

# Synthetic ACS Microdata: Considerations for the Census Bureau and Stakeholders

Leslie Reynolds and Jan Vink  
Cornell University  
Program on Applied Demographics



# Outline

- ❖ Background: PAD's role as data users and evaluators
- ❖ Introduction to ACS Microdata and the Upcoming DAS change
- ❖ Incorporating Use Cases
  - how do we handle these cases now?
  - what obstacles would come with synthetic data?
- ❖ Potential Barriers, and What we Need from the Census Bureau

# Our Background

- PAD is an applied demography group within the Cornell Population Center
- Contracted by the State Data Center to be experts on all things Census
  - Active members of the Federal State Cooperative for Population Estimates (FSCPE)
  - Lead demographer Jan Vink is FSCPE steering committee chair
- Frequent interactions with state and local government groups about data results and quality
- Heavily involved in feedback on the 2020 Census Differential Privacy Demonstration Data

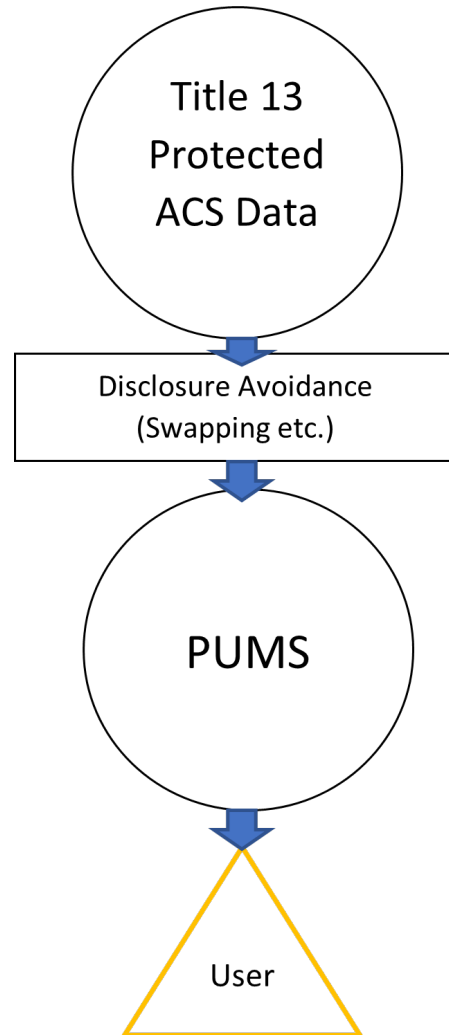
# What is ACS Microdata?

- Sample of full household responses
  - Public Use Microdata Series (PUMS) is NOT used in the creation of the published tables
- Allows for:
  - Creating detailed tabulations where ACS tables are aggregated,
  - custom cross-tabulations,
  - characteristics of different members within household,
  - statistical models based on individual level characteristics
- Weights
  - Population and household weights for representative results
  - Replicate weights (for standard errors and MOEs)
- Lowest level of geography is Public Use Microdata Area (PUMA)

