Estimating Survey Nonresponse Bias Using Tax Records

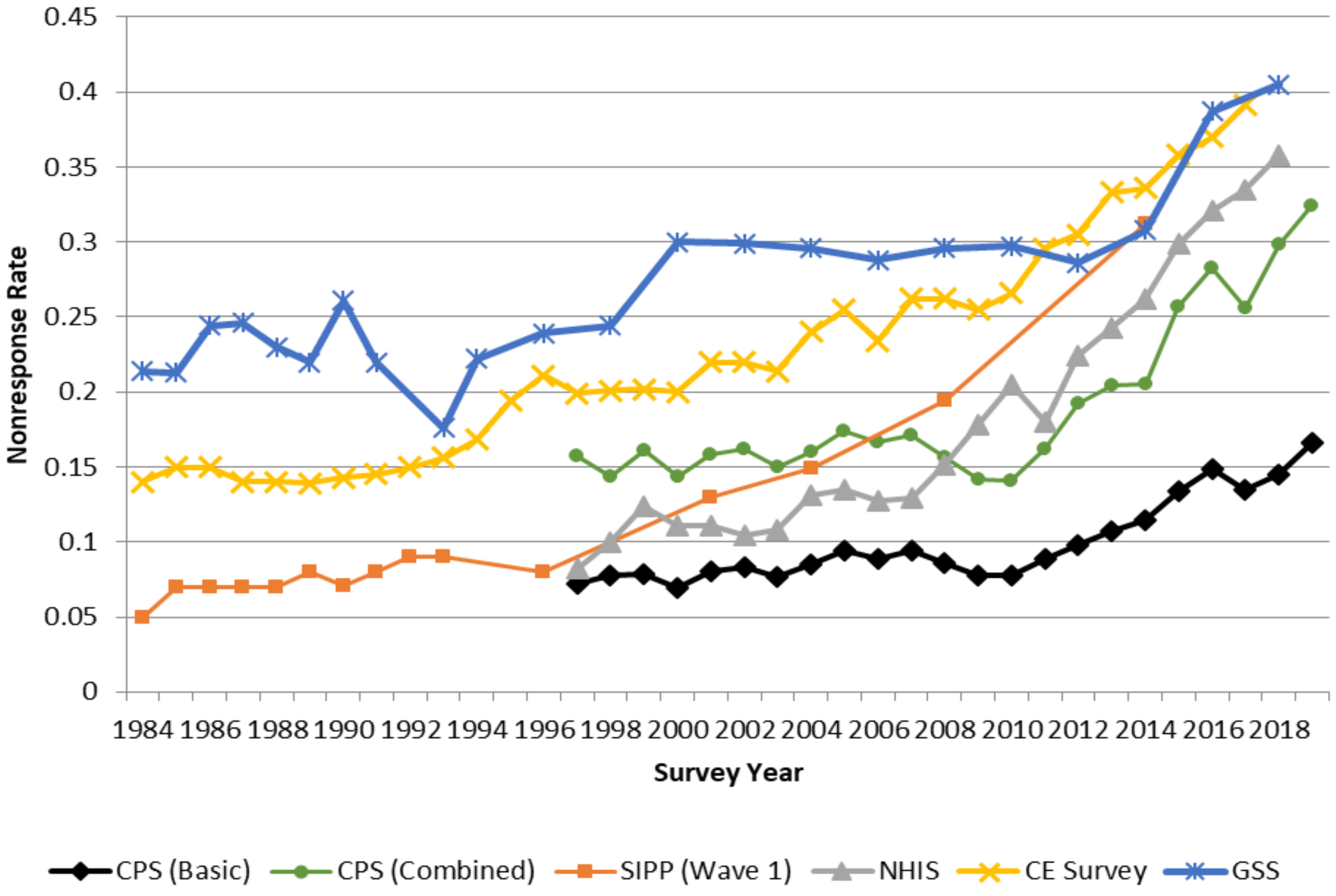
C. Adam Bee U.S. Census Bureau
Graton Gathright Amazon
Bruce D. Meyer University of Chicago, AEI,
U.S. Census Bureau and NBER

Federal Committee on Statistical Methodology Conference
September 2020

Preliminaries

Disclaimer:	The views expressed in this presentation do not necessarily reflect official positions or policies of the Census Bureau.
Note:	This project makes use of protected data. The analyses reported in the paper were done in secure settings at the Census Bureau headquarters in Maryland and at Census RDCs. The results presented here have been formally reviewed to ensure that no confidential information is disclosed. *Disclosure Review Board clearance memos dated 2015-05-13, 2015-08-13, 2016-04-11, 2016-04-27.*

Rates of Unit Nonresponse in Major Household Surveys



Motivation

- Unit nonresponse a focus of researchers and policy makers
  - Two recent panels of the National Academy of Sciences on nonresponse
  - Office of Management and Budget quality standards for federal surveys based on response rates
- Key question is extent of **bias** due to unit nonresponse
- In absence of evidence, nonresponse bias used as excuse
- New approach to assess bias by linking respondents **and** nonrespondents by address to individual tax returns
- Apply method to CPS Basic and CPS ASEC, source of official income and poverty statistics

Overview of Our Paper

Research Questions	• How do CPS respondents and nonrespondents differ on income and other characteristics? • Is ZIP code-level information sufficient for discerning income differences between respondents and nonrespondents?
Data	• 2011 CPS ASEC • Universe of IRS Form 1040s filed in calendar year 2011 • Public-use ZIP code-mean AGI data from IRS Statistics of Income program
Approach	• Link 1040s to CPS units by address • Compare linked tax information for respondents and nonrespondents • Compare unit-level results with ZIP code-level results

