

The Privacy/Accuracy Tradeoff: Respondents' Perspective

Casey Eggleston, Jennifer Hunter Childs,
and Aleia Clark Fobia

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

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

Shape
your future
START HERE >

CBDRB-FY20-CBSM002-001

United States[®]
Census
2020

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?

Purpose

To measure individuals' privacy risk tolerance with the goal of informing the privacy loss budgets allowed in mathematical privacy models for decennial data releases.

Responses to the questionnaire items provide information about respondent concerns surrounding each of the different pieces of individual and household-level information collected on the 2020 Census and included in statistical summaries released by the Census Bureau.

The Research Challenge

Build a questionnaire that addresses:

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

Research Timeline

Oct-Nov 2017:
Cog Test R1

Mar-Apr 2019:
Proof of
Concept

June 2020:
National
Survey

Aug-Nov 2018:
Cog Test R2

Nov 2019:
Web
Probing

National Privacy Concerns Survey

Decennial census questions

Concern about census items

Hacking and Reidentification

Privacy-Accuracy Trade-Off

Privacy Behaviors

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

Initial Questionnaire Development

Ask decennial items for context

Orients respondents to the information we refer to later

Start with a binary concern item to minimize over-reporting of concern, then a separate question about degree of concern

Focused on publicly released content

Asked about social security, income, full name for comparison

Re-identification items

Needed the questionnaire to acknowledge hacking/data breach before re-identification because that is the more familiar and concerning issue to respondents

Methods: Questionnaire Testing

2 rounds of in-person cognitive testing, think-aloud with intermittent probing

Proof of concept field test, half of Rs received additional web probing

Sample: 20,000 households, up to 3 emails per household, randomized national sample, Contact Frame

~750 responses

Online Web Probing, comparing versions of re-identification questions

Sample: 5,000 email addresses from an affinity panel maintained by the Census Bureau

~200 complete responses collected

Questionnaire Revisions

R1 cognitive testing showed minimal issues with concern items. Comprehension issues with re-identification questions.

Refined re-identification definition and example in R2 cognitive interviews. Added new items to get at privacy/accuracy tradeoff.

Re-identification still problematic.

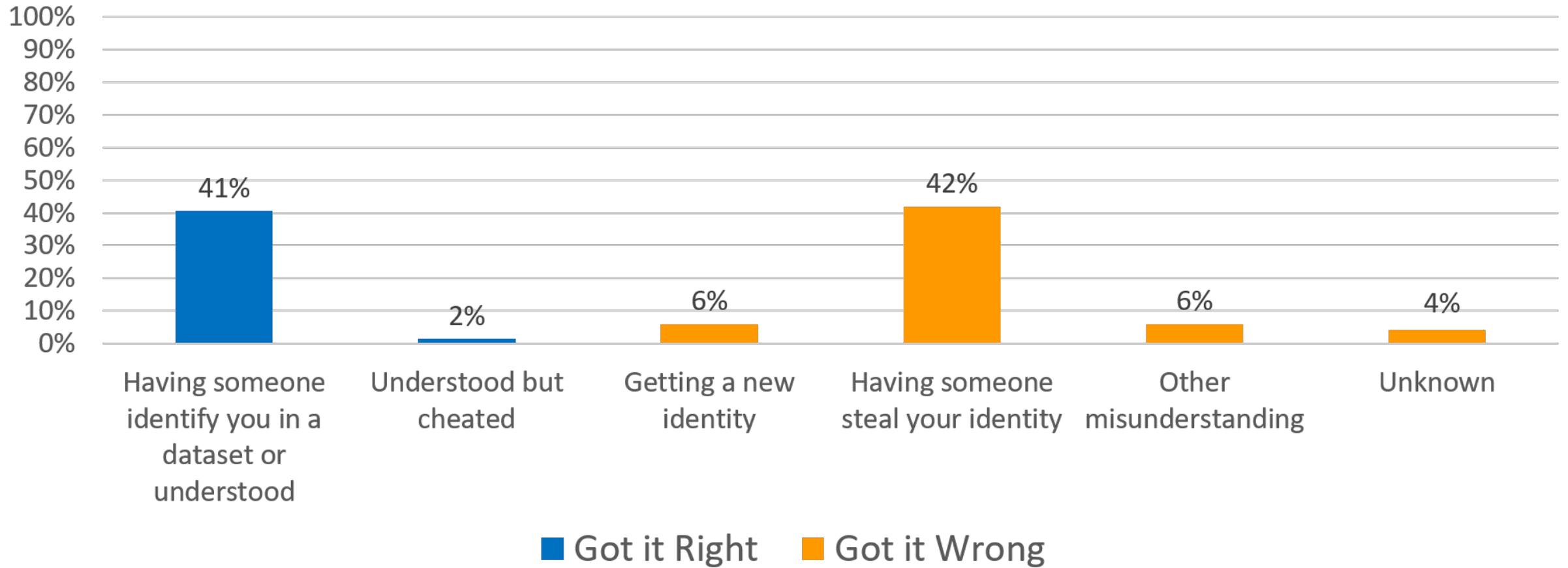
Pilot test web probing showed comprehension issues

Designed 3 different versions using alternative examples to explain the concept. Removed the term “re-identification”

Tested further in web probing and a small number of in-person interviews.

