Leveraging Instrument Test Paradata: Strategies for Managing Sensitive Data and Programming Challenges

Ana I. Sánchez-Rivera, Ph.D.
Rae Ellis, Ph.D.
Center for Behavioral Science Methods (CBSM), U.S. Census Bureau

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

Disclaimer: This presentation is released to inform interested parties of research and to encourage discussion. The views expressed are those of the authors and not necessarily those of the U.S. Census Bureau. The Census Bureau's Disclosure Review Board and Disclosure Avoidance Officers have reviewed this product for unauthorized disclosure of confidential information and have approved the disclosure avoidance practices applied to this release. DRB Approval Number: CBDRB-FY25-CBSM003-006

Agenda

- Intro to Paradata & Keyed Test Data
- Research Questions
- Data
- Findings
- Conclusions

Paradata as an Analysis Tool

- **Paradata:** Information about the process by which survey data are collected, does not refer to the content of the responses themselves.
- **Internet Paradata**
  - Relatively inexpensive to collect
  - Information about respondents' behavior, which may include aspects like use of help, backing, answer changes and error warnings
- **Challenges to using Internet Paradata**
  - Large, complex, unstructured data
  - Programming analysis takes time & requires the development of specialized skills
  - Title 13 protections

Keyed Test Paradata as Synthetic Paradata

- **What is Keyed Test data?**
  - Information generated during the User Acceptance Testing (UAT) when Subject Matter Experts (SMEs) are testing the instrument before it is released
    - UAT generates hundreds of synthetic records

- **SME's main task is to test their questions against the Specification file**
  - For typos in wording or response categories
  - To find bugs, errors in programmed paths and to test functionality
  - Most try to imitate behavior patterns of real respondents
    - May overrepresent problem behavior, small demographic groups or rare events to assure all pathways of the instrument work

Purpose & Research questions

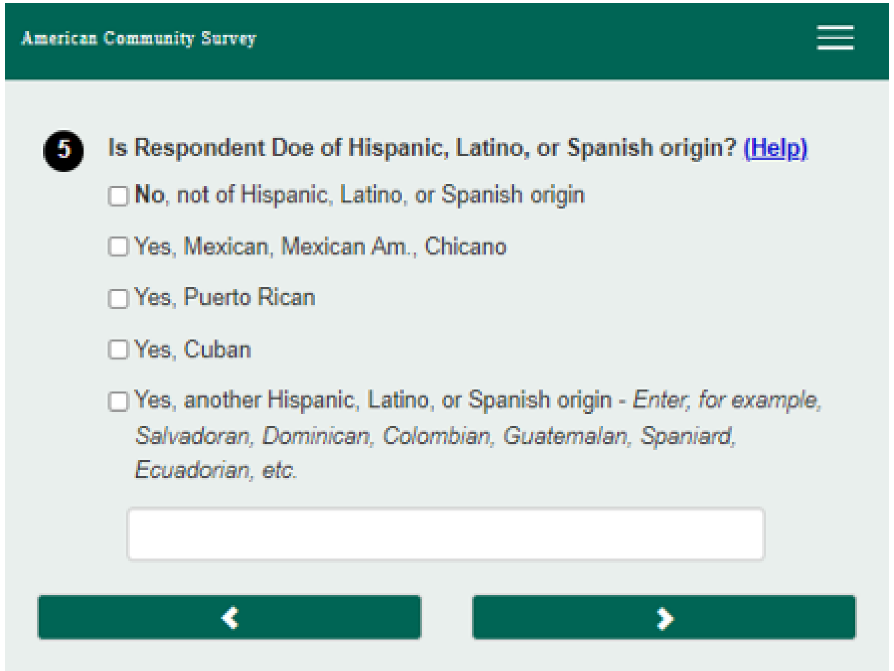
Purpose: To assess how Keyed Test (KT) paradata can support the development of a programming code to analyze the final paradata dataset.

1. Does the **structure and content** of the KT paradata align with those of the final dataset?
2. Does the KT paradata support an **accurate** development of processing code?
3. Is it efficient to use the KT paradata when developing processing code, in terms of **time and effort**?