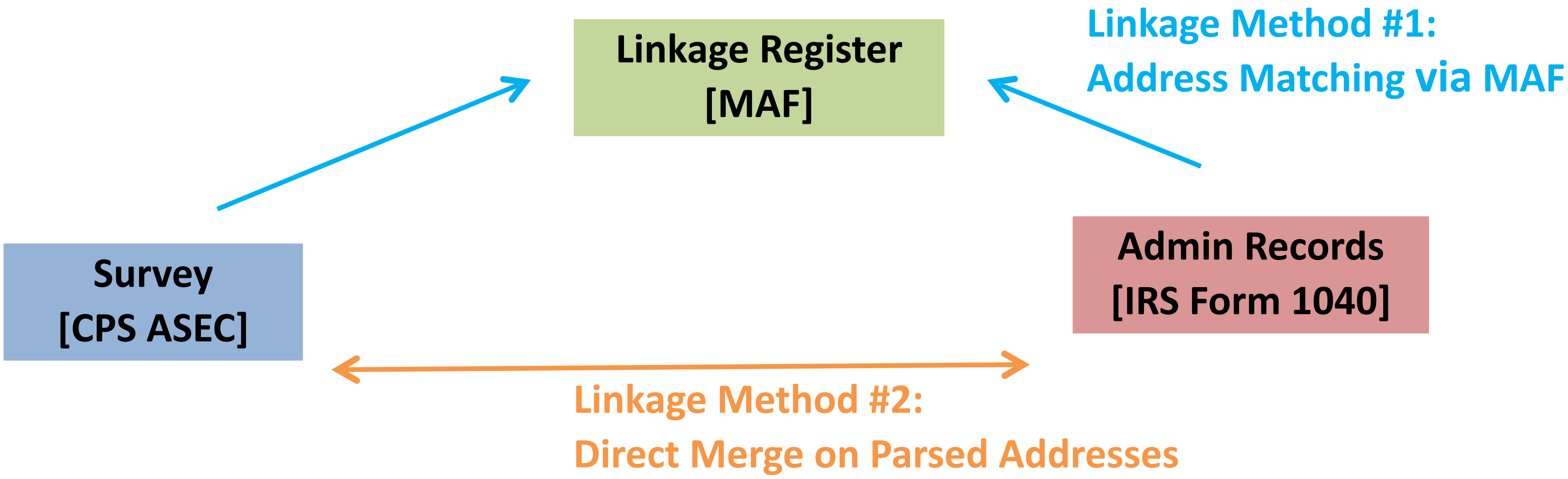
Data: 2011 CPS ASEC

- Annual Social and Economic Supplement (ASEC) of the Current Population Survey (CPS)
- Source of official US poverty rate estimates and household income statistics
  - Nationally representative (with survey weights)
  - 96,958 sampled in 2011 (81,737 eligible units)
  - 75,188 respondent units (Mostly by telephone, some in-person)
- Consider both nonrespondents to CPS Basic and “whole imputes,” who are respondents with entire ASEC imputed
- ASEC sample: March Basic CPS sample, other parts of ASEC
- Base weights account for probability of selection into CPS for all units
- Replicate weights to get SEs with clustering, stratification

Data: Tax Year 2010 IRS Form 1040

- Data from all IRS Form 1040 returns filed during calendar year 2011
- Provided to Census for survey improvement under Title 26, USC
- Nearly 140 million records
- Available information includes AGI, other income measures, marital status, number of dependents, indicators for forms filed, and address

Methods: Linking ASEC Units to 1040s by Address



Linkage Register
[MAF]

Linkage Method #1:
Address Matching via MAF

Survey
[CPS ASEC]

Admin Records
[IRS Form 1040]

Linkage Method #2:
Direct Merge on Parsed Addresses

Methods: Linking ASEC to IRS Form 1040 By Address

- A single 1040 per unit is the modal case.
- We resolve cases where multiple 1040s link to an ASEC household by taking the sum of the linked units’ AGI and the average of other characteristics across the linked 1040s.
  - As a check, also calculate results using average AGI
- We also reweight for non-linking using inverse of predicted probability of linking from a model using sample frame variables

Methods: Testing differences between respondents and nonrespondents

- Assumption: non-linking is not directly related to ASEC nonresponse. It may be related to ASEC or 1040 income or other characteristics as long as the relationship is same for respondents and nonrespondents
    - Implies size of test no higher than nominal size
- Power of tests: depends on relationship between non-linking and income

Response Rates

Table 1: 2011 CPS ASEC and Basic Response Rates by Sample Address List Variables

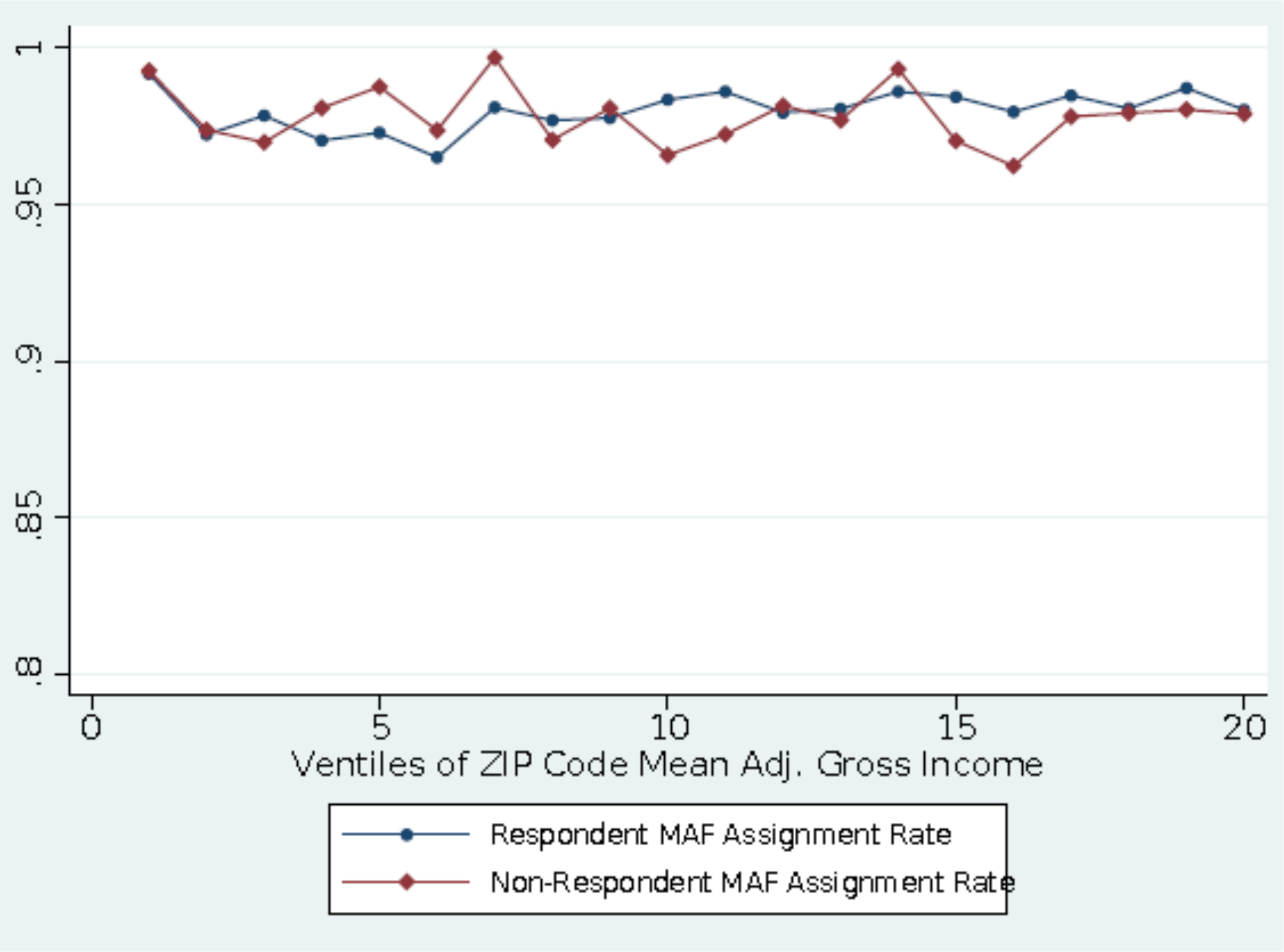
	ASEC response rate	Std. err.	Unwted household count		**March Basic sample response rate**	Std. err.	Unwted household count
Overall	0.921	0.001	81,500		0.911	0.001	58,500
Frame (Part of sample)							
Area	0.938	0.003	9,200		0.934	0.004	6,900
Group quarters	1.000	.	80		1.000	.	70
Permit	0.906	0.003	9,400		0.892	0.004	6,500
Unit	0.921	0.001	63,000		0.911	0.002	45,000
ASEC sample							
March Basic	0.911	0.001	58,500		0.911	0.001	58,500
Mar Hispanic from Nov	0.933	0.004	5,300		.	.	0
Feb month 9	0.893	0.005	4,500		.	.	0
Apr month 9	0.896	0.008	2,300		.	.	0
Feb month 4, 8 split path	0.953	0.004	4,500		.	.	0
Apr month 1, 5 split path	1.000	.	6,400		.	.	0
Tract poverty rate							
20% or more	0.931	0.003	12,500		0.917	0.004	8,000
Under 20%	0.919	0.001	69,500		0.910	0.001	50,500