Coded web probing results to choose best of re-identification question sets for national survey.

Pattern of concept interpretation



Shape
your future
START HERE >

United States[®]
Census
2020

Final wording for national survey

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.

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.

Methods: National Survey

Sample: KnowledgePanel is recruited via postal mailing utilizing the USPS Computerized Delivery Sequence File. All US residential non-institutional addresses are eligible for selection. To enable non-internet households to respond to online surveys, Ipsos provides non-internet households a tablet with a mobile data plan. The sample was selected to generate 10,000 responses reflecting the national distribution of gender, age, race, Hispanic ethnicity, home ownership, education and income according to Census Bureau benchmarks.

Recruitment: Panel respondents received an email invitation and reminders as well as a notification of the survey on the panel member portal.

Survey Administration: The questionnaire was administered online only using the survey platform Qualtrics.

Demographics

Demographic	Census Benchmark (Target)	Unweighted Final Survey Sample	Weighted Final Survey Sample
Sex	51.7% Female	49.1% Female	52.0% Female
18-24	11.8%	4.8%	8.3%
25-34	17.9%	14.2%	19.0%
35-44	16.3%	14.0%	17.6%
45-54	16.7%	14.3%	14.3%
55-64	16.8%	22.5%	19.5%
65+	20.5%	30.3%	21.4%
White Only	77.9%	81.6%	74.3%
Black/AA Only	12.6%	9.6%	12.4%
Hispanic (any race)	16.2%	11.6%	16.0%
Home Ownership	68% Own	74.7% Own	70.5% Own

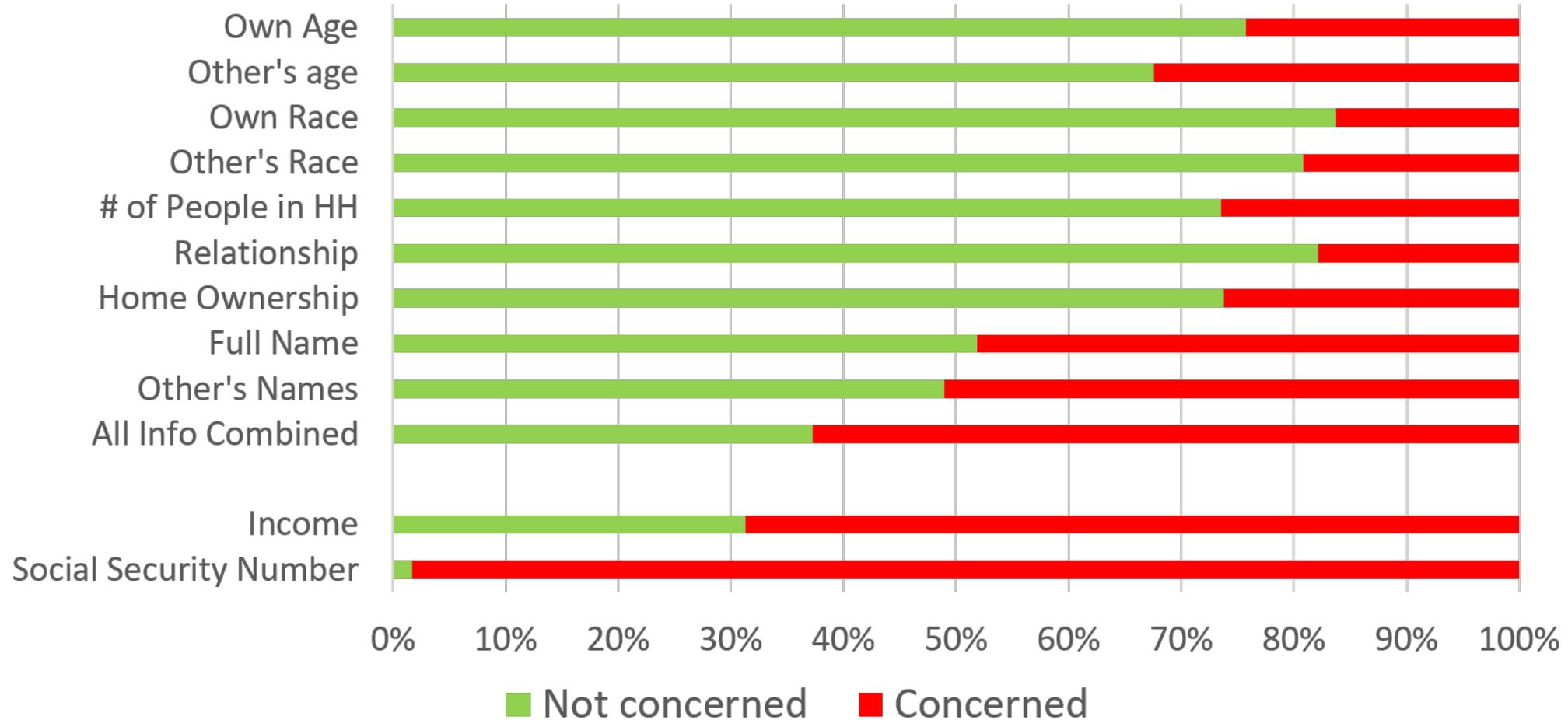
Source: Data from the National Privacy Concerns Survey (June 11th – June 29th, 2020) CBDRB-FY20-CBSM002-001.