Data

- **Hispanic origin question**
  - 2020 Census Analysis of Census Internet Self-Response Paradata by Language found that the Hispanic origin question was one of the most difficult questions to answer on the 2020 Census.
  - Top 5 pages with more: Breakoffs, trigger Edit warnings, use of Help text, one of the most frequently backed (to include verify).

- **Paradata datasets sliced by Hispanic origin question**
  - 2024 ACS Keyed Test Data
    - Total of 222 cases (rep. households), 410 Persons (rep. household members)
  - 2022 ACS Paradata
    - 847,736 Households, 2,137,112 household members



5 Is Respondent Doe of Hispanic, Latino, or Spanish origin? (Help)

□ No, not of Hispanic, Latino, or Spanish origin

□ Yes, Mexican, Mexican Am., Chicano

□ Yes, Puerto Rican

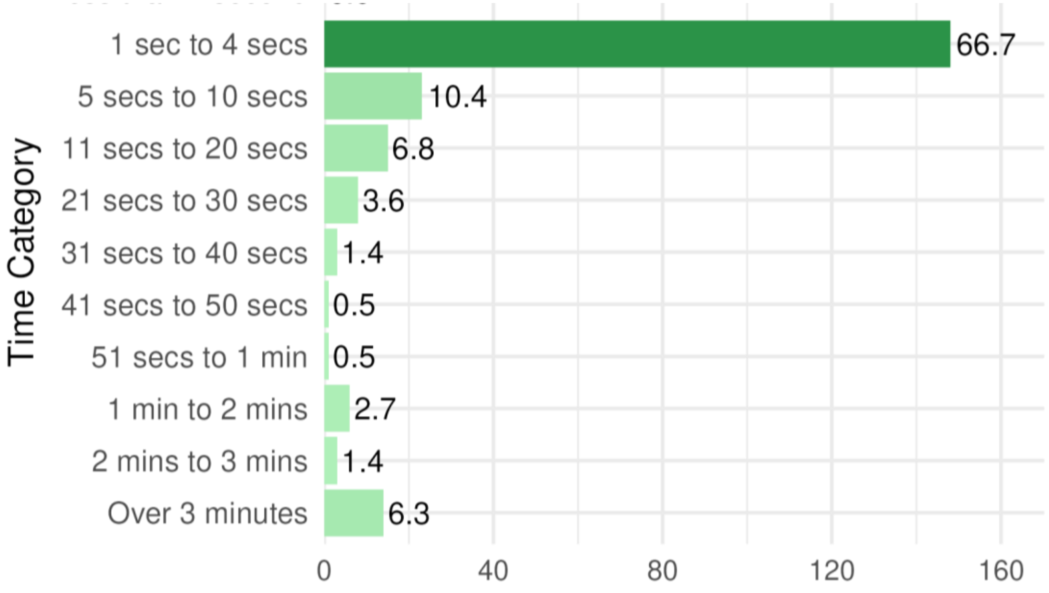
□ Yes, Cuban

□ Yes, another Hispanic, Latino, or Spanish origin - *Enter, for example, Salvadoran, Dominican, Colombian, Guatemalan, Spaniard, Ecuadorian, etc.*