Rates of MAFID assignment for ASEC

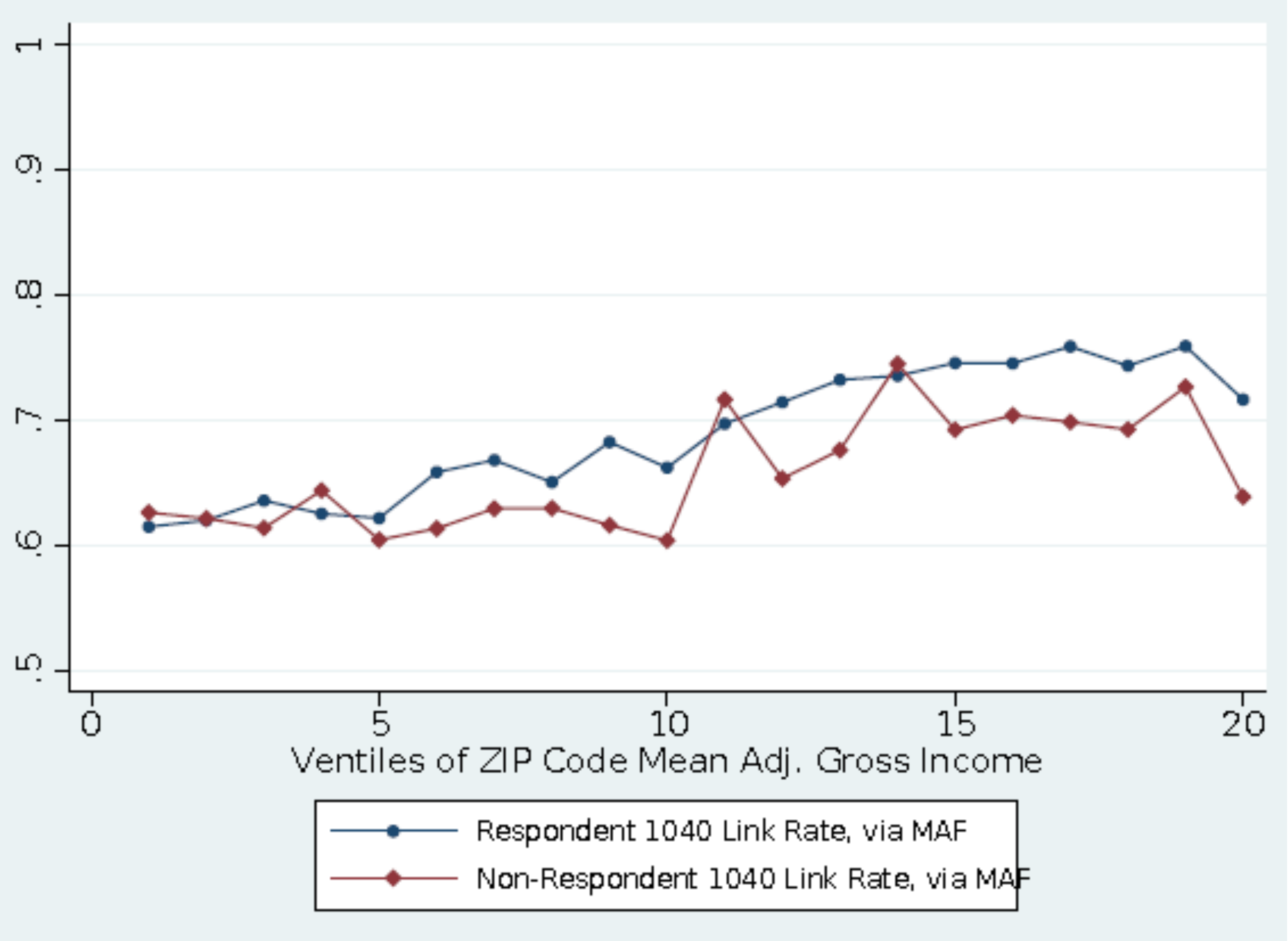
Table 2: Proportions of CPS Households that Link to the Master Address File