Demographics

Demographic	Quota	Unweighted Final Survey Sample	Weighted Final Survey Sample
Less than HS	10.9%	4.9%	8.9%
HS or Equivalent	28.7%	24.2%	28.0%
Some college	28.1%	29.3%	28.3%
Bachelors or more	32.3%	41.6%	34.8%
Under \$25,000 <i>*Ranges for survey slightly different</i>	14.6%	13.3%	15.3%
\$25,00-\$49,999*	19.1%	17.8%	18.3%
\$50,000 or more*	66.3%	68.9%	66.4%

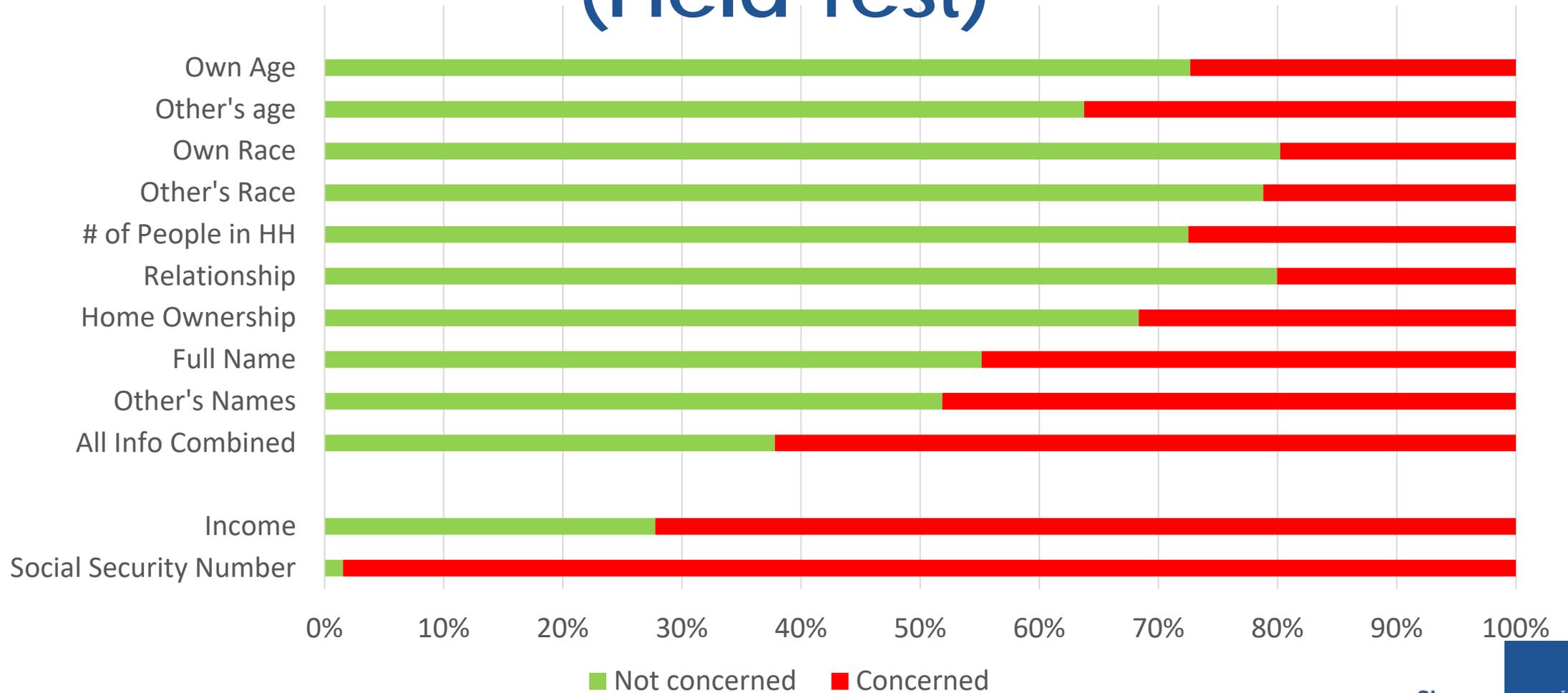
Source: Data from the National Privacy Concerns Survey (June 11th – June 29th, 2020) CBDRB-FY20-CBSM002-001.