Findings

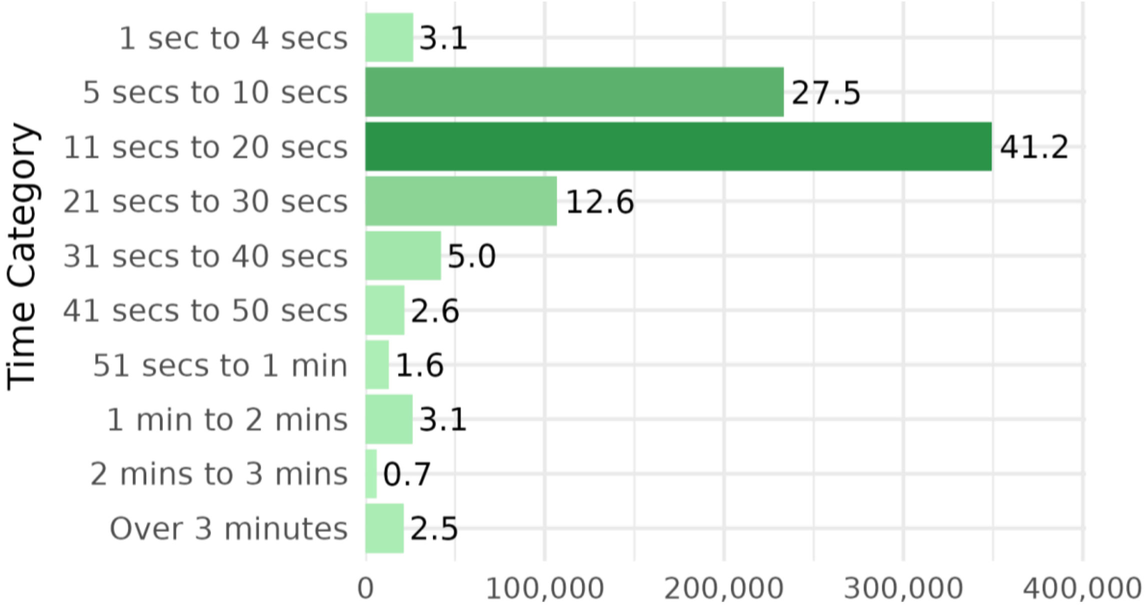
Time spent in the Hispanic origin question