	Non-Imputed Respondents	Whole-Imputed Respondents	All Respondents	Nonrespondents	p: (1)=(2)	p: (1)=(4)	p: (3)=(4)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Overall	0.979	0.978	0.979	0.977	0.801	0.229	0.266
Region							
Northeast	0.989	0.990	0.989	0.984	0.832	0.210	0.188
Midwest	0.984	0.986	0.984	0.980	0.519	0.464	0.423
South	0.968	0.964	0.967	0.966	0.317	0.665	0.757
West	0.983	0.981	0.982	0.982	0.702	0.815	0.866
Urban							
Urban	0.986	0.986	0.986	0.980	0.956	0.086	0.090
Rural	0.950	0.948	0.950	0.957	0.721	0.502	0.498
Tract poverty rate							
20% or more	0.974	0.975	0.974	0.987	0.861	0.073	0.081
Less than 20%	0.979	0.978	0.979	0.975	0.721	0.039	0.049
Number of households	47,500	6,000	53,500	5,300			

Rate of MAFID Assignment for ASEC By Ventile of ZIP Code-Mean AGI



Rates of ASEC Linking to 1040s by Ventile of ZIP Code-Mean AGI



Rate of Linking of ASEC to 1040s

Table 3: Proportions of CPS Households that Link to a Form 1040 Record via the Master Address File

	Non-Imputed Respondents	Whole Imputes	All Respondents	Nonrespondents	p: (1)=(2)	p: (1)=(4)	p: (3)=(4)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Overall	0.687	0.670	0.685	0.654	0.032	0.000	0.000
Region							
Northeast	0.666	0.635	0.662	0.639	0.047	0.147	0.211
Midwest	0.724	0.714	0.723	0.719	0.495	0.735	0.791
South	0.675	0.655	0.673	0.641	0.139	0.017	0.025
West	0.685	0.679	0.685	0.638	0.710	0.004	0.005
Urban							
Urban	0.708	0.692	0.706	0.669	0.047	0.000	0.000
Rural	0.604	0.587	0.602	0.580	0.336	0.213	0.246
Tract poverty rate							
20% or more	0.578	0.579	0.578	0.505	0.962	0.001	0.001
Less than 20%	0.706	0.685	0.704	0.678	0.007	0.001	0.003
Number of households	47,500	6,000	53,500	5,300			
Number of linked households	32,000	4,000	36,000	3,400			

What explains the link rate?

- Non-filers: Mortenson et al. (2009) and Heim et al. (2014) estimate that 10-12 percent of individuals and 17 percent of tax units do not appear on 1040s.
- Late filers
- Nonresidential addresses on returns (PO boxes, preparers)
- Complicated/bad addresses

Methods: Testing differences between respondents and nonrespondents

- Assumption: non-linking is not directly related to ASEC nonresponse. It may be related to ASEC or 1040 income or other characteristics as long as the relationship is same for respondents and nonrespondents
- Power of tests: depends on relationship between non-linking and income

Results: Distribution of AGI

Table 4: Characteristics of CPS Respondents and Nonrespondents as Recorded in Tax Records

	Non-Imputed Respondents	Whole Imputes	All Respondents	Nonrespondents	*p*: (1)=(2)	*p*: (1)=(4)	*p*: (3)=(4)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Percentiles of AGI							
1	12	0	7	0	0.991	0.994	0.997
	(41)	(1,098)	(32)	(1,863)			
5	6,959	7,074	6,977	7,761	0.793	0.252	0.264
	(165)	(404)	(137)	(673)			
10	12,587	11,935	12,544	12,792	0.282	0.676	0.619
	(184)	(570)	(175)	(469)			
25	26,932	27,214	26,989	27,626	0.732	0.322	0.356
	(257)	(777)	(237)	(674)			
50	55,115	55,031	55,098	54,746	0.949	0.790	0.797
	(421)	(1,204)	(407)	(1,459)			
75	94,834	95,899	94,971	94,722	0.551	0.946	0.891
	(635)	(1,802)	(629)	(1,949)			
90	144,874	148,196	145,268	150,907	0.362	0.181	0.209
	(1,138)	(3,608)	(1,113)	(4,299)			
95	194,107	198,691	194,656	204,606	0.531	0.108	0.126
	(2,109)	(7,136)	(2,119)	(6,355)			
99	393,341	395,645	393,862	485,099	0.931	0.136	0.134
	(12,999)	(28,181)	(11,953)	(54,670)			
Joint equality test at given percentiles					0.769	0.507	0.577

Non-Income Characteristics of Respondents and Nonrespondents

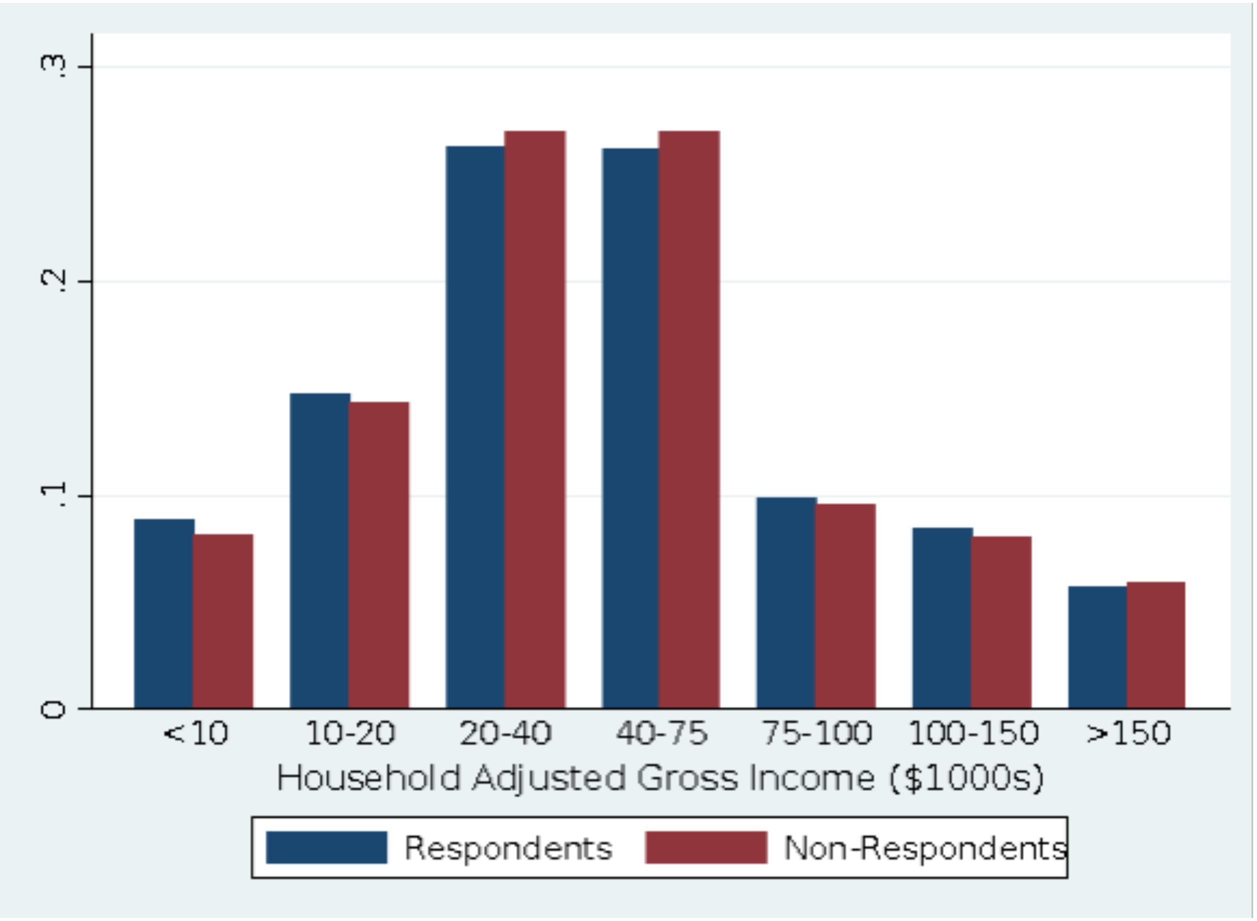
- Mostly fit model that respondent households are those more likely to have someone at home
- Married, those with more children, those on social security more likely to respond
- Households with wage and salary income more likely to respond