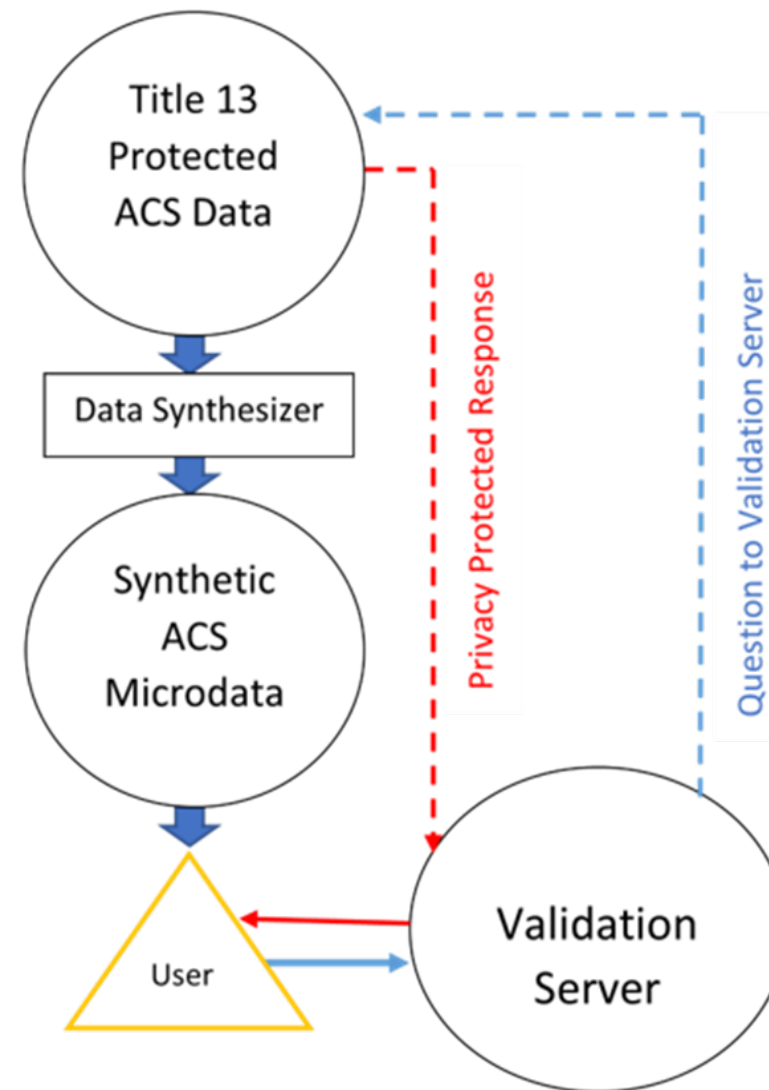
Thinking about switching to synthetic data – what are the expectations?

# Layout of Current and Future Access to ACS Microdata

Current ACS Workflow



Synthetic ACS Workflow



# American Community Survey Disclosure Avoidance

Share



## FAQs

[EXPAND ALL](#) | [COLLAPSE ALL](#)

- ☐ **Will data users have an opportunity to provide feedback on the disclosure avoidance changes made to the ACS?**

The Census Bureau is committed to transparency in its decision-making regarding privacy and confidentiality. As such, the Census Bureau is planning a **public test** of a **potential synthetic data and validation service** for the ACS public use microdata. We will **gather feedback from that test** to inform the decision-making and design of an improved disclosure avoidance solution for the ACS. Currently this test is **planned for mid-2025.**

### Related Information

[Disclosure Avoidance](#)

[FACT SHEET](#)

[What Are Synthetic Data?](#)

[Statistical Safeguards](#)

# Data quality

## Utility

Timeliness  
Accessibility  
Granularity



## Objectivity

Accuracy  
Coherence



## Integrity

Credibility  
Privacy



How would this format  
impact use cases?

What are potential barriers  
to making an external test  
successful?

- Many unknowns
- How to test for accuracy  
and fitness-for-use?





# Example: *Cross-tabulations*

## Language spoken at home

Use: Agencies need to know what languages are used across populations to provide services

- Languages are grouped in tabulated data
- Microdata has detail on language spoken at home, BUT is not enough to resolve ambiguities
  - e.g. Portuguese spoken in Portugal and in Brazil are different
  - Some languages are written using different alphabets depending on location
  - Need to include data on ancestry and country of birth to resolve ambiguities

Place of birth (Recode) (POBP)	Household population in households that speak Portuguese	At least one person in the household 14 and over speaks English only or speaks English 'very well'	No one in the household 14 and over speaks English only or speaks English 'very well'
<b>Total in New York</b>	71,473	64,788	6,685
New York/NY	28,486	27,337	1,149
Brazil	21,238	17,956	3,282
Portugal	7,832	6,244	1,588
Connecticut/CT	1,040	1,040	0

# Example: *Cross-tabulations (continued)*

## Language spoken at home

- Detailed household language has 132 responses
- Ancestry has 234 responses (times two for first and second responses)
- Place of birth has 224 responses

 Giant cross tabulation where cell sizes vary greatly

Can a synthesizer reproduce this?

Can we also get Margins or Error? How do they compare with PUMS – MOEs?

How do we compare cross tabulations from PUMS and from synthetic data and measure fitness-for-use?

How do we prioritize fitness-for-use analyses of (cross-)tabulations?

Will the output of a validation server provide usable detail after applying Disclosure Avoidance?  
Including Margins of Error?

