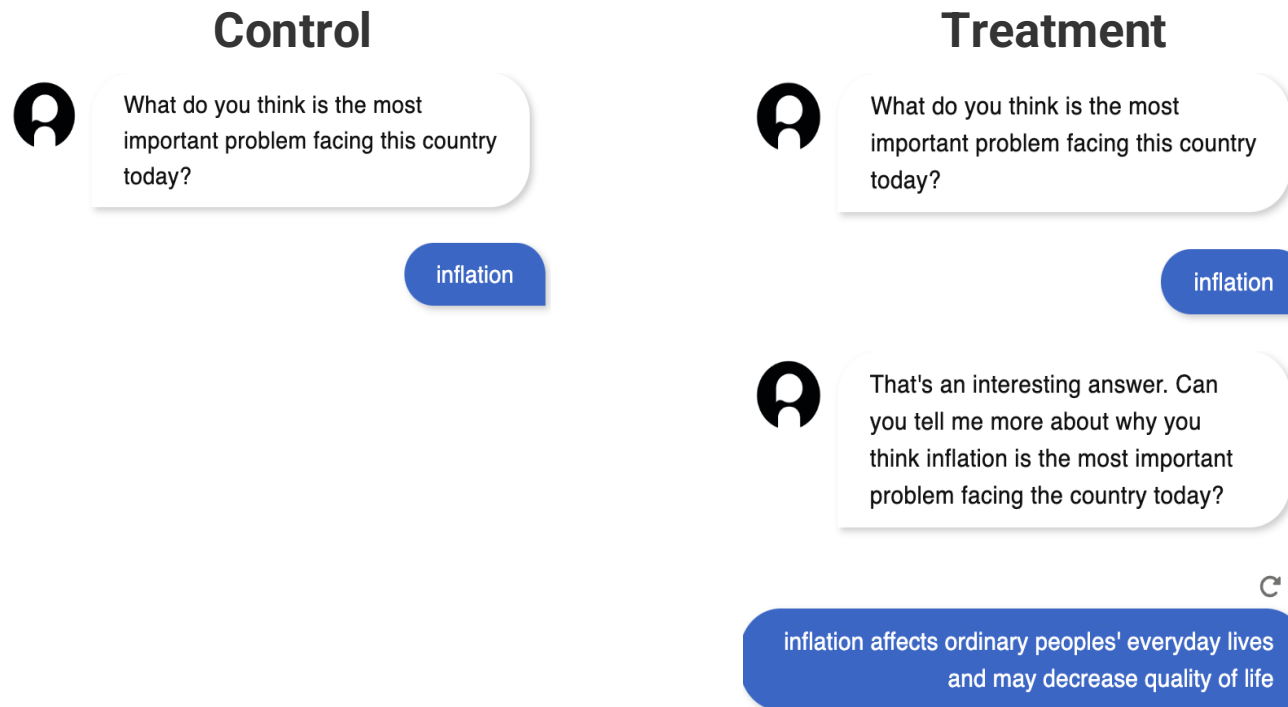
An Overview of the Survey Lifecycle



Study

- Conduct a pilot experiment to compare using chatbots across a range of question topics and types
- Promising results
 - Up to 96% accuracy across a range of questions
 - More specific and explanatory responses
 - Minimal to no impact on user experience depending upon type of automated probe