Concern about Census Items (Nationally Representative Panel)



Income and SSN are not asked on the decennial census. Source: Data from the National Privacy Concerns Survey (June 11th – June 29th, 2020) CBDRB-FY20-CBSM002-004.

Concern about Census Items (Field Test)

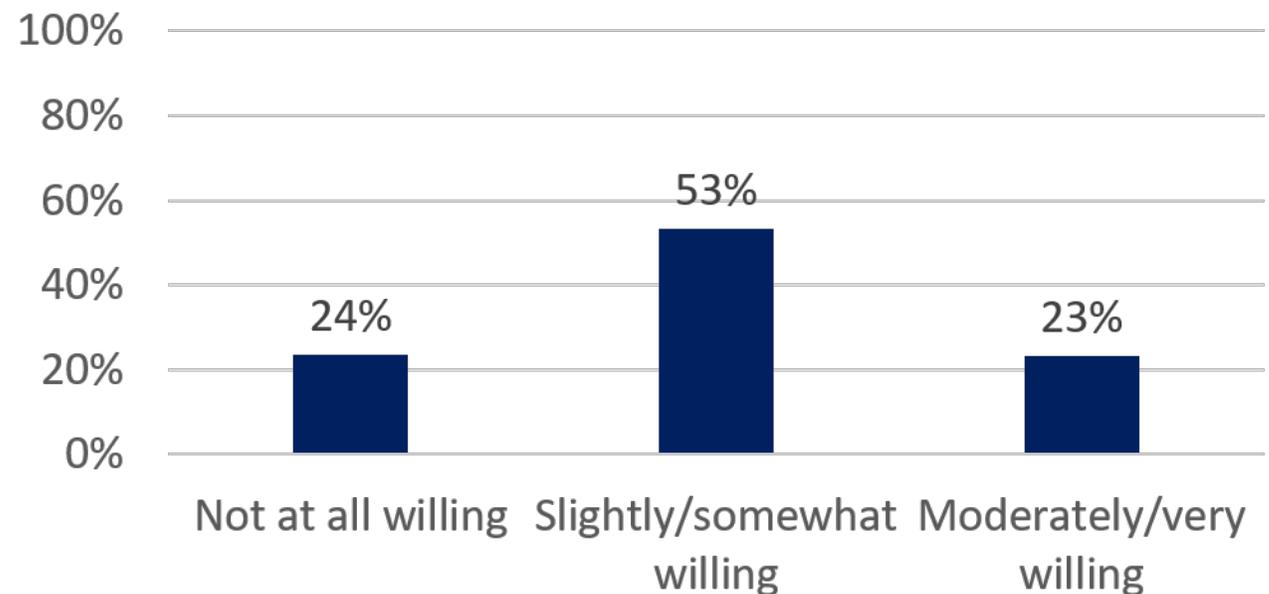


Privacy-Accuracy Trade-Off Question 1

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, for example, 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.

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?

