Risks and Rewards: How Respondents Balance (or Don't) Concerns about Privacy, Benefits of Accurate Data, and the Possibility of Privacy Loss

Casey Eggleston,

Aleia Clark Fobia and Jennifer Hunter Childs

Center for Behavioral Science Methods, U.S. Census Bureau

United States Census Bureau

U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

"Any views expressed are those of the author(s) and not necessarily those of the U.S. Census Bureau."

What is re-identification?

- Re-identification is a risk for publicly released data
  - **Data re-identification** is the practice of matching anonymous data (also known as de-identified data) with publicly available information, or auxiliary data, in order to discover the individual to which the data belong to.
- Existing practices are already designed to address such risks
  - E.g., Census Bureau Statistical Quality Standard S1 which recommends practices such as top-coding, cell suppression, and noise infusion to protect confidentiality of publicly-released data
- However, advances in availability of data and computing power pose a challenge to traditional techniques

What is differential privacy? (Nissim & Wood, 2017)

- All analyses of personal data inevitably "leak" some amount of identifying information, and the risk accumulates with additional releases or analyses
- Differential privacy is a way of formally defining privacy that has emerged from the theoretical computer science literature
  - The goal of differential privacy is to ensure that "Any information-related risk to a person should not change significantly as a result of that person's information being included, or not, in the analysis."
  - Differential privacy makes it possible to quantify maximum privacy loss and to measure the cumulative risk of multiple computations/releases
  - Differential privacy is attained by adding "*carefully crafted* random noise" into computations, but there are many different methods/algorithms that can be used

Defining “acceptable” privacy loss
(Abowd & Schmutte, 2015)

- Applying differential privacy techniques requires the data curators to make decisions about the tradeoff between data accuracy and data privacy
  - These decisions cannot be determined from the data themselves, they are essentially policy questions.
- ε (epsilon) sets a value for “worst-case” privacy loss or “leakage”
  - It “can be treated as a ‘privacy budget’ which is consumed as analyses are performed” (Nissim & Wood, 2017)
- How can a data owner decide what level of privacy loss is “acceptable”?