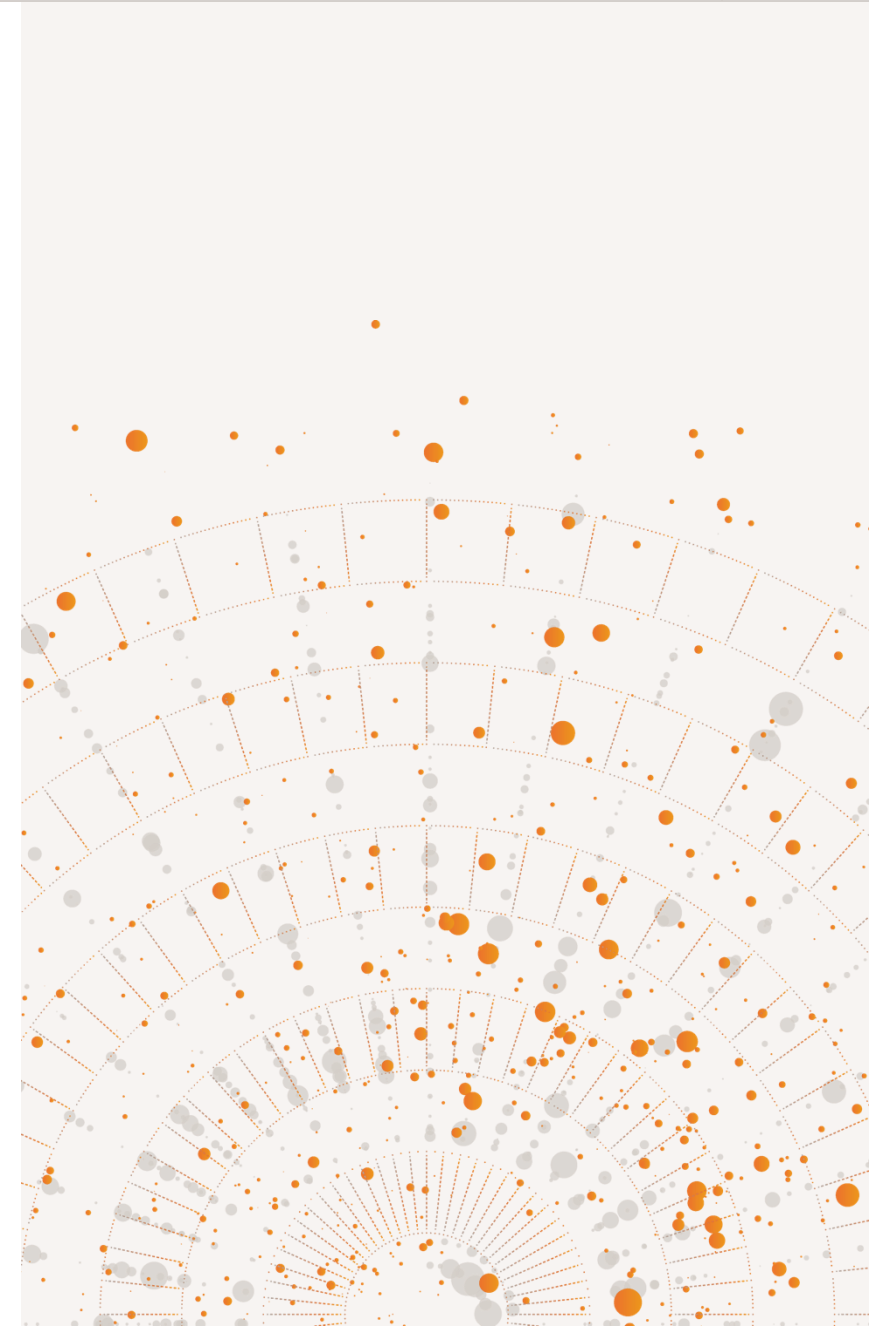
Can AI agents improve data quality and improve the respondent experience?



Respondent experience was captured in measures of burden and satisfaction.