Willingness to risk confidentiality so government can produce useful statistics



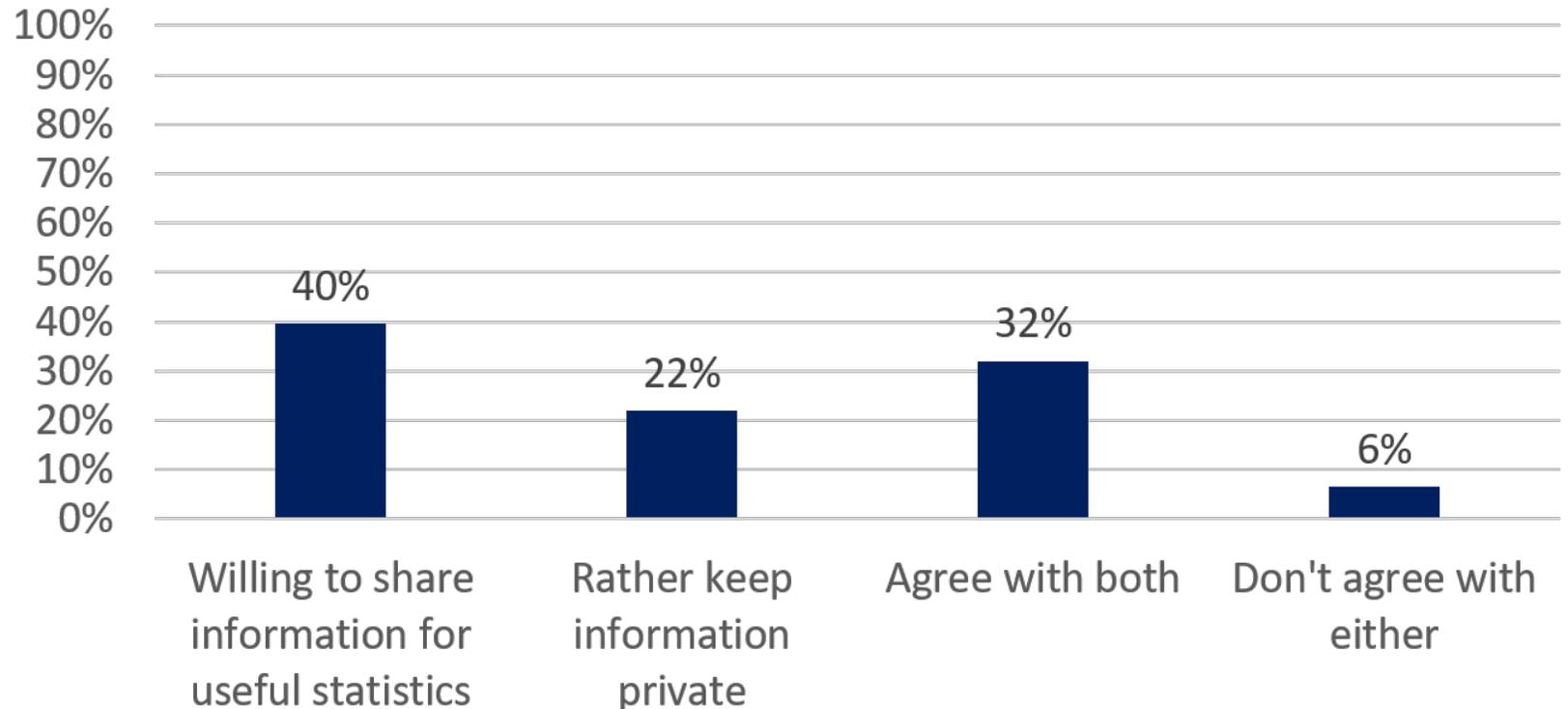
Source: Data from the National Privacy Concerns Survey (June 11th – June 29th, 2020) CBDRB-FY20-CBSM002-004.

Privacy-Accuracy Trade-Off Question 2

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

Preference for sharing vs. keeping info private



Source: Data from the National Privacy Concerns Survey (June 11th – June 29th, 2020) CBDRB-FY20-CBSM002-004.