Keyed Test Paradata



Median time in question: 9 seconds

2022 ACS Paradata



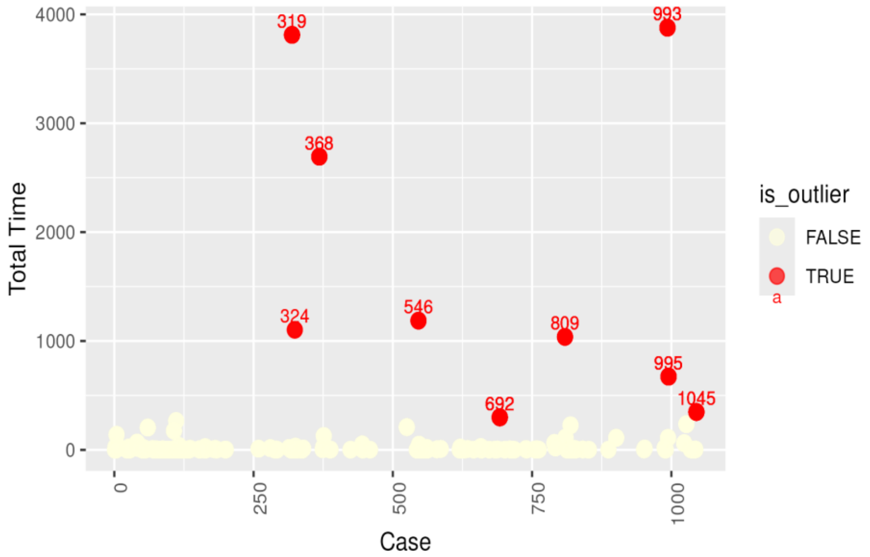
Median time in question: 16 seconds

Long/Short Time Outlier Cases in the Hispanic origin question

Keyed Test Data

Cases with unusually Long/Short Times in the Hispanic origin question

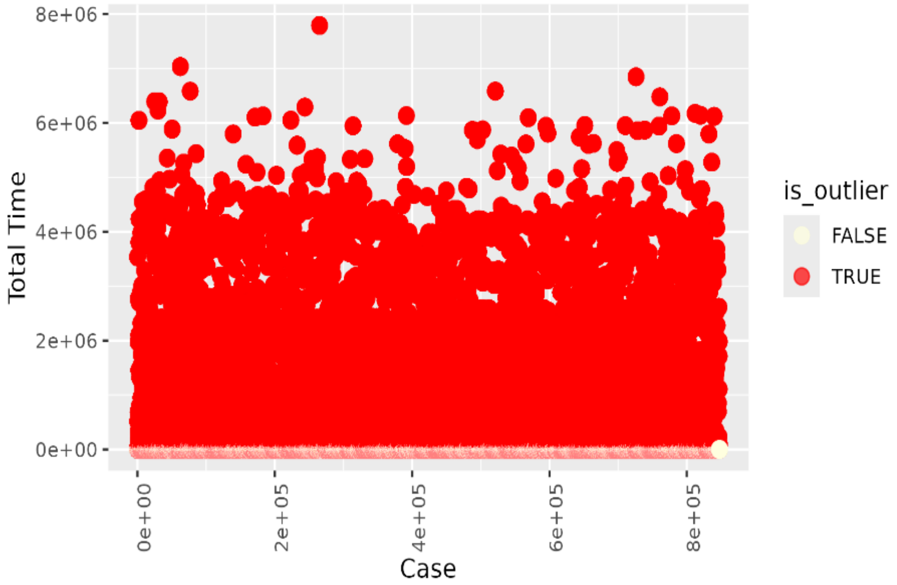
204 Cases Identified as Outliers



2022 ACS Paradata

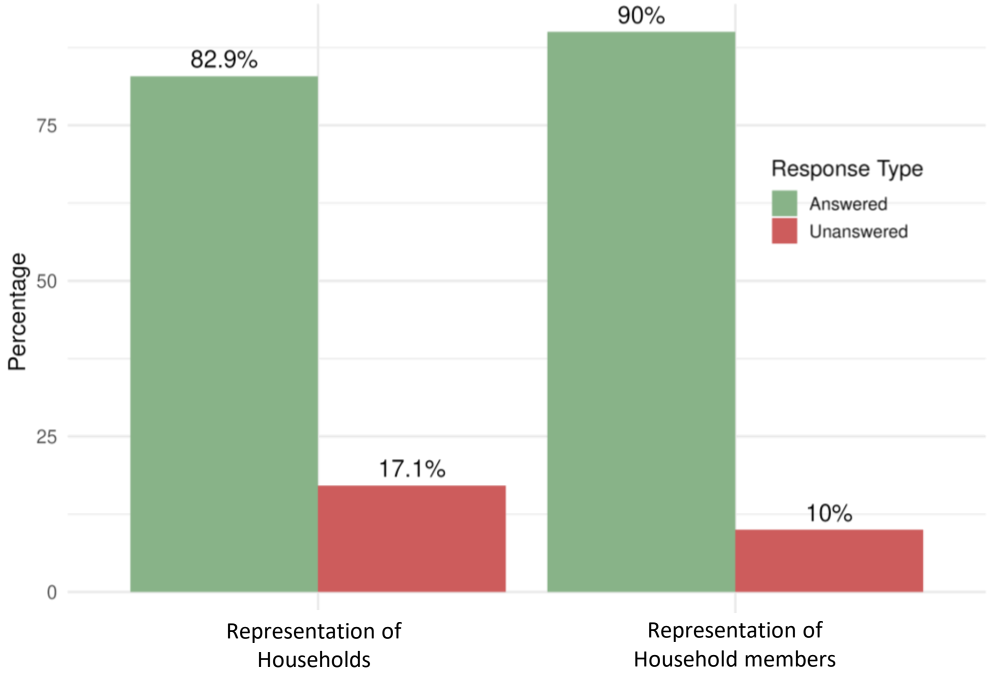
Cases with unusually Long/Short Times in the Hispanic origin question