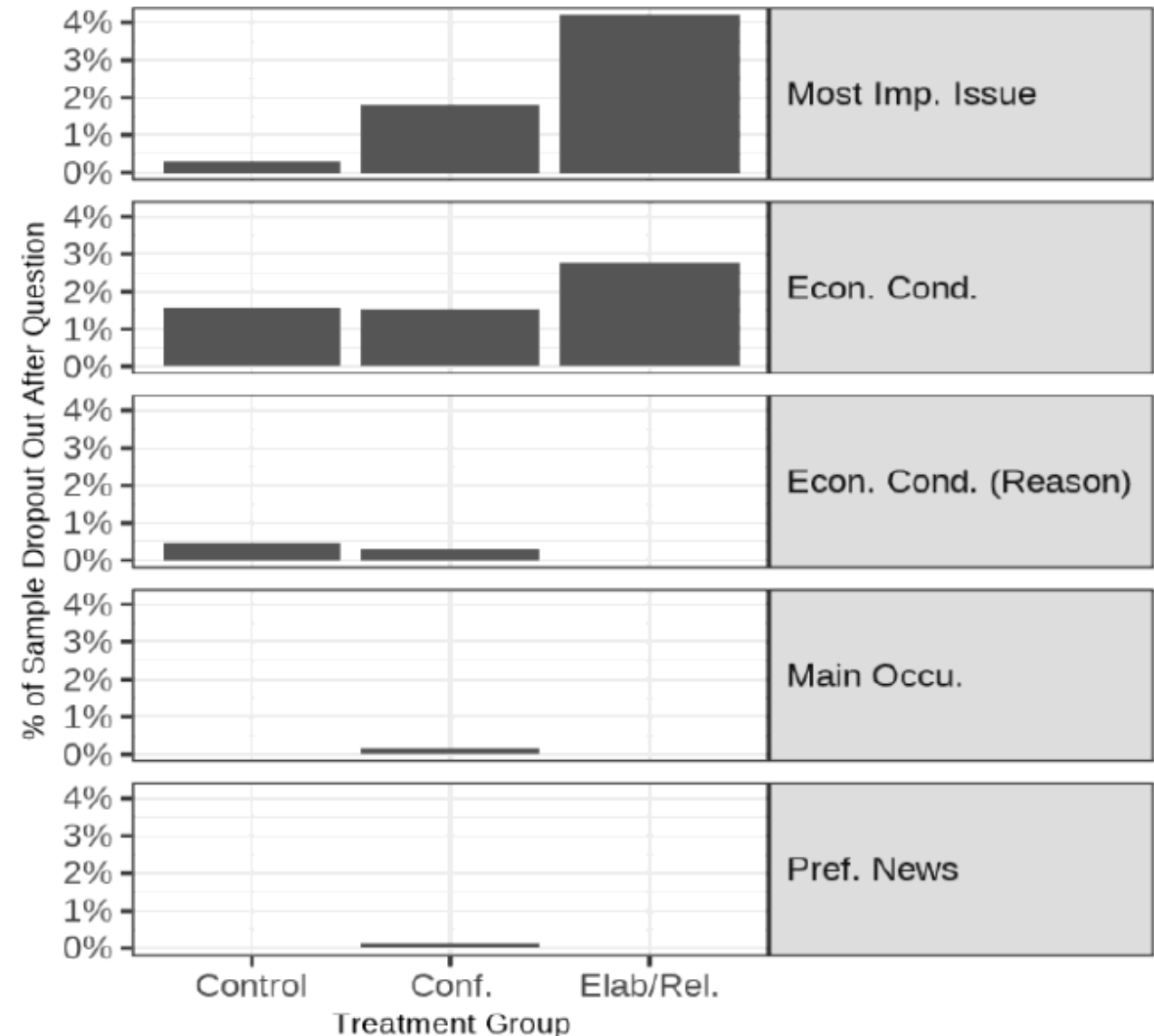
UX Metrics

- Breakoffs
- Self-reported experience measures:
 - Quality of survey responses
 - Ease of completion
 - Frustration
 - Overall satisfaction