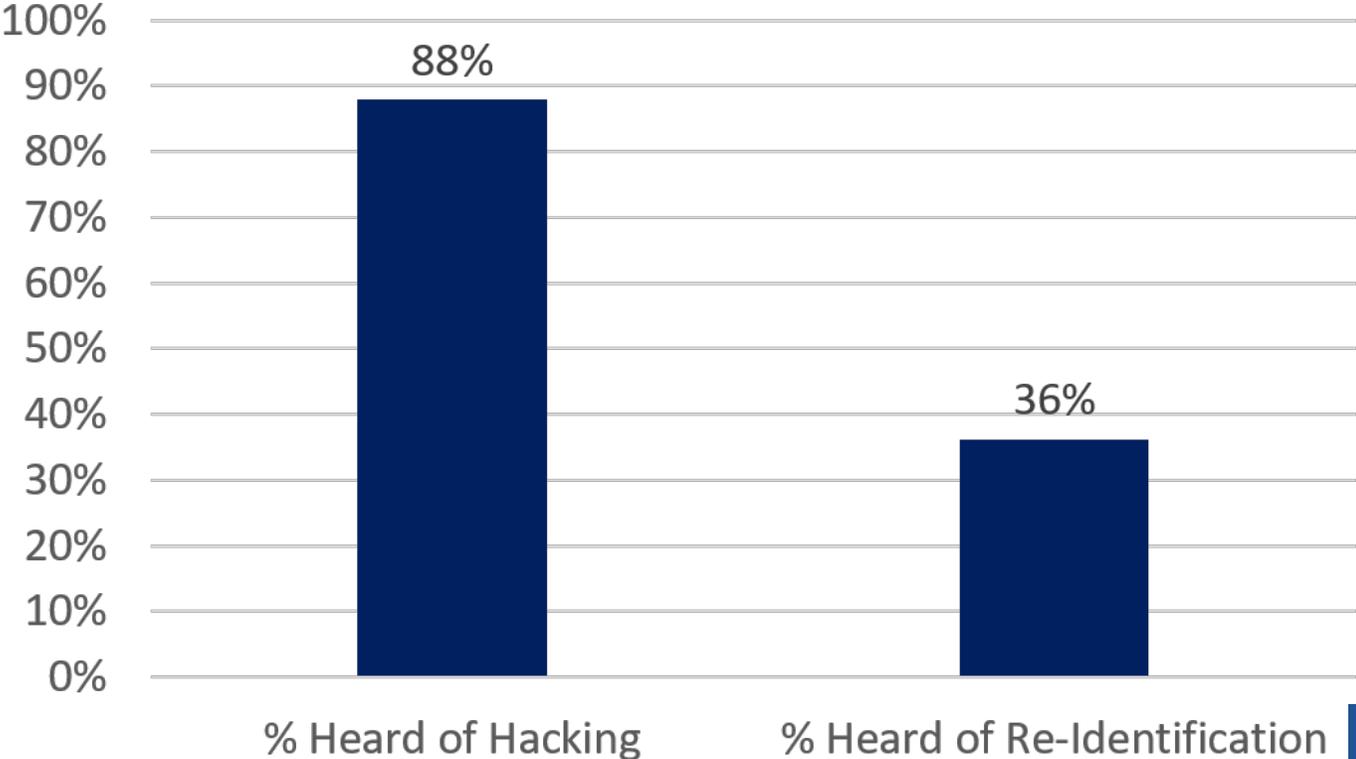
Familiarity with Data Breach vs. Re-Identification

Have you ever heard of information being stolen through hacking or a data breach?

Yes

No

Familiarity with Concepts



Concern about Census Data Breach and Re-Identification

How worried are you about information you give to the Census Bureau being stolen through hacking or a data breach?