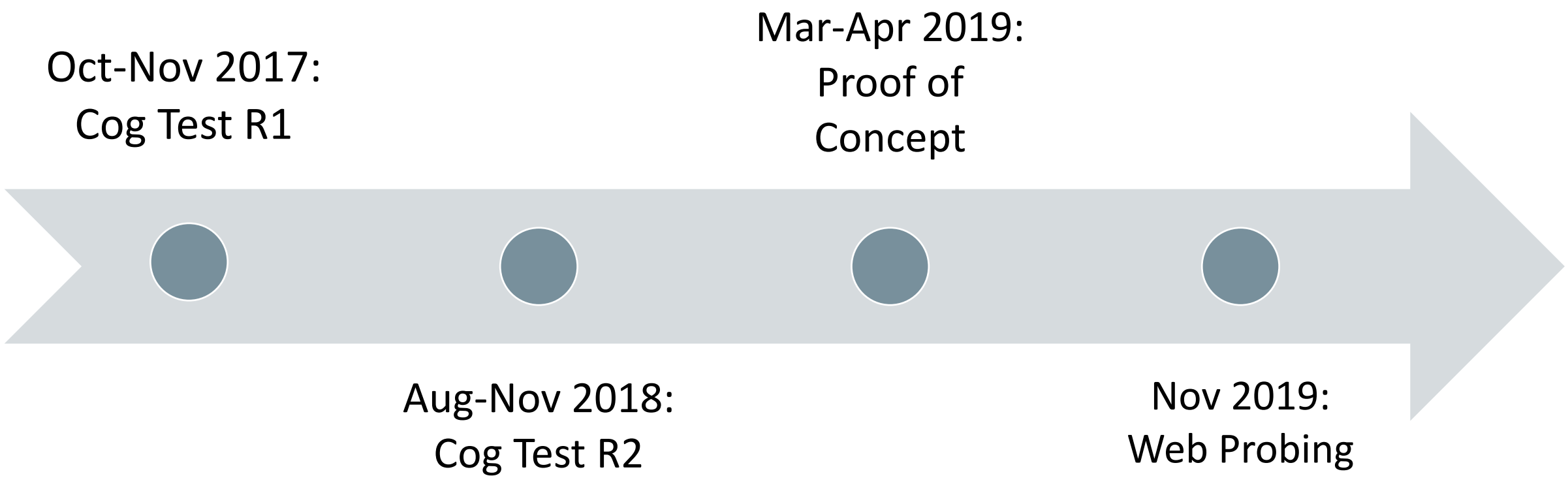
The Research Challenge

- The Census Bureau plans to apply a differential privacy system to 2020 Census data releases, but there are many considerations
- The Center for Behavioral Science Methods was tasked with tackling the policy questions surrounding ε and the privacy loss budget
  - How might respondents value the confidentiality of their census data?
  - Are respondents worried about re-identification?
  - Do respondents prefer more privacy at the cost of less accuracy of publicly released data or are they willing to risk privacy for more accurate and useful data?
- The terms and concepts are familiar to economists, data scientists, and survey researchers but are not something respondents have had to think about

Relevance to Health Context

- Compared to census data, potential risks and benefits are higher (and generally easier to explain) for a health context, especially as regards research on treatment for specific conditions
- Health information provides an excellent context for understanding individual privacy concerns, comprehension of and perceived risks associated with re-identification, and privacy/accuracy tradeoff preferences
- Concepts of privacy and risk of identification are deeply embedded in laws that govern the protection of health information and the ethical standards and principles of medical research
- Health researchers have asked similar questions to those we are now asking for the census context (e.g., Clerkin et al., 2013; O'Brien et al., 2020)

Research Timeline



Methods

- Goal: Large, nationally representative sample survey
- In-person cognitive testing, think-aloud with intermittent probing
  - Round 1: 27 interviews
  - Round 2: 17 interviews
- Proof of concept field test
  - Qualtrics instrument
  - Sample: 20,000 households, up to 3 emails per household, randomized national sample
  - 727 responses after cleaning
  - Half of sample had web probes

Data has been approved for release by the Census Bureau's Disclosure Avoidance Review Board (CBDRB-FY19-CED002-B0003; CBDRB-FY19-CED001-B0015.).

Methods (cont)

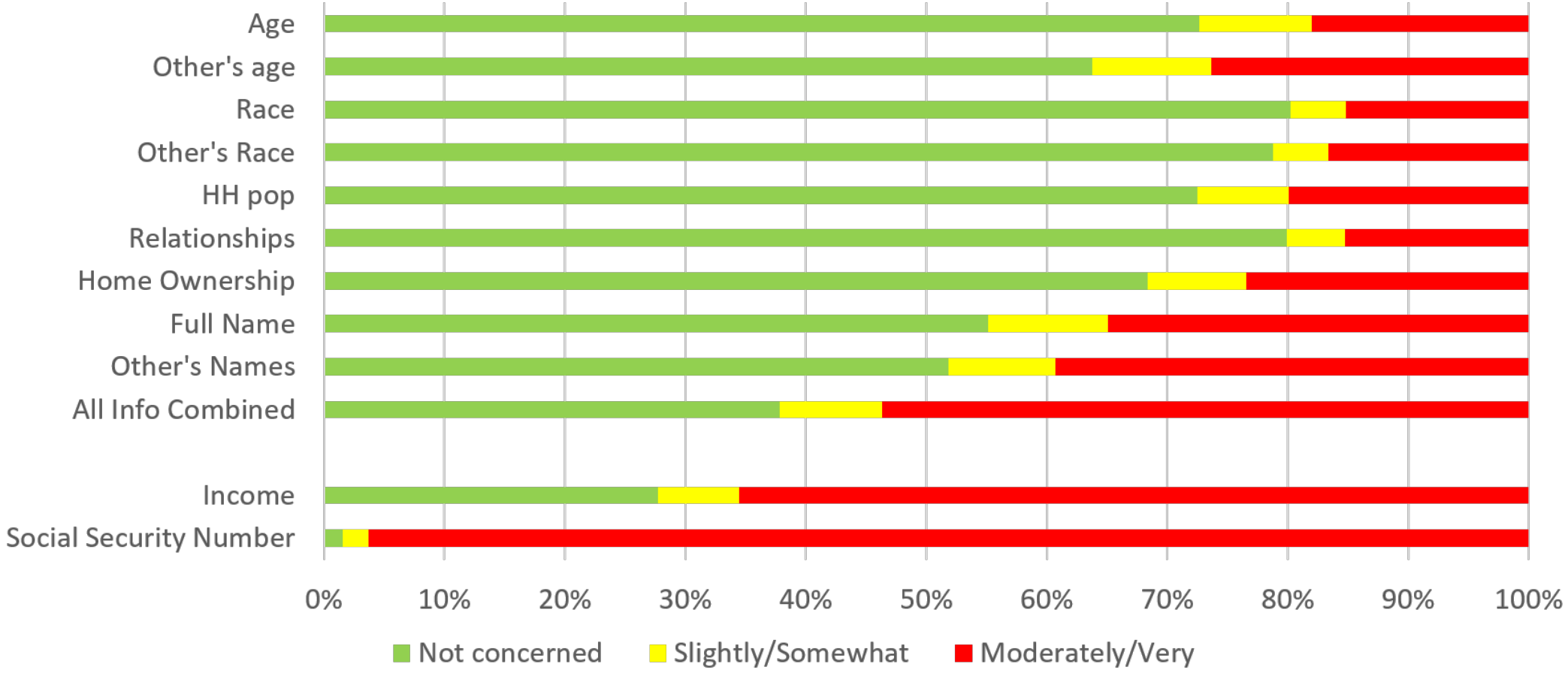
- Online Web Probing
  - Qualtrics instrument
  - Sample: 5,000 email addresses from an affinity panel maintained by the Census Bureau
  - Around 200 complete responses collected
  - Randomly presented one of 3 different versions of re-identification questions