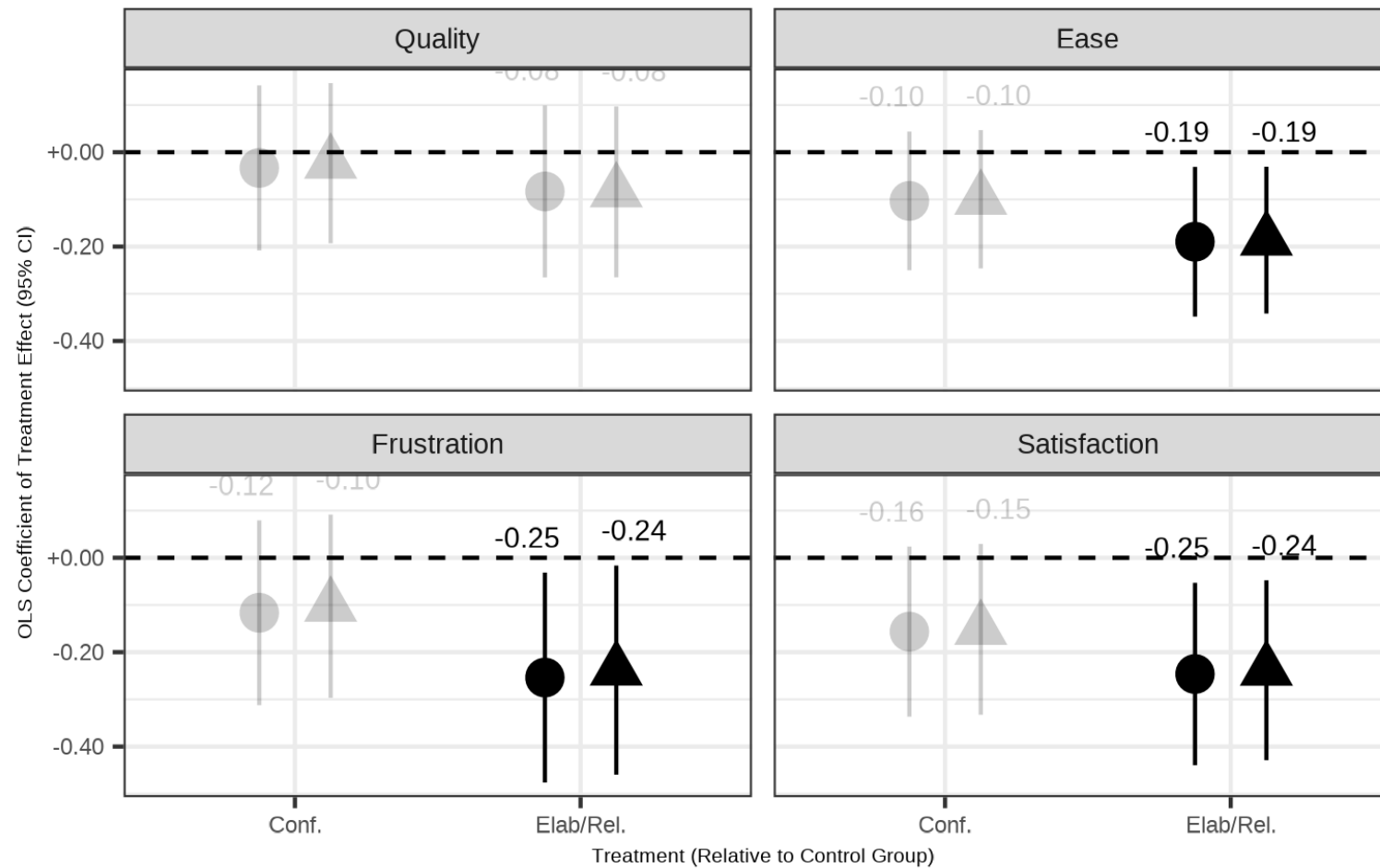


Breakoff is concentrated in **early questions** and among **mobile** respondents

- Q1 most important issue
- Q2 perception of economic conditions
- 1-4% of sample per question
- Respondents were more likely to drop out after the elaboration and relevance probes than the confirmation probes.
- Almost all of these breakoffs were on mobile



Probing **slightly** increases self-reported frustration, decreases satisfaction



P-Value: ● p < 0.05 ○ p >= 0.05 Specification: ● No Controls ▲ Controls

Chatbots can provide useful survey data with minimal respondent burden.

Findings

- Probes yielded accurate codes and more detailed, better-quality responses.
- Chatbot surveys are minimally burdensome
 - They are worse for mobile respondents
- Respondents reported slightly more negative responses to ease of use, frustration, and satisfaction measures

Basic UX Principles for Chatbots

1. **Tell respondents what you're doing and why.**

- “For some questions, this survey might use a chatbot to make answering the questions easier for you and to help us get a more accurate understanding of your experience.”

2. **Limit usage of chatbots to a subset of questions**

3. **Use buttons; chatbot questions don't have to be open-ended**

4. **Use a chat format when a conversation makes sense**

- True open-ended questions
- Items where a conversational approach helps recall
 - Medical history, residential history, pregnancy and childbirth

