Correcting for Misreporting of Government Benefits

Online Appendix

Nikolas Mittag

September 12, 2018

Appendix A: Tables and Figures

Figure A1: Map of Counties in east (dark) and west (light) NY Sample

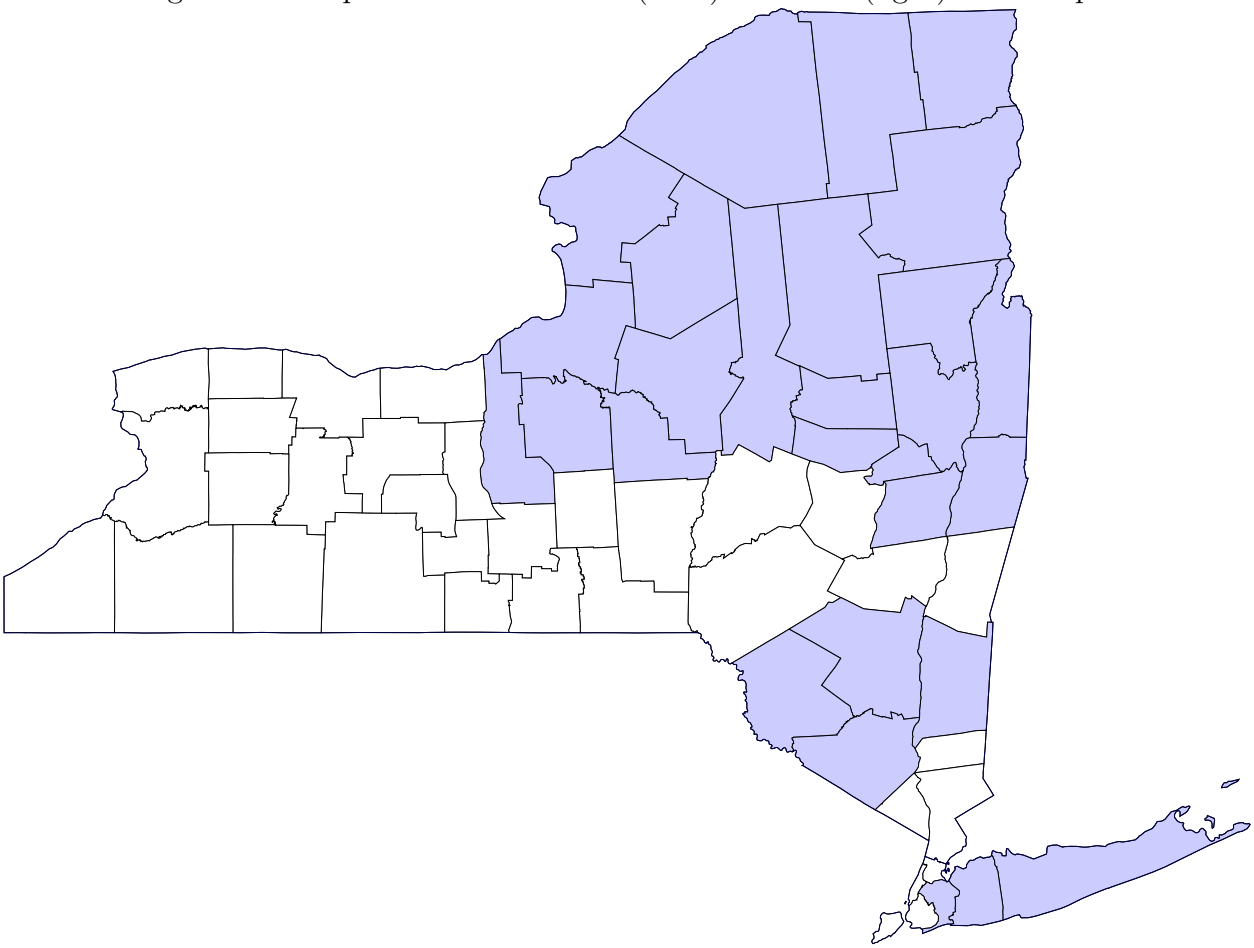

Figure A2: SNAP by Income Relative to the Poverty Line, U.S. 2010

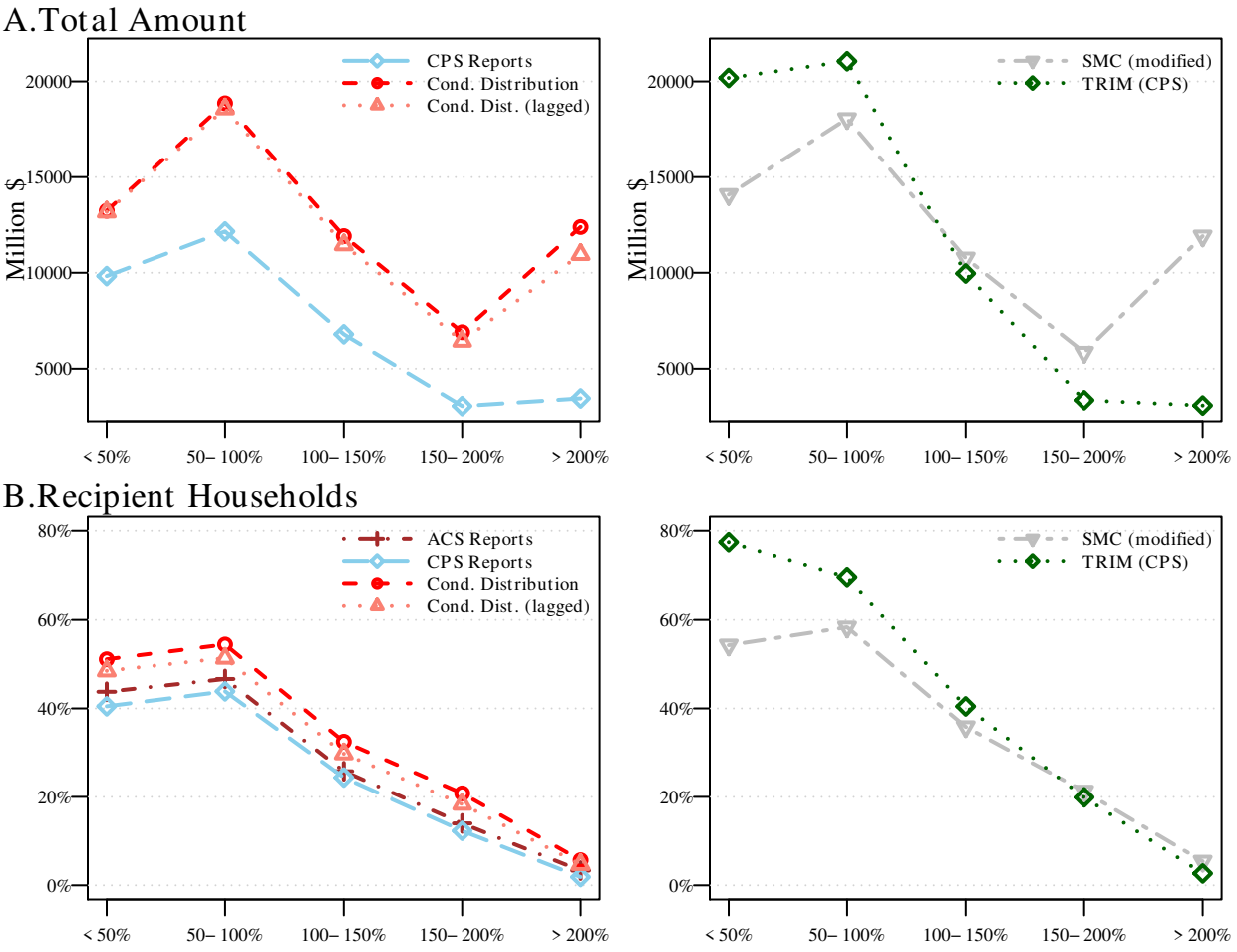

Table A1: The Determinants of an ACS Household having a PIK, Probit Coefficients

	(1)	(2)	(3)	(4)
Year	2009		2010	
	Coef.	SE	Coef.	SE
Income/Poverty Line	0.015	0.002	0.010	0.002
Age 18-29	-0.128	0.029	-0.023	0.032
Age 30-39	-0.098	0.026	-0.095	0.027
Age 50-59	0.143	0.026	0.089	0.027
Age 60-69	0.275	0.030	0.159	0.031
Age 70 or Older	0.360	0.032	0.218	0.034
Number of Persons in HH	0.050	0.010	0.070	0.011
Number of Persons Under 18	0.019	0.018	-0.017	0.019
Not Married, no Children	-0.133	0.035	-0.103	0.036
Not Married, Children	0.194	0.036	0.045	0.036
Married, no Children	-0.045	0.037	-0.077	0.038
Unemployed	0.030	0.038	0.087	0.040
Not in Labor Force	-0.104	0.023	-0.071	0.024
Not a U.S. Citizen	-0.511	0.024	-0.325	0.027
Hispanic	0.038	0.024	0.060	0.026
White	0.103	0.018	0.071	0.019
Less Than High School	-0.089	0.027	-0.081	0.029
High School Degree	-0.164	0.021	-0.124	0.022
College or More	0.086	0.022	0.060	0.023
Disability	0.070	0.045	0.075	0.047
Disabled, not Working	0.061	0.052	0.056	0.055