Results: Non-Income Characteristics

Table 4, continued: Characteristics of CPS Respondents and Nonrespondents as Recorded in Tax Records

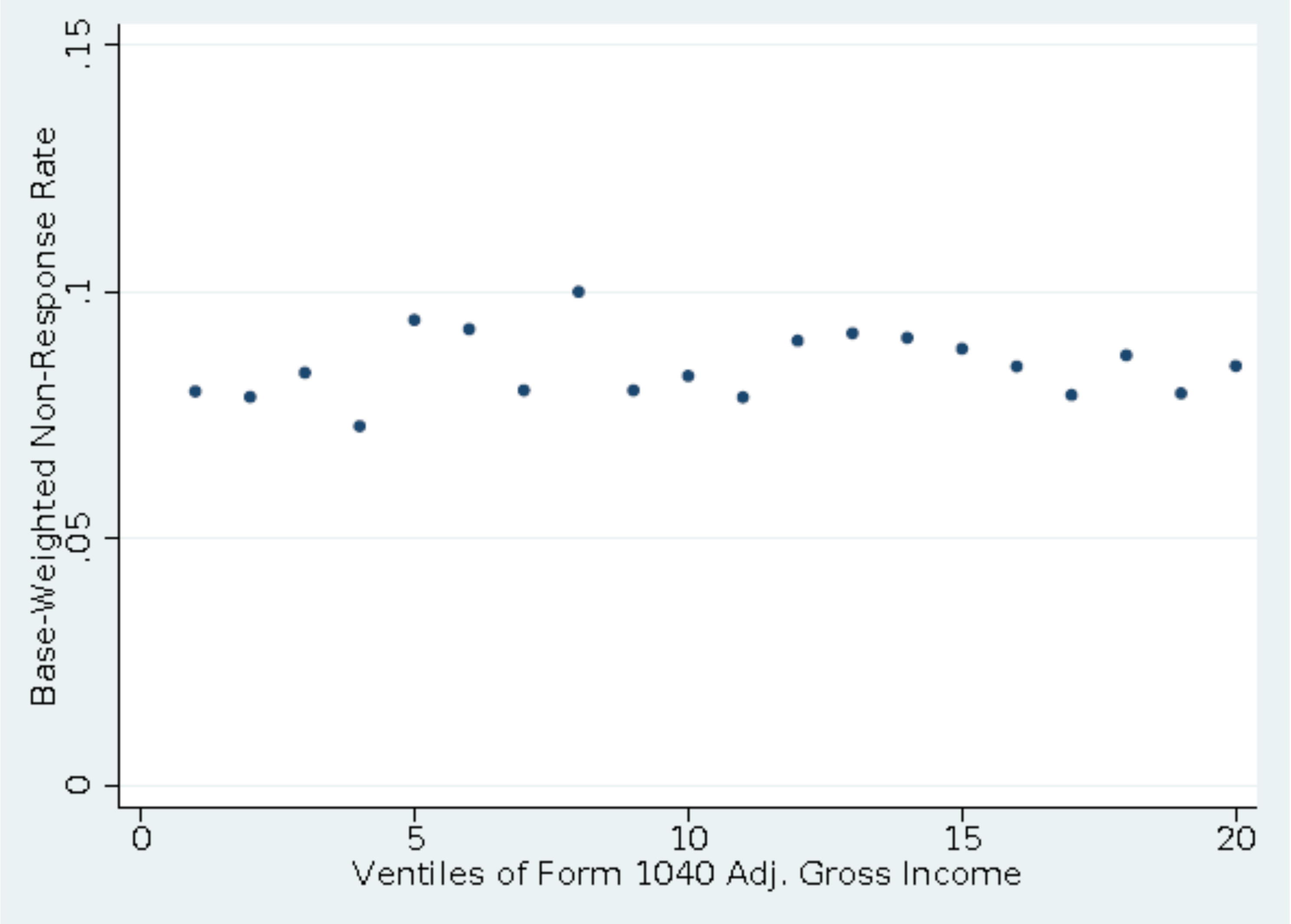
	Non-Imputed Respondents	Whole Imputes	All Respondents	Nonrespondents	*p*: (1)=(2)	*p*: (1)=(4)	*p*: (3)=(4)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Inequality measures							
90/10 ratio	11.5	12.4	11.6	11.8	0.143	0.608	0.736
	(0.2)	(0.6)	(0.4)	(0.5)			
Gini coefficient	0.486	0.505	0.488	0.493	0.323	0.678	0.764
	(0.008)	(0.018)	(0.007)	(0.016)			
Means							
Adjusted gross income	75,328	78,503	75,680	77,184	0.462	0.516	0.594
	(1,237)	(3,936)	(1,114)	(2,569)			
Married filing jointly	0.463	0.452	0.462	0.404	0.211	0.000	0.000
	(0.003)	(0.008)	(0.003)	(0.009)			
Number of child exemptions	0.633	0.655	0.635	0.582	0.212	0.008	0.005
	(0.007)	(0.017)	(0.006)	(0.018)			
Receipt of income sources							
Wage and salary	0.816	0.837	0.818	0.853	0.001	0.000	0.000
	(0.002)	(0.006)	(0.002)	(0.006)			
Interest and dividends	0.490	0.451	0.485	0.460	0.000	0.004	0.013
	(0.003)	(0.008)	(0.003)	(0.010)			
Social security	0.216	0.198	0.214	0.150	0.013	0.000	0.000
	(0.003)	(0.007)	(0.003)	(0.007)			
Number of households	32,000	4,000	36,000	3,400			