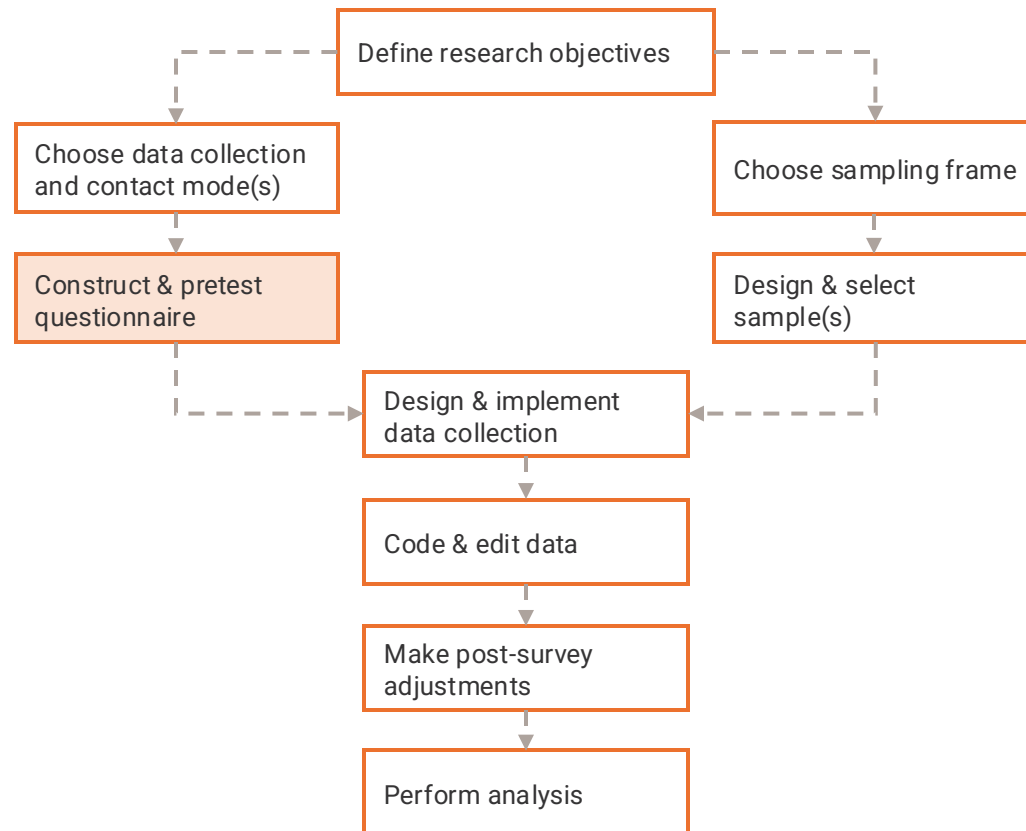
5. **Optimize for mobile users**

- Buttons, voice entry

Generative AI for Question Appraisal

Can AI help methodologists more quickly and cost effectively design high quality survey questions?

An Overview of the Survey Lifecycle



Study

- Testing the quality of questionnaire reviews conducted by chat-based generative AI tools (ChatGPT, Copilot)
- Used the Question Appraisal System (QAS-99) to guide the review.

The Questionnaire Appraisal System (1999) evaluates each survey item for issues in 7 categories:

- 1. Reading
- 2. Instructions
- 3. Clarity
- 4. Assumptions
- 5. Knowledge/Memory
- 6. Response Categories
- 7. Sensitivity/Bias

QAS form for Knowledge/Memory section

STEP 5 - KNOWLEDGE/MEMORY: Check whether respondents are likely to <i>not know</i> or have trouble <i>remembering</i> information.	
5a. KNOWLEDGE may not exist: Respondent is unlikely to <i>know</i> the answer to a factual question.	YES NO
5b. ATTITUDE may not exist: Respondent is unlikely to have formed the attitude being asked about.	YES NO
5c. RECALL failure: Respondent may not <i>remember</i> the information asked for.	YES NO
5d. COMPUTATION problem: The question requires a difficult mental calculation.	YES NO

Prompt 1:

“Using the QAS-99 approach linked here: (ResearchGate publication link), please conduct an appraisal of each of the following survey questions. Ask me any additional questions you need to refine your appraisal.”

Response from ChatGPT included:

“Follow-up Questions for You:

What’s the target population for these questions (general population, clinical sample, specific age group)?

Mode of administration: Online, phone, paper, in-person interview?

Are any of these questions required to align with previously validated instruments, or can they be modified freely?

Would you like me to suggest revised versions of any questions based on this analysis?

Let me know and I’ll tailor the next step!”

Prompt 2:

“The target population is general United States adult population. Surveys will be conducted in English and Spanish. It will be a self-administered web survey. Questions do not need to align with previously validated instruments, but please show any similarities to existing validated instruments. Please conduct your QAS review using the clarification I provided above. And yes, please suggest revised versions of each question where necessary.”