## Example: *Household relationships and derived variables*

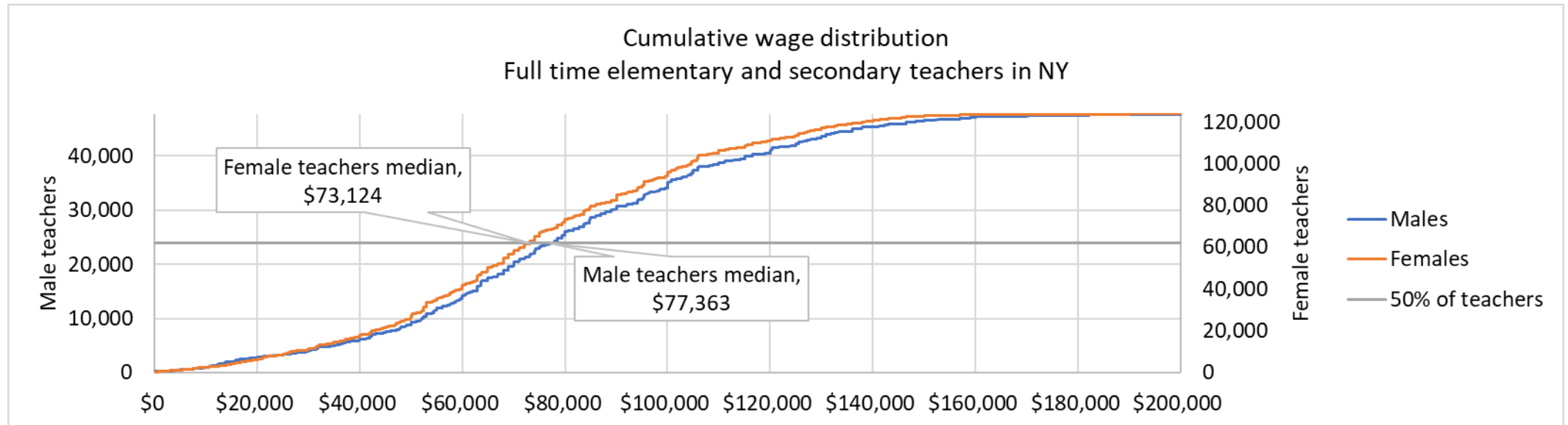
How many children with full-time working parent(s) live in poverty?

- PUMS procedure:
  - Select households with children
  - Select if householder worked full time (hours per week and weeks worked)
  - Select if spouse/partner present and worked full time
  - Count children with family income to poverty ratio  $0 \leq \text{ratio} \leq 100$
- Requirements for synthetic data (and validation server)
  - Household identifiers
  - Similar correlations between work status of parents
  - An income-to-poverty variable
    - Derived from synthesized households and incomes or synthesized ratio?
- How to test for fit-for-use?

## Example: *Distributions*

How do wage distributions differ between male and female teachers?

- Need for metrics to compare distributions:
  - Between male and female teachers
  - Between PUMS, **synthetic data** and **validation server** data
- Construction of a full distribution requires microdata; **will the validation server return microdata?**



# Example: *Regression*

## Modeling migration status

- Example from “*Analyzing US Census Data: Methods, Maps, and Models in R*”, Kyle Walker, 2023
- The example models whether an individual in the labor force aged between 25 and 49 changed residences in the past year, as a function of educational attainment, wages, age, class of worker, and family status in Rhode Island.
- Will the resulting model be the same or better when built with synthetic data?
  - What do we mean by “the same”? How will we know?
  - Wages are top-coded in PUMS- what about in synthetic microdata?
- Can we run the same model through the validation server? What will be returned? How does that compare to the PUMS result?





Objectivity

Integrity

Utility



# Barriers and opportunities

- Testing for accuracy and usability and documenting findings is HARD
  - Many stakeholders don't have sufficient skills, resources, or time to participate in a test
- Lack of a common language
- Comparison between PUMS and synthetic microdata focused on use is needed
- Transparency of Census Bureau decision making
  - What are the risks and potential costs of keeping on the current Disclosure Avoidance path?
  - What do internal tests of the synthetic data reveal? (honest and complete assessment)
  - Relay strengths and weaknesses of selected implementation
  - Publish requirements of synthesizer and validation server and be open to feedback
- Communication plan
  - Feedback is more effective if there is room for a back-and-forth
    - Potential for a working group
  - Facilitate communication between stakeholders



We Can Find  
the Balance.