Approx. 1,220,000 Cases Identified as Outliers

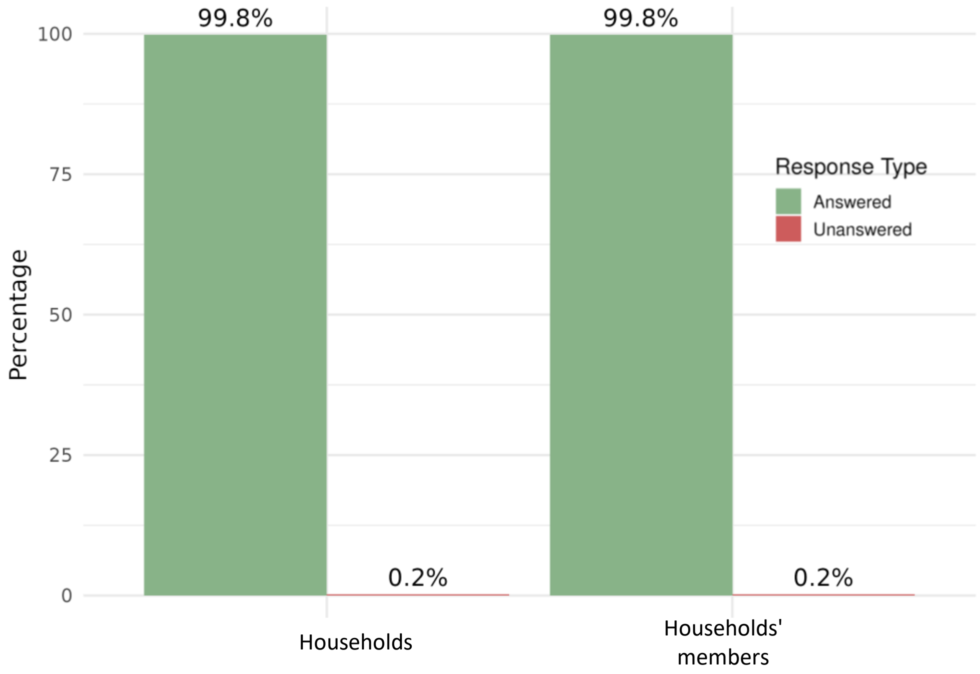


Cases that Answered the Hispanic origin question

Keyed Test Paradata

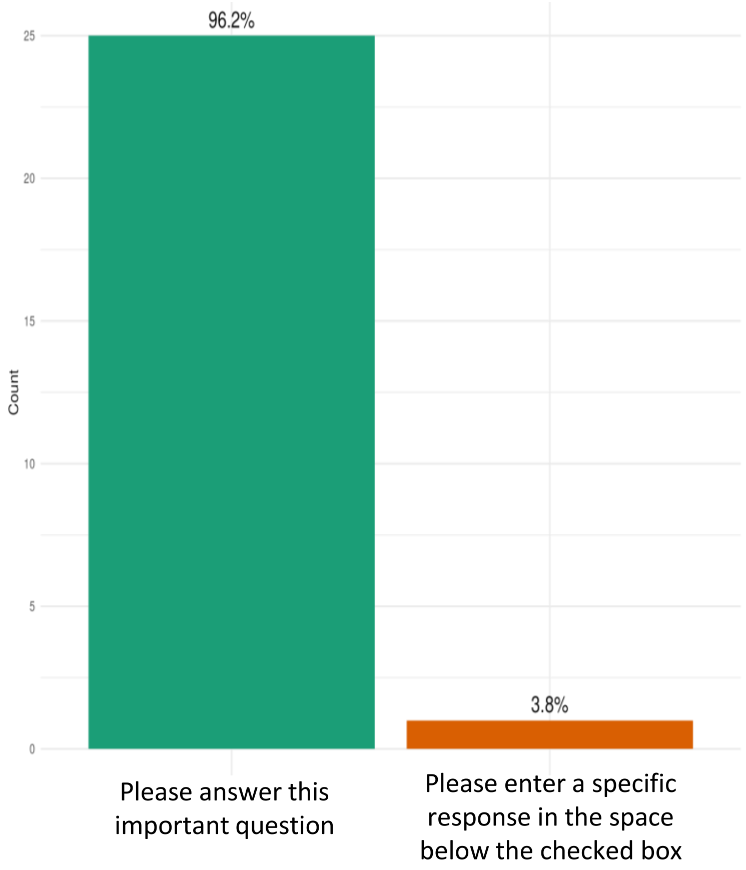


2022 ACS Paradata

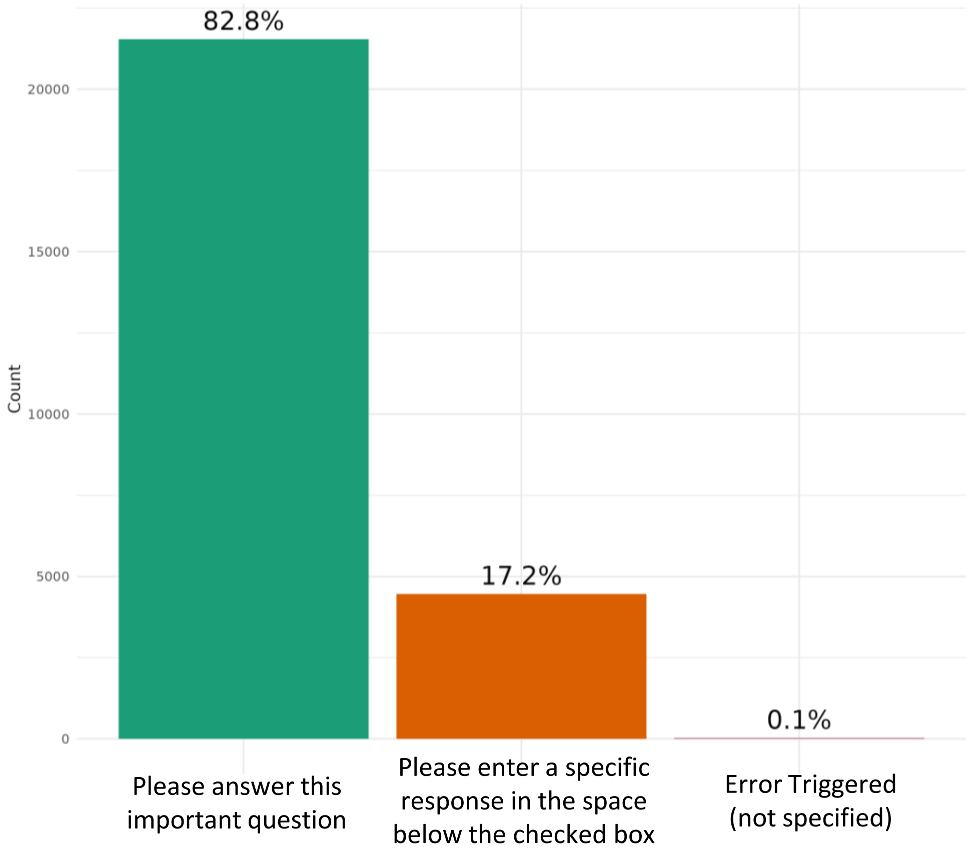


Type of error messages in the Hispanic origin question

Keyed Test Paradata



2022 ACS Paradata



Summary

*1. Does the **structure and content** of align with those of the final dataset?*

- KT paradata closely **mirrored the structure** because response categories aligned across datasets.
- KT paradata **did not always mirror the content** because of the differences in frequency distributions as they limited our ability to create meaningful response groupings.

*2. Does it support an **accurate** development of processing code?*

- 2024 KT paradata allowed us to **write a correct and valid** code even when the 2022 ACS paradata was older and did not fully align the new way the Census Bureau is collecting paradata.

*3. Is it efficient to use KT data when developing processing code, in terms of **time and effort**?*

- It is **efficient to save both time and effort** when you receive the real dataset**, but it requires** several **adjustments** after.

Conclusions

Advantages:

- Preparing for data wrangling
- Coding for data visualizations
- Coding to develop research questions
- Reducing time analyzing production data once available
- Teaching and Sharing