Respondents and nonrespondents are the same

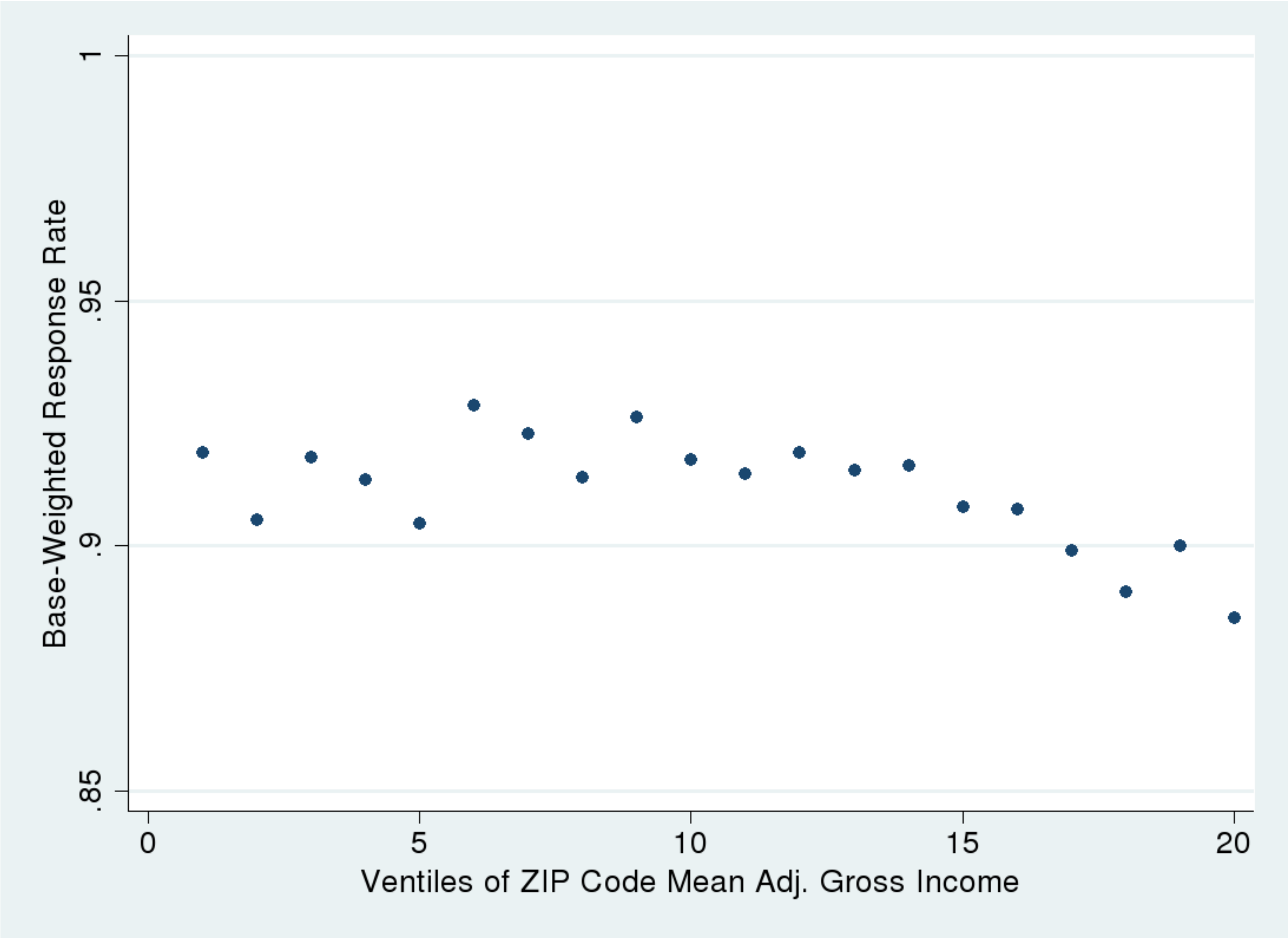


Conclusion: Nonresponse doesn’t bias official income and poverty measures

Results: Response Rate by Ventile of AGI for CPS ASEC units linked to 1040s



Results: Response Rate by Ventile of *ZIP Code-Mean* AGI for 1040-Linked ASEC Units



Source of ZIP Code v. Unit Difference

- Low response rate by low and middle income households in high income ZIP codes
- We tabulate nonresponse rate for cells defined by the interaction of quintiles of ZIP code-level AGI with household-level AGI.
    - Typical nonresponse rate of about 8 percent.
    - 6 cells have a nonresponse rate over 10 percent. 4 are the bottom quintiles of household AGI for those in the top quintile of ZIP code-level AGI.

