Testing Questions to Measure Perceptions of Disclosure Risks

Alfred D. Tuttle, Casey Eggleston, and Aleia Clark-Fobia
United States Census Bureau
FedCASIC, April 17, 2018

Any views expressed are those of the authors and not necessarily those of the U.S. Census Bureau.

United States™ Census Bureau
U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Introduction

- Re-identification – type of disclosure risk for statistical data products
- Pre-testing questions intended to gauge:
  - Awareness of re-identification
  - Perceptions of likelihood and impact
  - Sensitivity of data collected by the decennial census

Re-identification

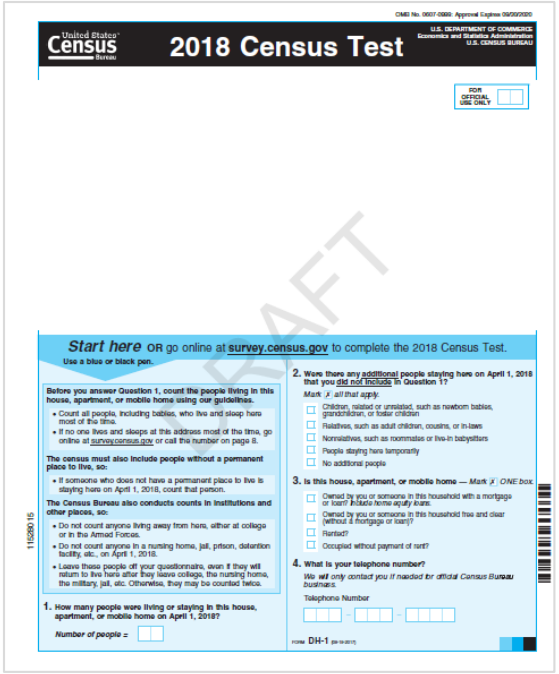
- Survey results are “anonymized” (PII removed, disclosure avoidance methods) to protect confidentiality
- Respondents can be “re-identified” in published statistics by combining them with other data sources

Cognitive interviews

- October-November 2017
- Self-administered paper questionnaire
- N=28
- Mix of age, race/ethnicity, education, single-person/multi-person households
- D.C. metro area, Florida, West Virginia
- ~ 1 hour
- $40 cash incentive

Procedures

- First administered decennial questionnaire
  - 2018 Census Test draft questionnaire
  - Provide context, specificity of collected data
- Then administered test questionnaire
  - Thinkaloud with retrospective probes
  - Probed comprehension of concepts and how respondents arrived at their answers



Disentangling disclosure

- Assumption: Re-identification is probably not a familiar concept
- Previous research – Hacking and identity theft are salient, but statistical concepts and procedures are not
- Added “data breach/hacking” to distinguish from and focus R’s’ attention on “re-identification”

Defining data breach/hacking

Every 10 years, the Census Bureau collects information about people living in the United States. Before making any of this information available to the public, the Census Bureau removes information that could be used to identify individuals (such as names and addresses). Although the Census Bureau makes every effort to protect your information, it may be possible for someone to link information you provide to the Census Bureau to your personal identity or address. One way that this could happen is if your information was stolen from the Census Bureau with your identifying details still connected to your responses. This is known as a **DATA BREACH** or **HACKING**.

Defining data breach/hacking

Before making any of this information available to the public, the Census Bureau removes information that could be used to identify individuals (such as names and addresses).

…One way that [disclosure] could happen is if your information was stolen from the Census Bureau with your identifying details still connected to your responses. This is known as a DATA BREACH or HACKING.

Defining re-identification

Though hacking and data breaches have received a lot of media attention lately, they are not the only way that your identifying information could be accessed. It could also be possible for someone to use anonymous Census Bureau data and combine it with another information source to figure out identifying information about a person or address. For example, someone could combine Census data about a small geographic area with other publicly-available information to find out that the household on a particular block with seven people living in it has three renters or two adopted children. This practice of finding out identifying information by combining anonymous information with some other data is called **RE-IDENTIFICATION**.

Defining re-identification

…possible for someone to use anonymous Census Bureau data and combine it with another information source to figure out identifying information about a person or address.