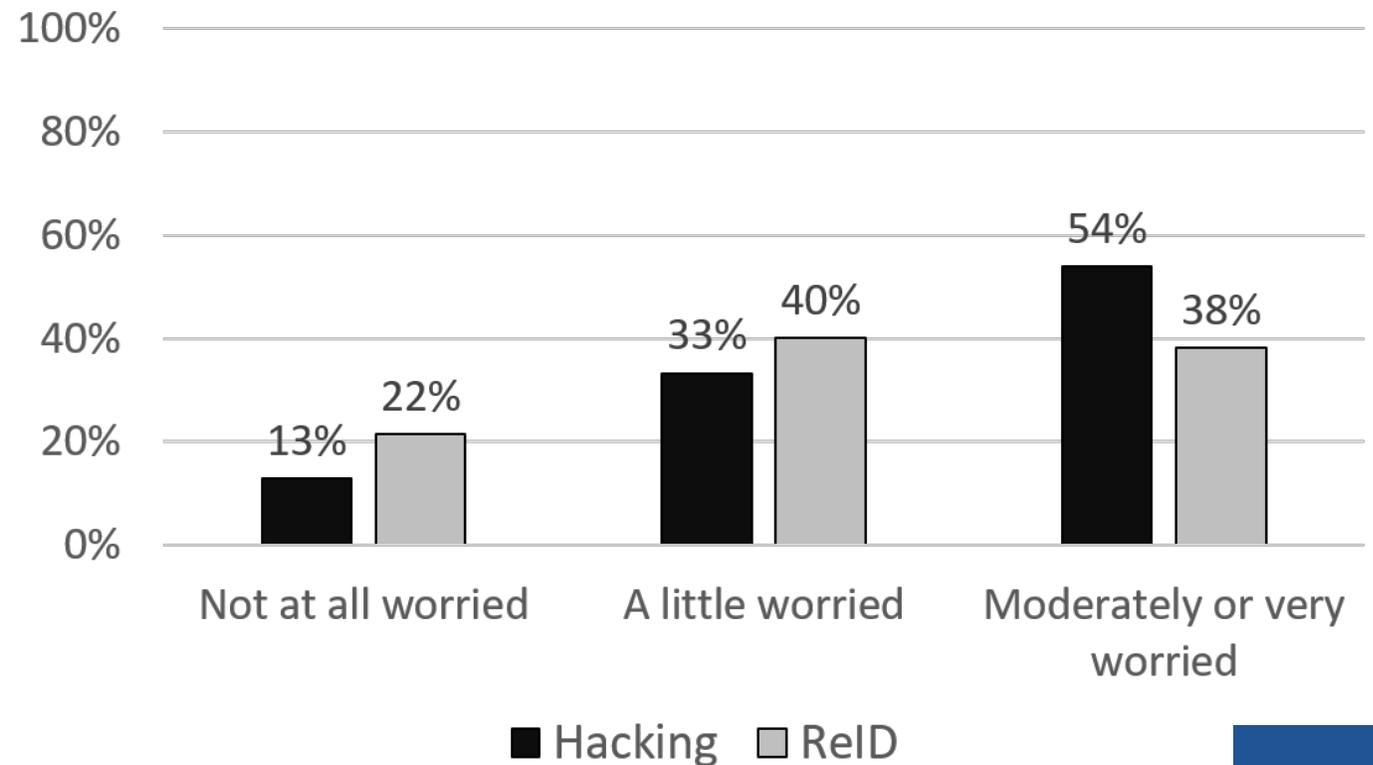
Not at all worried

A little worried

Moderately worried

Very worried

How worried about...

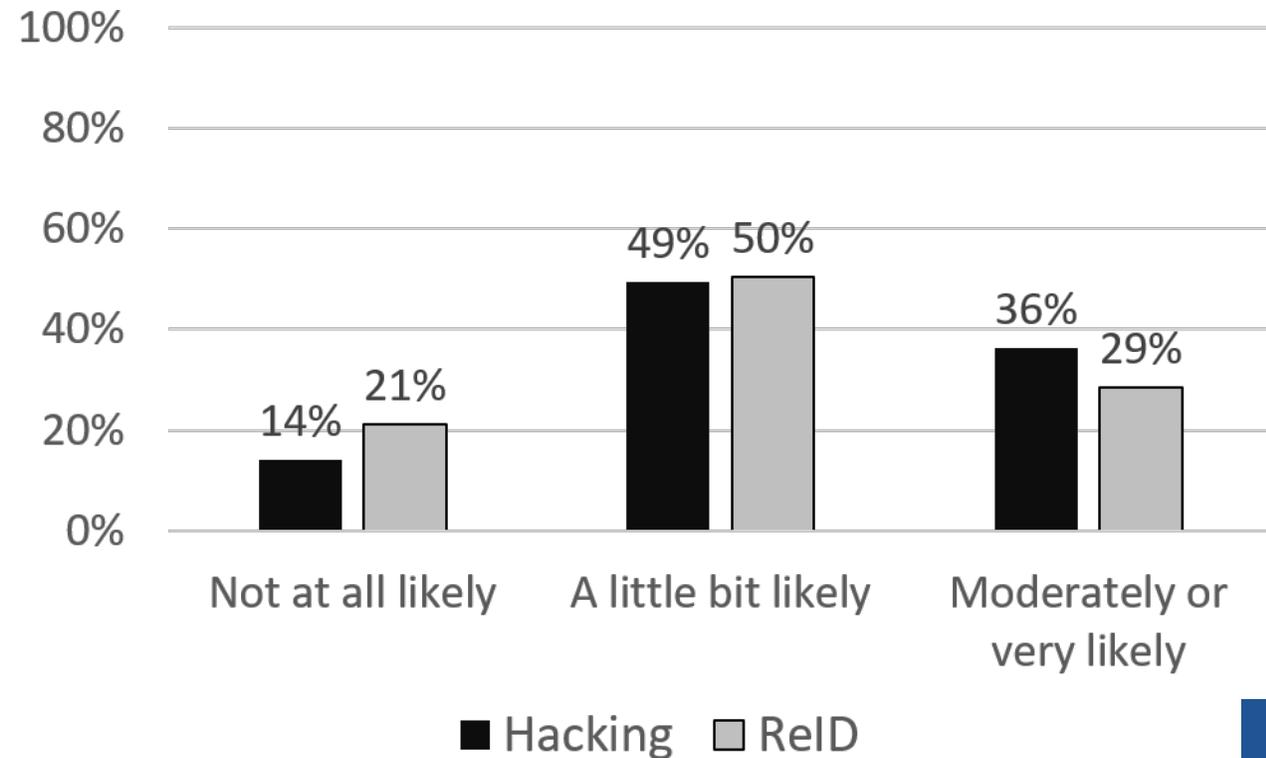


Reported Likelihood of Data Breach and Re-Identification

How likely do you think it is that information you give the Census Bureau will be stolen through hacking or a data breach?

- Not at all likely
- A little bit likely
- Moderately likely
- Very likely

How likely....



Future Participation in Circumstance of Data Breach or Re-Identification

If the information you gave the Census Bureau was stolen through hacking or a data breach, how likely would you be to participate in future Census Bureau surveys?