Data has been approved for release by the Census Bureau's Disclosure Avoidance Review Board (CBDRB-FY20-174).

Questionnaire Development Process

1. Assumed we needed to ask the decennial items for context.
2. Realized we could ask concern questions with very minimal background/context. Decided to start with a binary concern item to minimize over-reporting of concern, then a separate question about degree.
3. Then moved on to the difficult subject of re-identification.
4. Needed the questionnaire to acknowledge hacking/data breach before re-identification because that is the more familiar and concerning issue to respondents. Designed questions to address this.
5. R1 cognitive testing showed minimal issues with concern items. Comprehension issues with re-identification.
6. Refined re-identification definition and example in R2 cognitive interviews. Added new items to get at privacy/accuracy tradeoff.
7. Re-identification still problematic. Designed 3 different versions using alternative examples to explain the concept. Tested further in web probing.
8. Currently coding web probing results to choose best of re-identification question sets and decide whether to drop any privacy/accuracy tradeoff items.

Level of Concern about Census Items (Proof of Concept Survey)



Data has been approved for release by the Census Bureau's Disclosure Avoidance Review Board (CBDRB-FY19-CED002-B0003; CBDRB-FY19-CED001-B0015.).

Comparison with Web Probing Survey

- Main point of web probing survey was to evaluate options for asking about re-identification, but also replicated some of the same items from the proof of concept.
- Opportunity to add questions relevant to the health context:

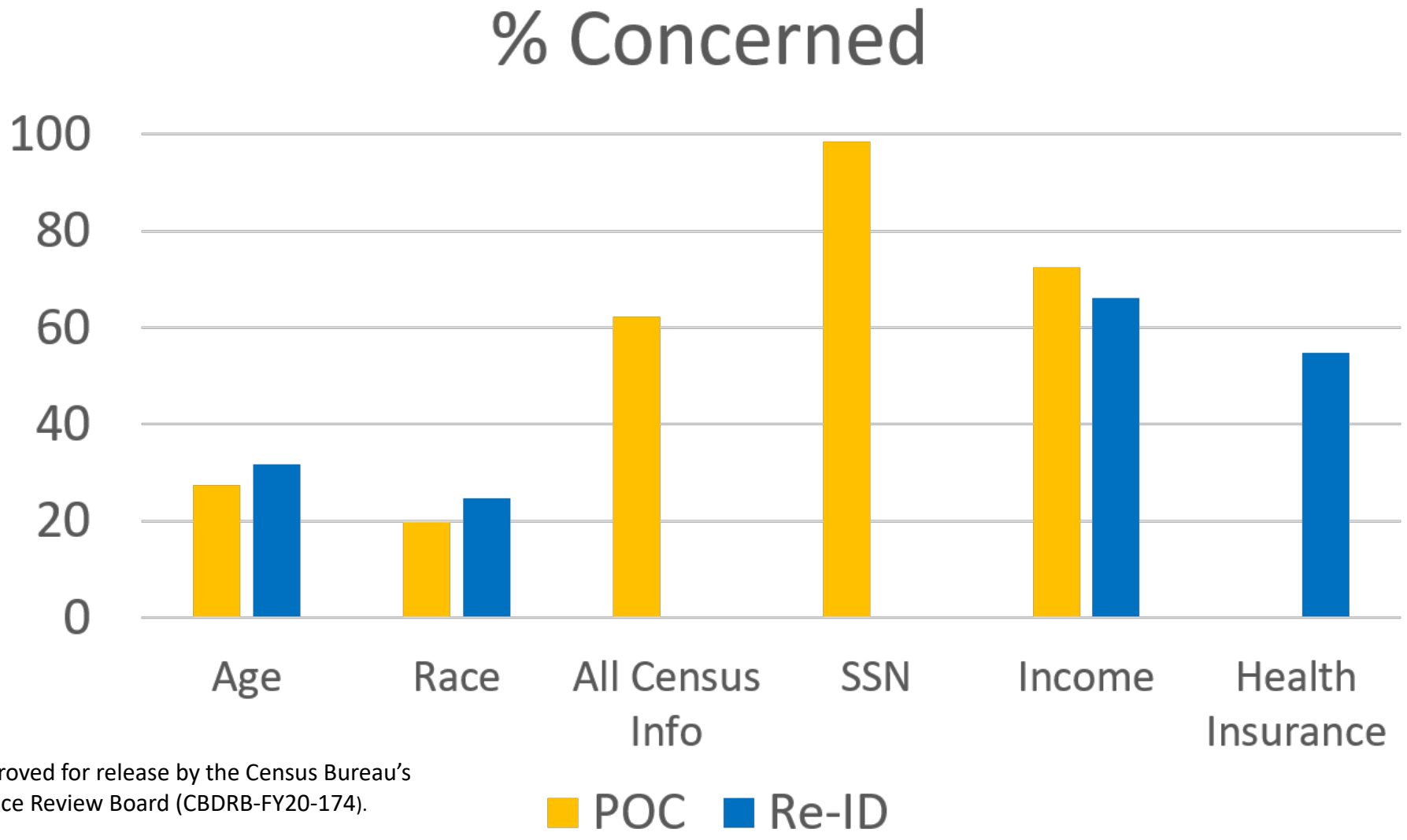
Would it concern you if someone was able to find out **whether or not you currently have any kind of health insurance or health care coverage**?

- ○ Yes
- ○ No

If someone was able to find out **whether or not you currently have any kind of health insurance or health care coverage**, how concerned would you be?

- ○ Slightly concerned
- ○ Somewhat concerned
- ○ Moderately concerned
- ○ Very concerned

Level of Concern (Comparison)



Data has been approved for release by the Census Bureau's Disclosure Avoidance Review Board (CBDRB-FY20-174).