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

Re-identification questionnaire

- First section introduced disclosure concepts
  - For each concept, R's were asked
    - Have you heard of it?
    - Have you or others you know experienced it?
    - How worried are your about it with regard to data you give to the Census Bureau?
    - How likely is it to happen?
- Second section – Concerns about the specific types of data collected on the census questionnaire

Findings: Data breach/hacking

Data breach/hacking is salient

- Most R's had personal experience
  - Stolen credit card numbers
  - Funds stolen from bank accounts
  - Credit reporting agencies, banks, retailers, government agencies
  - Social media and other accounts compromised

Data breach/hacking is salient

- News reports, films, reality TV

"Seems like every other day there's a story in the news about some company or bank or something… where somebody hacked into their system and was able to access information they should not have been able to access."

Data breach/hacking is salient, and nuanced

"Either through hacking or computer error or personnel error, someone releases information about me onto the internet."

"Someone either hacking into the system or someone who works for the system that gave someone else access to get in."

Data breach/hacking is salient because of risk

- Personal risk makes it salient
  - Identity theft
  - Financial loss
  - Stalking
  - Etc.

Data breach/hacking is salient

- Most R's were generally not concerned about the Census Bureau's ability to protect their data
- Though hacking is generally perceived to be possible for any institution

Findings: Re-identification

Factors that contribute to misinterpretation

- Re-identification is not a common term
- Ambiguity combined with familiar linguistic components
- Questionnaire context effects
- Lack of knowledge about statistics

Re-identification is not a common term

“I’m just hearing that for the first time… I'm assuming that's how people carry out identity theft… But until today I'd really never, I've heard of identity theft but I hadn't heard it described as re-identification.”

Re-identification is ambiguous, yet familiar

- Linguistic components are familiar: “re-”, “identification”

"It doesn’t sound like re-identification. Your identity is who you are, the things that identify you. Who you are, your name, what you own.”

"‘Re-‘ means to do again, so… recreate your identification? Change your information to… save yourself from being identified? A Re-do? Recreating identification… Unfortunately, you can re-create a new card, but you can't take away your name and birth date if somebody has that information."

Salience of identity theft

- Most common interpretation of re-identification is identity theft

"It would be similar to someone getting your social security number and your address and trying to create an identification off of what information they can get from you."

Questionnaire context effects

"Maybe because this talks about data breach and hacking, I was already in the vein of stolen identity, so when I see re-identification, someone is taking my identity and they are going to act in my identity."

Statistics are not salient

- General lack of knowledge of statistics and Census Bureau data products
  - Most respondents were not users of statistics
  - No context for understanding statistical products, potential disclosure risks

Re-identification – Correct interpretations

"When somebody takes information from a number of sources and somehow is able to either with a computer or otherwise sort it in such a way to pinpoint or more closely identify individuals or organizations that supposedly were anonymous when the information was given."

“Different sources may have your information, partial information here, partial information here (gesturing) and then combining it so that they can figure you out.”

Potential for alarming misinterpretation

"I sorta get it. It reads clear, but when you think about it… it's just that easy to go on the Census Bureau web site and you got like a search engine where you can find out these things… I know you **can't just go on Google** and just say, on this block, who lives on this block, what races live on this block… **but I guess on you all's site you can do that**… I'm **just thinking about how crazy that is**."

Section 2: Concern, likelihood, and impact of disclosure

Perceived risk of disclosure of decennial census items

Would it concern you if someone was able to find out [your name, age, etc./the name, age, etc. of people you live with]? Yes/No

If someone was able to find out [your ___/the ___ of someone you live with], **how much of a negative effect would it have** on your life?

No negative effect

Minor negative effect

Moderate negative effect

Severe negative effect

Perceived risk of disclosure of decennial census items

- Individual census items generally not considered risky
- R's bothered more by *idea* of hacking their information than that disclosure of census data pose a risk
- "…Someone finding out some combination of your data" and "all of your data" elicited increased concern

Perceived risk of disclosure of decennial census items

- Caveats
  - R's often were not thinking of the decennial census questionnaire in answering risk questions. Tended to think about "your information" more generally.
  - Cognitive interview context – Discussion of disclosure risks prior to answering sensitivity questions

Summary

- Re-identification not easy to explain
- Salience of hacking, identity theft overpowered attempts to present re-identification in SAQ
- Ambiguous familiarity of “re-identification”
- Respondents generally do not have knowledge about statistics, processes to protect confidentiality
- Incorrect interpretations may cause unnecessary alarm
- Decennial census items not thought of as risky

Next steps

- Revision and additional pre-testing
  - Establish more direct connection between questions about risk and specific census questions
  - Perhaps avoid “re-identification” and focus on riskiness of census data items
- Pilot survey

Thanks!

Alfred (Dave) Tuttle

alfred.d.tuttle@census.gov

We welcome feedback!