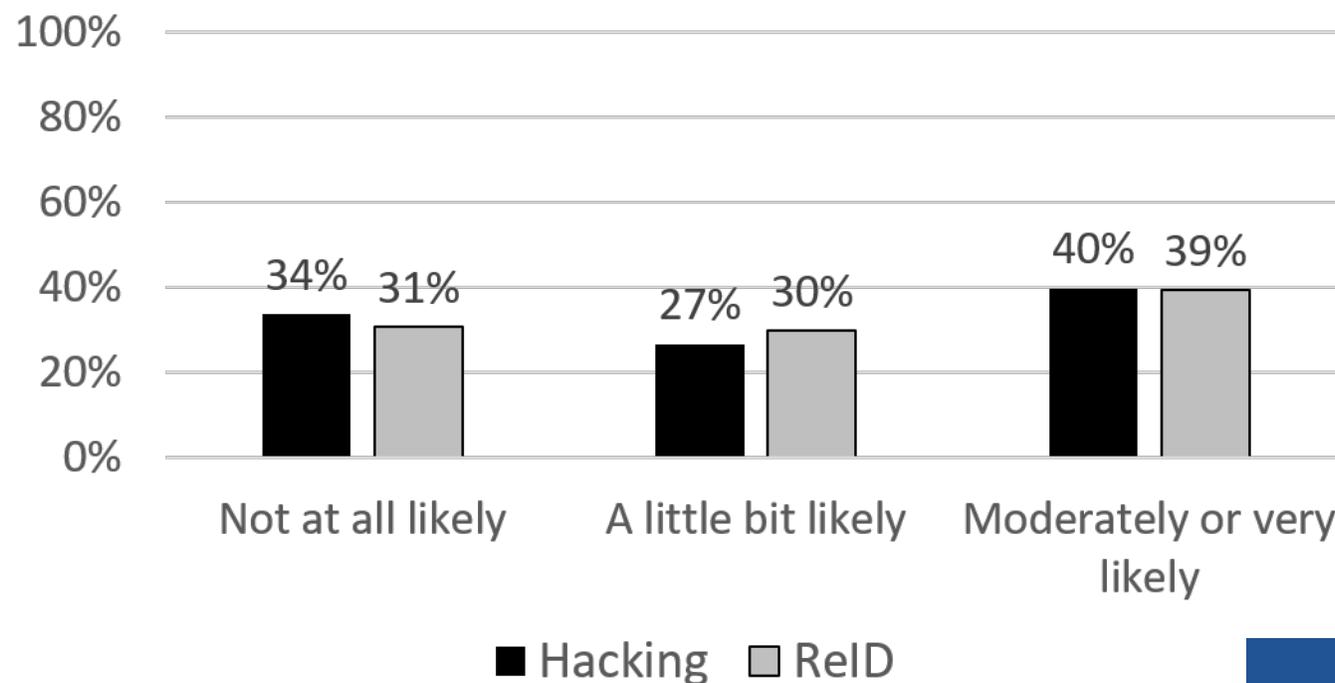
Not at all likely

A little bit likely

Moderately likely

Very likely

How likely to participate in future surveys...



Predictors of Concern

In a model predicting overall concern (concern for all decennial items), beliefs and attitudes were strong predictors.

- Those more worried about hacking and reidentification were more likely to report concern, while those more willing to share private information to benefit federal statistics were less likely to report concern.
- One practical indicator of privacy concerns, non-reporting of income, was associated with a greater likelihood of reporting concern.

Some demographics also predicted greater likelihood of reporting concern, including age 18-24 and higher education

- Those who identified as Hispanic or Black/African American alone were less likely to report concern on the set of all decennial items together

Predictors of Concern

Pattern relatively consistent across other concern items:

- Worry about hacking and reidentification stood out as strong and consistent predictors
- The two tradeoff attitude questions also frequently emerged as significant predictors, particularly the second forced-choice item
- Non-report of income as well as age and education were consistent predictors in the same direction (income non-reporters, 18-24 year olds, and those with a bachelor's degree or more consistent were more likely to report concern across items)
- Race and ethnicity were often non-significant in these models, and when they were significant the direction varied by the particular concern

Conclusions

- Other than name, most items on the decennial census are not considered particularly sensitive by the majority of respondents.
- Attitudes and behaviors regarding privacy predicted concern more than demographics.
- Depending on how it is asked, 40-76% of participants are willing to give up some confidentiality protection for the public good.
- About a quarter of respondents surveyed were not willing to sacrifice data confidentiality for the public good.

Next Steps

- Others at Census Bureau will set Privacy/Loss Budget
- Continue to monitor public opinion concerning privacy and confidentiality

Thank you!

aleia.yvonne.clark.fobia@census.gov

casey.eggleston@census.gov

jennifer.hunter.childs@census.gov