Implication is that ZIP code approach in Sabelhaus et al. (2015) may be misleading

Source of ZIP Code v. Unit Difference

Nonresponse rate

Population share

	Quintiles of ZIP-Code-Mean Adjusted Gross Income					
Quintiles of Household Adjusted Gross Income	Bottom quintile	2nd quintile	Middle quintile	4th quintile	Top quintile	Overall
Bottom quintile	0.085	0.078	0.069	0.090	0.121	0.085
	0.061	0.047	0.039	0.030	0.023	0.200
2nd quintile	0.086	0.086	0.080	0.093	0.123	0.090
	0.053	0.048	0.042	0.032	0.025	0.200
Middle quintile	0.083	0.083	0.085	0.095	0.126	0.092
	0.041	0.045	0.043	0.041	0.031	0.200
4th quintile	0.084	0.071	0.075	0.074	0.105	0.082
	0.030	0.037	0.044	0.047	0.041	0.200
Top quintile	0.113	0.090	0.085	0.090	0.086	0.089
	0.015	0.023	0.033	0.050	0.079	0.200
Overall	0.087	0.081	0.079	0.088	0.105	0.088
	0.200	0.200	0.200	0.200	0.200	1.000

Combined Sample and Nonresponse Adjustment

	Respondents Base-Weighted (1)	Respondents with Non-Interview Adjustment (2)	Respondents and Nonrespondents Combined Base-Weighted (3)	*p:* (1)=(3) (4)	*p:* (2)=(3) (5)
Percentiles of AGI					
1	7	10	1	0.838	0.763
	(32)	(33)	(27)		
5	6,977	6,977	7,014	0.765	0.765
	(137)	(137)	(131)		
10	12,544	12,554	12,562	0.920	0.963
	(175)	(178)	(164)		
25	26,989	27,009	27,028	0.870	0.937
	(237)	(244)	(238)		
50	55,098	55,228	55,088	0.988	0.736
	(407)	(411)	(412)		
75	94,971	95,224	94,934	0.950	0.642
	(629)	(619)	(610)		
90	145,268	145,712	145,659	0.722	0.952
	(1,113)	(1,108)	(1,069)		
95	194,656	195,365	195,594	0.637	0.903
	(2,119)	(2,031)	(1,965)		
99	393,862	394,506	396,616	0.807	0.833
	(11,953)	(12,126)	(11,531)		

Income by household type (marital status and presence of children)

- Married without children—no significant differences in percentiles or mean
- Unmarried without children—significant differences at middle percentiles (differences are $1-3 thousand), nonrespondents have higher income
- Married with children—only significantly different at 25th percentile; nonrespondents have higher income
- Unmarried with children—no significant differences

Robustness—direct linking

- We standardized and parsed addresses and linked directly using SAS DQ
- Similar results, slightly lower link rate
- Considered trying to increase link rate where we though SAS was having trouble with certain types of addresses, but thought that was too involved a process

Robustness—PIK linking

- Only for non-imputed respondents v. imputed respondents (whole imputes)
- PIK linking has
  - higher link rate,
  - more power for high income households,
  - no significant income differences between non-imputed respondents and whole imputes

Robustness—alternatives to unit sum

- We examine number of linked 1040s by household type
- For full sample, no significant income differences at any percentile when we average 1040s
- For married households, only 5 percent have more than one 1040. Using sum, mean, or max makes little difference
- For unmarried households with children, there are a few percentiles that have whole impute or non-respondent income percentiles significantly different from those for non-imputed respondents when we average 1040s

Robustness—full sample (not just March)

- Results for March cleanest: no way to weight nonrespondents comparably to respondents when bring in other sampled households
- March relevant for monthly: weekly earnings, etc.
- Full sample used in studies of annual earnings; weights not exactly right
- Significant differences between non-imputed respondents and imputed respondents go away

Conclusions

- Little or no evidence from 1040s of bias from unit nonresponse in measurement of income using the CPS Basic or ASEC. Some small differences for whole imputes. Some small differences within household type
- Differences between respondents and nonrespondents on some demographic and economics characteristics
- Fairly different results between household-level and ZIP code-level analyses

Future Work

- Formal bounding arguments
- Linking improvements
  - Additional sources: Information returns, SNAP, etc.
  - Checks on links
  - Resolving multiple link choices
- Contact History Instrument
- Ineligible units

Thank You!

Adam Bee:	charles.adam.bee@census.gov
Graton Gathright:	ggathright@gmail.com
Bruce Meyer:	bdmeyer@uchicago.edu

References

Bollinger, Christopher R., Barry T. Hirsch, Charles M. Hokayem, and James P. Ziliak. 2019. “Trouble in the Tails? Earnings Non-Response and Response Bias across the Distribution.” Journal of Political Economy.

Groves, Robert M. and Emelia Peytcheva. 2008. “The Impact of Nonresponse Rates on Nonresponse Bias.” Public Opinion Quarterly 72: 167-189.

Mah, Ming-Yi and Dean Resnick. 2009. “Preliminary Analysis of Medicaid Enrollment Status in the Current Population Survey.” U.S. Census Bureau.

Sabelhaus, John, David Johnson, Stephen Ash, David Swanson, Thesia Garner, John Greenlees, and Steve Henderson. 2015. “Is the Consumer Expenditure Survey Representative by Income? In *Improving the Measurement of Consumer Expenditures.* University of Chicago Press.

Extra Tables

Selected Related Literature

Prior work on Non-response	Groves & Peytcheva (2008)	More variation in bias from unit nonresponse across estimates within surveys than across surveys.
	Sabelhaus et al. (2015)	Suggest income in top quintile understated in CE Survey because nonresponse higher for those from ZIP Codes with high mean AGI. We are not aware of a study linking to nonrespondent addresses in a major survey
Other approaches	King et al. (2009)	Uses late respondents as proxy for CE Survey nonrespondents.
	Heffetz and Reeves (2016)	Uses difficult to reach respondents as proxy for nonrespondents; use method in several surveys.

Selected Related Literature

Prior work linking to survey frame	Several authors	Special samples
	Mah and Resnick (2009)	Medicaid receipt
	Lin and Schaeffer (1995)	Child support awards
	Kreuter et al. (2010)	Welfare receipt
Recent Census Bureau Papers Adopting our Approach	Mattingly et al. (2016)	Examines the Survey of Income and Program Participation Wave 1 of the 2008 panel. Finds small and insignificant differences between respondent and nonrespondent income mean and percentiles.
	Brummet et al. (2018)	Examines the Consumer Expenditure Interview Survey collected 2013-14. Finds that mean income is higher among nonrespondents than respondents and finds higher nonresponse rates in the extreme tails of income distribution.

Selected Related Literature

Other Related Literature	Bollinger et al. (2019)	*Item* non-response in CPS earnings is higher in the tails of the distribution. Briefly looks at “whole imputes” in an online appendix.
	Hokayem et al. (2016)	*Item* non-response and “whole imputes” in CPS earnings lead to understatement of poverty rate.