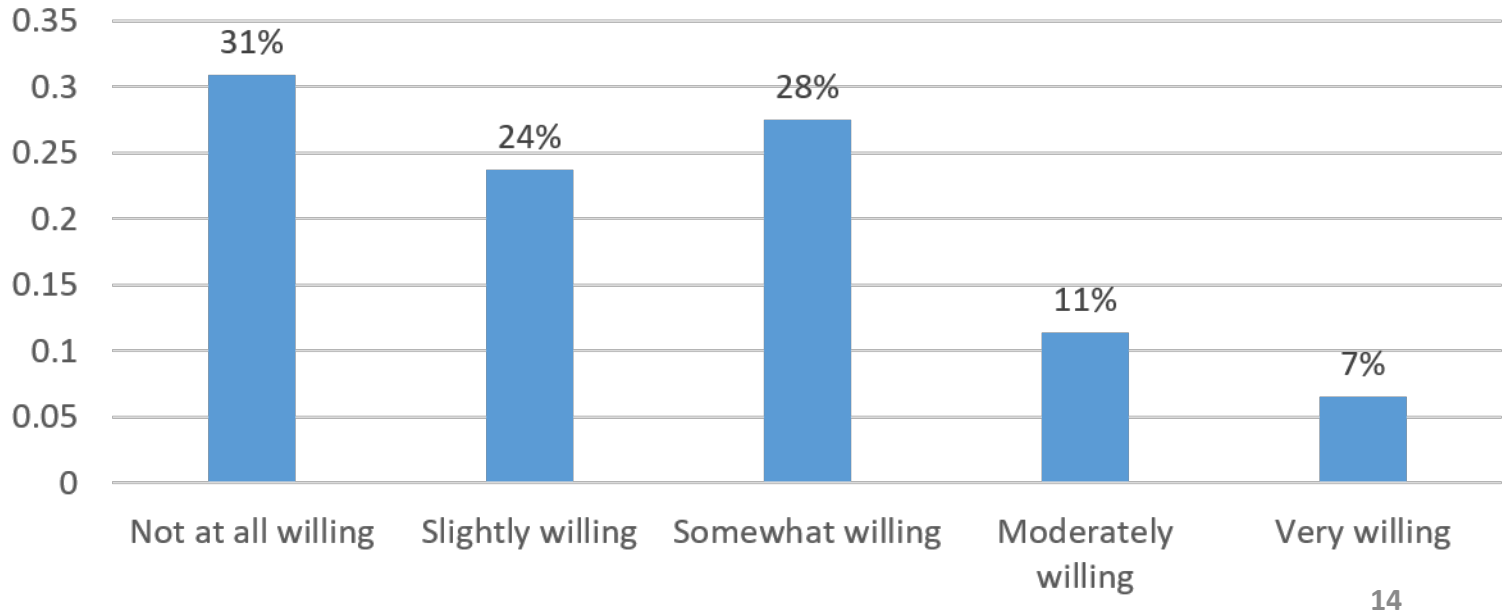
Privacy-Accuracy Trade-Off Questions

Policy makers, businesses, and researchers use information collected from government surveys to make important decisions. The more detailed the data provided by households like yours, the more useful that information is. This might mean reporting data by ZIP code instead of by state. But providing more detail may increase the risk that an individual household's information will be identified, even if that risk is low.

Data has been approved for release by the Census Bureau's Disclosure Avoidance Review Board (CBDRB-FY19-CED002-B0003; CBDRB-FY19-CED001-B0015.).

In general, how willing are you to risk your confidentiality so the government can produce useful data and statistics for policy makers, businesses and researchers to use?

- ☐ Not at all willing
- ☐ Slightly willing
- ☐ Somewhat willing
- ☐ Moderately willing
- ☐ Very willing