Results

- While ChatGPT successfully identified question problems in the spirit of the QAS, it did not actually code the categories according to the taxonomy.
- The categories it named were not always consistent with the QAS, or even within ChatGPT's own responses:
 - Clarity (A) ☒
 - Assumptions (B) ☒
 - Reference Period (C), Recall/Reference Period (C) ☐
 - Mapping Answers to Questions (F) ☐
 - Response Categories (G) ☒
 - Sensitivity/Social Desirability (H), Sensitivity (H), Social Desirability (H) ☐
 - **(Categories D, E, and I were not specified)** ☐

Actual QAS categories:

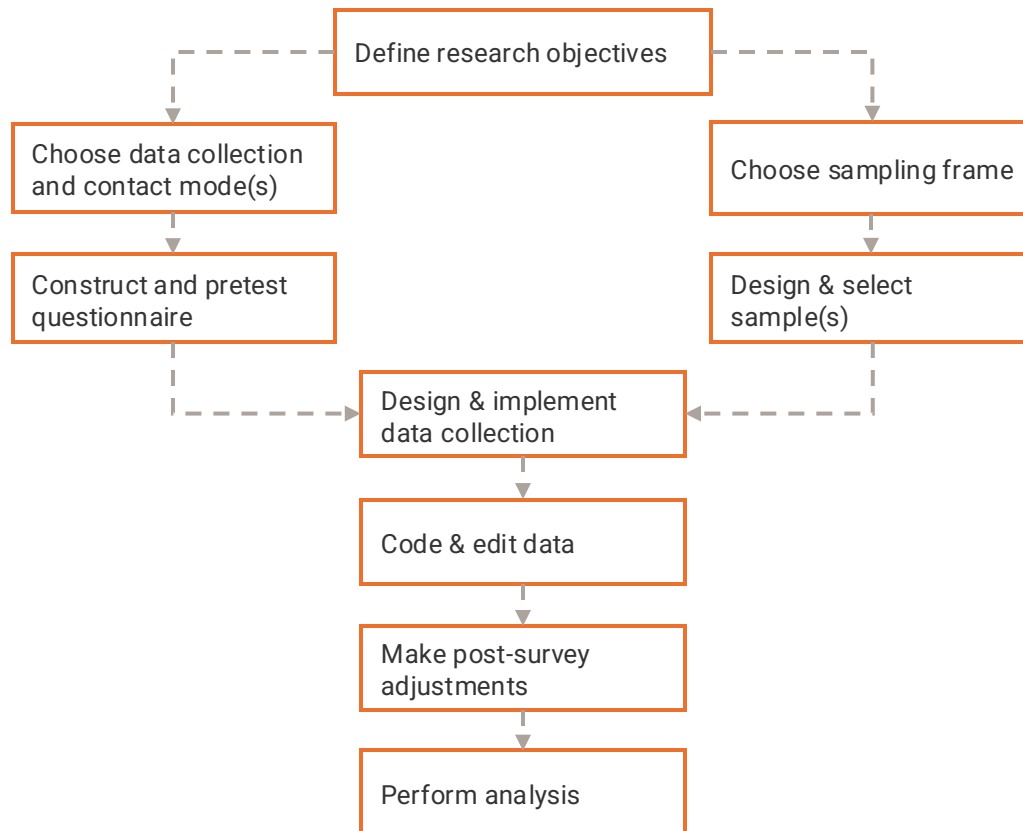
1. Reading
2. Instructions
3. Clarity
4. Assumptions
5. Knowledge/Memory
6. Response Categories
7. Sensitivity/Bias

UX and Using Generative AI for Question Appraisal

1. **We the researchers are the users.**
2. **Using generative AI for questionnaire design, evaluation, and testing is inevitable. How can we ensure it generates more usable results?**
 - Prompt it better to get it to deliver actionable results?
 - Adapt our methodologies?
 - Always have a human comparison?
 - Wait?
3. **Treat it more like a research assistant and less like a programmable tool.**
4. **Adapt our methodologies for AI research assistance**

What's next?

An Overview of the Survey Lifecycle



A User-Centered Approach

- User personas of AI enhancements depend on the phase of the survey lifecycle
- Consider UX from the respondent, questionnaire designer, data collector, analyst, etc. perspectives

Thank you.

Elizabeth Dean
Principal Research
Methodologist
Dean-elizabeth1@norc.org

 Research You Can Trust™

 **NORC** at the
University of
Chicago