Speaks Poor English	-0.273	0.034	-0.180	0.036
Speaks no English	-0.387	0.053	-0.333	0.055
Public Assistance (Reported)	0.065	0.046	0.065	0.049
Constant	1.290	0.046	1.367	0.049
Number of Observations	107,237		106,655	

Notes: The dependent variable is an indicator of whether someone in the household was assigned a PIK. All analyses are conducted using household weights. Individual characteristics refer to the household head. The omitted family type is married with children, the omitted age category is 40-49 and the omitted education category is some college.

Table A2: Summary Statistics Linked Data

Sample	Linked NY Sample				West NY	
Year	2009		2010		2010	
	Mean	SD	Mean	SD	Mean	SD
Admin. SNAP Receipt	0.156	0.363	0.179	0.383	0.174	0.379
Admin. SNAP Amount	435.7	1,320.9	601.9	1,662.7	575.1	1,592.1
SNAP Receipt Reported	0.128	0.334	0.141	0.348	0.139	0.346
SNAP Receipt Imputed	0.009	0.093	0.013	0.113	0.014	0.116
East NY Sample	0.487	0.500	0.487	0.500	–	–
Income/Poverty Line	5.042	6.715	4.809	6.249	4.494	4.922
Age 18-29	0.095	0.293	0.102	0.302	0.087	0.281
Age 30-39	0.170	0.376	0.168	0.374	0.163	0.369
Age 50-59	0.209	0.407	0.210	0.407	0.217	0.412
Age 60-69	0.150	0.357	0.153	0.360	0.154	0.361
Age 70 or Older	0.157	0.364	0.158	0.364	0.160	0.366
Any Income From Capital	0.246	0.431	0.222	0.416	0.215	0.410