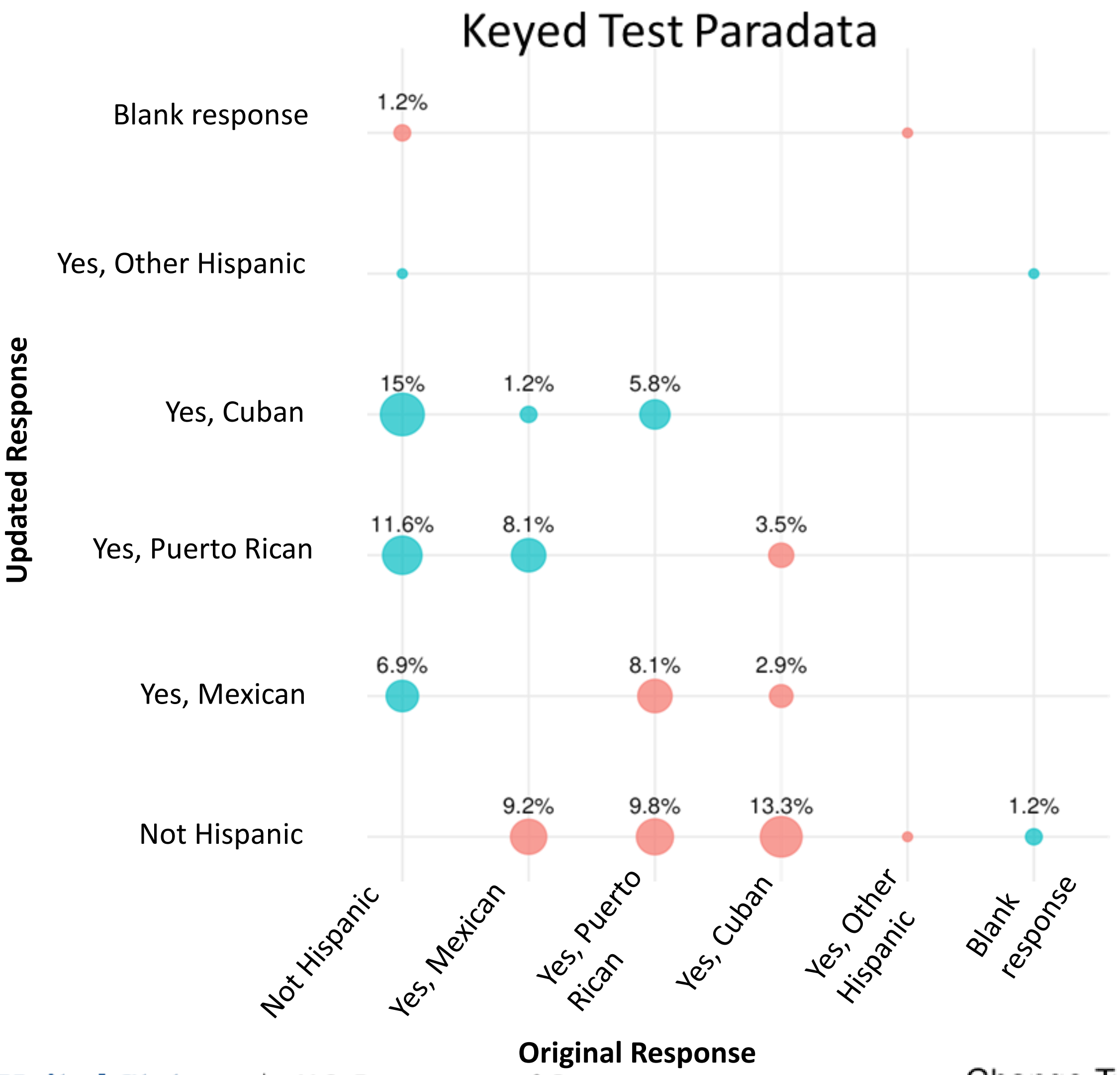
Limitations:

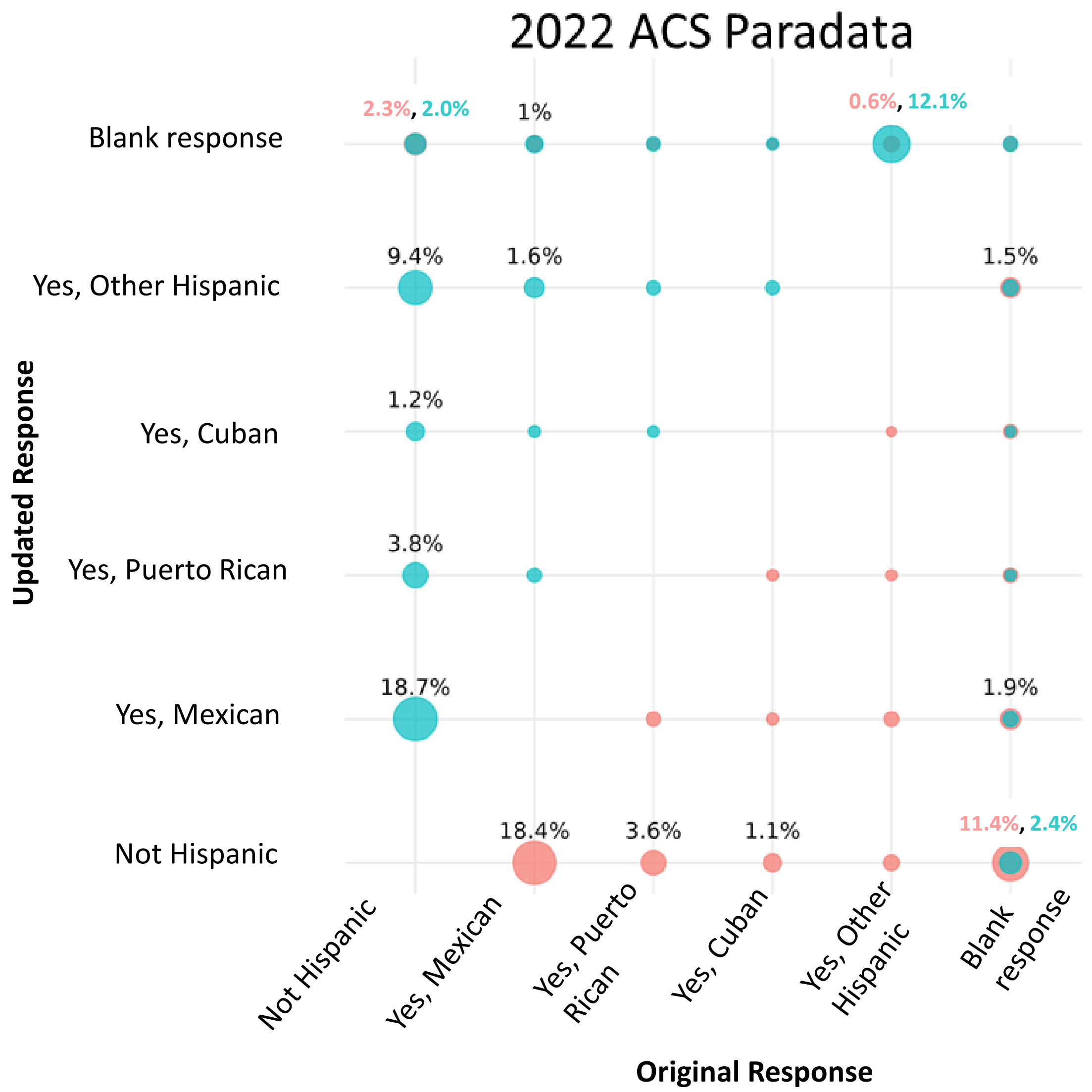
- Keyed data cannot be used to confirm research questions or hypotheses
- Data does not accurately mimic patterns
- Sample size differences need to be considered when developing graphics

Next Steps

- Apply this framework in ongoing projects
  - Example: Analysis by device type
- Develop training materials so other analysts can effectively use the paradata
  - Census Bureau is looking to standardize paradata structure (e.g. Centurion 2.0)
- Categorize observed patterns that can help us contextualize behavior
- Define analytical parameters, such as:
  - What constitutes *short* vs. *long* time on a question?
  - Do the parameters change depending on the question or question type?
  - Are the patterns observed across surveys?

Next Steps





Change Type

- ● Changed response from
- ● Changed response to

Thank you!

Contact Information:

- ana.i.sanchez.rivera@census.gov
- renee.ellis@census.gov