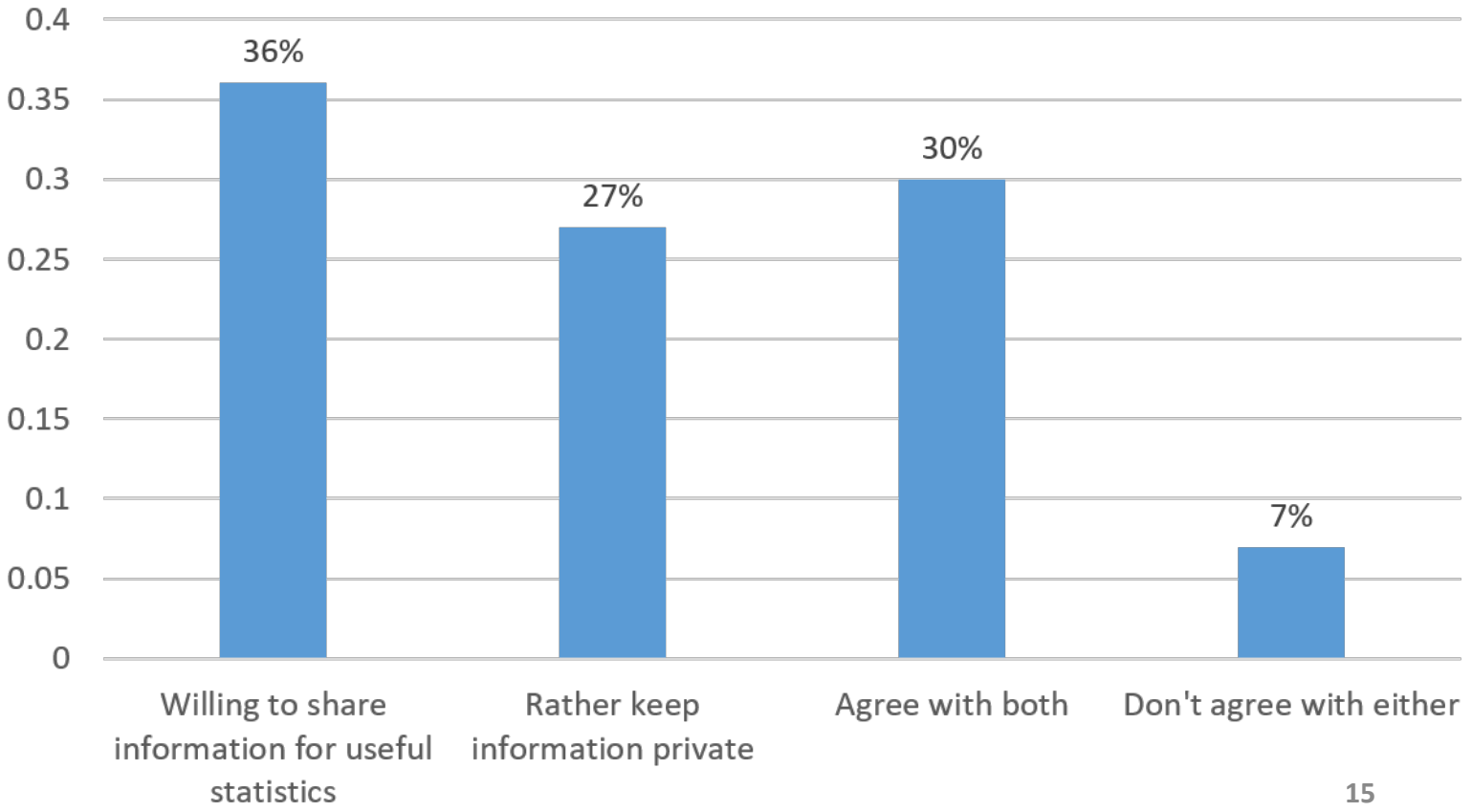
☐ A. I am willing to share information about me and my household with some government agencies (like the Census Bureau) so the government can produce more useful data and statistics, even if it means having less control over that information.

☐ B. I would rather keep information about me and my household private even if it means the data and statistics produced by the government are less useful.

☐ C. I agree equally with both

☐ D. I don't agree with either

Data has been approved for release by the Census Bureau's Disclosure Avoidance Review Board (CBDRB-FY19-CED002-B0003; CBDRB-FY19-CED001-B0015.).



Refining Re-Identification Questions

- Version 1: No Example

Though hacking and data breaches have received a lot of media attention lately, they are not the only way that the privacy of your information is at risk. When governments or other institutions release data, they remove identifying information such as your name, address, and birthdate. However, it could be possible for someone to combine the anonymous data with another information source and match the information with its true owner. If this happens, your private information may be identified.

Refining Re-Identification Questions

- Version 2: Taxi Cab Example

Basic definition +

For example, the City of New York released an anonymous database of all taxicab rides in a year. Using public photos of people climbing into cabs, someone was able to match individual people with times, locations, fares, and tips for specific rides.

Refining Re-Identification Questions

- Version 3: Census Example

Basic definition +

For example, someone could combine Census data about a small geographic area with other publicly-available information and find out that a specific household on a particular block has seven people living in it, including three unrelated people and two adopted children.

Discussion

- Respondents vary a lot in privacy concern for different personal information
- The term “re-identification” is a problem
  - Respondents seemed to understand the behavior and definition but the term was confusing
  - In the process of coding the open-ended web probing responses defining re-identification in respondents’ own words to decide which of the 3 versions of our re-identification definition is understood best
- Next steps
  - Working on finalizing instrument and sample for larger, more representative study

Implications for Health Survey Research

- The conversation surrounding differential privacy and other formal privacy methods impacts everyone who releases public statistics, and health survey data is no exception
- The approach we have taken to understand and measure individual privacy/confidentiality concerns could easily be replicated for other types of data (such as private health information)
- Although it obviously poses policy challenges, the need to formally define a privacy loss budget in differentially private systems increases visibility of often-overlooked issues such as individual privacy concerns and variable sensitivity of personal information
- Our hope is that this research contributes positively to efforts of the survey research field to practice ethically, communicate transparently, and produce useful data

Questions?

Casey Eggleston (casey.m.eggleston@census.gov)