Assessing Nonresponse Bias with Linked Data

Y_i^s survey report for unit *i*, not always observed

D_i = 1 when *i* responds, 0 when nonrespondent

Test null that respondent distn $(Y_i^s \mid D_i = 1)$ same as nonrespondent distn $\left(Y_i^s \middle| D_i = 0\right)$

Want link to administrative data such that in linked data nominal size of test (preset size) no greater than true size

When L_i = 1 observe Y_i^a, true value from administrative data

For simplicity initially assume $Y_i^s \equiv Y_i^a$

Key Condition

Theorem 1: If linking satisfies the independent linkage condition

if $(Y_i^a \mid D_i = 1)$ equals $\left(Y_i^a \middle| D_i = 0\right)$ then

$(Y_i^a \mid D_i = 1, L_i = 1)$ equals $\left(Y_i^a \middle| D_i = 0, L_i = 1\right)$

then conventional tests of equality of the respondent and nonrespondent distributions will have the right size.

Violated if linkage depends on D_i but fine if it depends on Y_i^a

Power will depend on the linkage rate and the range of the variable covered

Extension to "Double Sampling"

Let $Y_i^s = Y_i^a + \varepsilon_i$

Results above hold if ε_i is classical measurement error, i.e., is independent of Y_i^a

Now let D_i have three values, 1 for respondents, 0 for nonrespondents,

and 2 for reluctant or late respondents (Groves or Heffetz and Reeves)

Condition for test to have good properties

True distribution Y_i^a same for nonrespondents and reluctant respondents,

i.e., for $D_i = 0$ and $D_i = 2$, and

the distribution of ε_i must not vary with D_i.

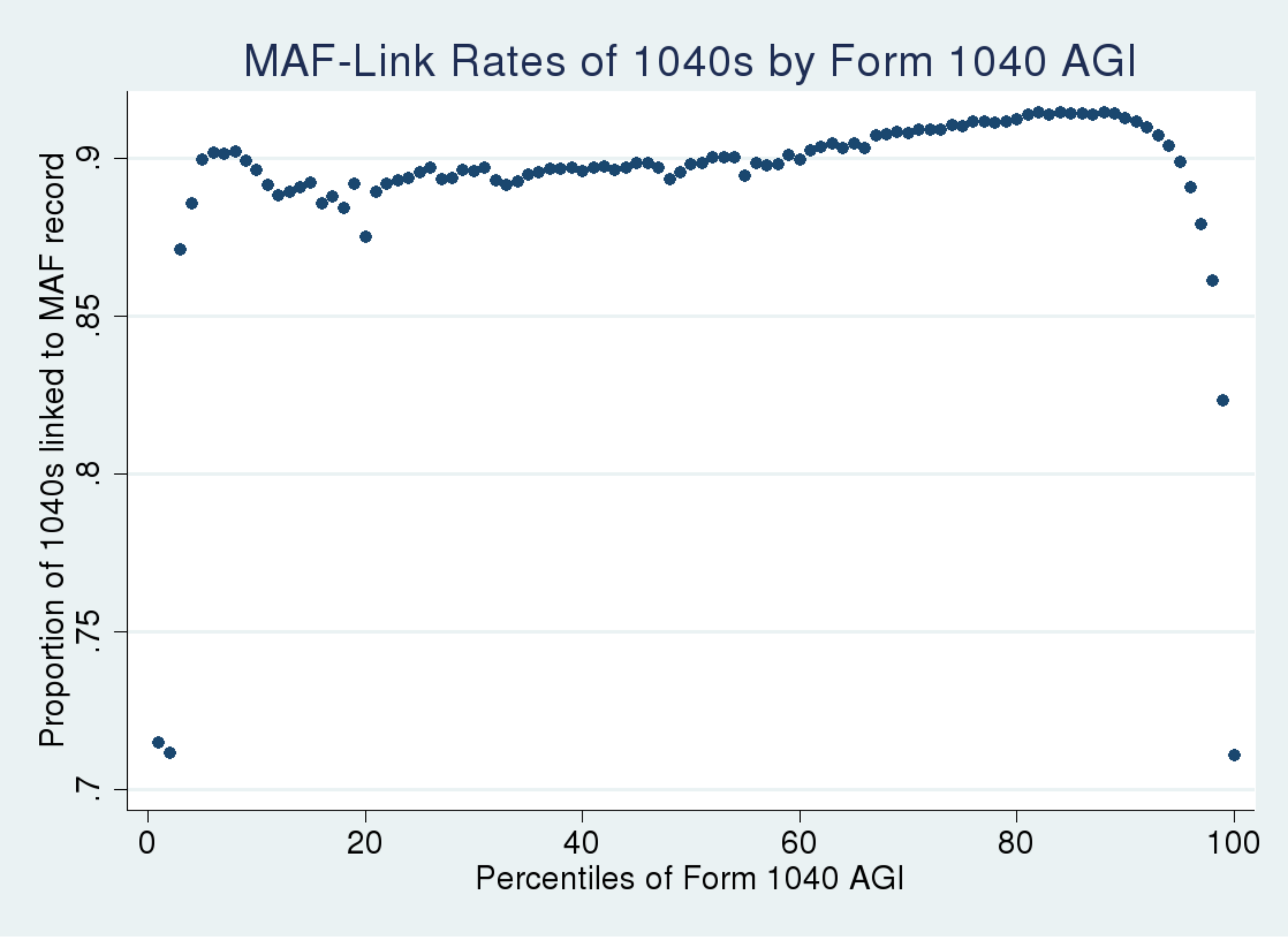
Relaxation of Key Condition and Bounds

Suppose linking is independent except that a fraction α of the population is "off the grid", i.e. not in administrative records or survey
These unlinkable nonrespondents have D_i=0 and L_i=0.

Then α is $P(D_i=0)\ (P(L_i=1|D_i=1) - P(L_i=1|D_i=0))/\ P(L_i=1|D_i=1)$,
Consistently estimated by sample value of proportional difference in link rates between respondents and nonrespondents times the nonresponse rate

We use this expression to provide bounds on percentiles of full-population income distribution. We obtain lower bound by assuming this share has zero AGI and upper bound by assuming this share has AGI of $1,000,000.

Rate of MAFID Assignment for 1040s By Percentile of AGI



ASEC-Reported Income for Linked and Not-Linked Units

ASEC-reported household income	1040-Linked ASEC Respondents	ASEC Respondents Not Linked to 1040	p-value
Mean	$ 74,573	$ 42,341	<.001
Percentiles			
1	$ 0	$ 0	n/a
5	9,605	2,157	<.001
10	15,500	7,280	<.001
25	30,000	13,157	<.001
50	56,080	26,000	<.001
75	96,020	53,288	<.001
90	147,904	94,208	<.001
95	191,680	126,899	<.001
99	338,100	239,067	<.001
Observations	59,000	16,500	