# of Persons in HH	2.512	1.507	2.544	1.538	2.662	1.563
# of Children in HH	0.597	1.036	0.596	1.037	0.642	1.047
Not Married, no Children	0.455	0.498	0.455	0.498	0.417	0.493
Not Married, Children	0.098	0.297	0.099	0.298	0.105	0.306
Married, no Children	0.254	0.435	0.255	0.436	0.269	0.444
Linguistic Isolation	0.078	0.268	0.078	0.269	0.080	0.271
# of Persons Employed	1.715	1.479	1.710	1.486	1.792	1.523
Anyone in HH Employed	0.770	0.421	0.766	0.423	0.772	0.420
Elderly or Disabled in HH	0.443	0.497	0.442	0.497	0.459	0.498
Single Household	0.294	0.456	0.288	0.453	0.258	0.438
Unemployed	0.045	0.207	0.051	0.219	0.052	0.223
Not in Labor Force	0.317	0.465	0.322	0.467	0.323	0.468
Female	0.501	0.500	0.505	0.500	0.496	0.500
Not a U.S. Citizen	0.092	0.289	0.096	0.294	0.099	0.298
White	0.726	0.446	0.713	0.452	0.706	0.455
Less than High School	0.135	0.342	0.133	0.340	0.138	0.344
High School Degree	0.256	0.437	0.254	0.435	0.270	0.444
College or More	0.347	0.476	0.349	0.477	0.313	0.464
Disabled	0.154	0.361	0.148	0.355	0.151	0.358
Disabled, not Working	0.120	0.325	0.118	0.323	0.120	0.325
Speaks English Poorly	0.015	0.122	0.015	0.123	0.015	0.121
Speaks No English	0.050	0.219	0.051	0.220	0.054	0.226
Number of Observations		101,335		101,683		49,577

Notes: Individual characteristics refer to the household head. All statistics are at the household level using household weights adjusted for incomplete linkage.

Table A3: Summary Statistics ACS Public Use Data, 2010

Sample	NY		West NY		U.S.	
	Mean	SD	Mean	SD	Mean	SD
SNAP Receipt Reported	0.138	0.345	0.135	0.342	0.119	0.323
SNAP Receipt Imputed	0.016	0.125	0.016	0.127	0.013	0.114
East NY Sample	0.489	0.500	–	–	–	–
Income/Poverty Line	4.822	5.655	4.537	4.635	4.251	4.326
Age 18-29	0.102	0.302	0.088	0.283	0.115	0.319
Age 30-39	0.169	0.375	0.164	0.371	0.172	0.377
Age 50-59	0.210	0.407	0.217	0.412	0.205	0.404
Age 60-69	0.152	0.359	0.153	0.360	0.152	0.359
Age 70 or Older	0.157	0.364	0.159	0.366	0.152	0.359
Any Income From Capital	0.221	0.415	0.213	0.409	0.218	0.413
# of Persons in HH	2.539	1.538	2.655	1.561	2.520	1.483
# of Children in HH	0.595	1.037	0.640	1.048	0.624	1.060
Not Married, no Children	0.455	0.498	0.419	0.493	0.417	0.493
Not Married, Children	0.099	0.298	0.106	0.308	0.097	0.296
Married, no Children	0.255	0.436	0.268	0.443	0.286	0.452
Linguistic Isolation	0.083	0.275	0.083	0.276	0.046	0.210
# of Persons Employed	1.704	1.485	1.787	1.516	1.714	1.482
Anyone in HH Employed	0.765	0.424	0.773	0.419	0.767	0.423
Elderly or Disabled in HH	0.441	0.497	0.457	0.498	0.444	0.497
Single Household	0.291	0.454	0.260	0.439	0.274	0.446
Unemployed	0.051	0.219	0.052	0.222	0.055	0.228
Not in Labor Force	0.323	0.468	0.322	0.467	0.314	0.464
Female	0.505	0.500	0.494	0.500	0.470	0.499
Not a U.S. Citizen	0.096	0.295	0.100	0.300	0.065	0.246
White	0.707	0.455	0.700	0.458	0.782	0.413
Less Than High School	0.135	0.342	0.140	0.347	0.125	0.331
High School Degree	0.252	0.434	0.265	0.441	0.262	0.440
College or More	0.348	0.476	0.314	0.464	0.303	0.460
Disabled	0.149	0.356	0.152	0.359	0.164	0.370
Disabled, not Working	0.120	0.325	0.122	0.327	0.126	0.332
Speaks English Poorly	0.052	0.222	0.055	0.229	0.030	0.171
Speaks no English	0.015	0.123	0.014	0.119	0.010	0.101
Number of Observations	74,105		36,192		1,203,777	

Notes: Individual characteristics refer to the household head. All statistics are at the household level using household weights.

Table A4: Summary Statistics CPS

Sample	NY		U.S.	
	Mean	SD	Mean	SD
SNAP Receipt Reported	0.116	0.320	0.100	0.300
Reported SNAP Amount	388.1	1,470.6	298.7	1,225.5
Income/Poverty Line	4.421	4.836	4.212	4.456
Number of Observations	6,689		151,368	

Notes: To make the time period comparable to the 2010 ACS, this table pools the 2010 and 2011 CPS ASEC. All statistics are at the household level using household weights adjusted for pooling years.

Table A5: Parameter Estimates of the Conditional Distribution, Linked NY ACS, 2009 and 2010

Year	2009				2010			
	Mass Point		Amounts		Mass Point		Amounts	
	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE
SNAP Receipt Reported	-2.51	0.002	1,156.9	6.9	-2.50	0.002	1,109.8	6.8
SNAP Receipt Imputed	-0.50	0.006	534.5	20.8	-0.42	0.005	386.6	18.4
Income Relative to Poverty Line Intercepts								
50-100 %	-0.17	0.015	376.0	29.2	0.58	0.015	383.4	30.6
100-150 %	-0.24	0.021	-777.9	53.3	-0.29	0.020	-532.7	54.9
150-200 %	-0.38	0.032	427.5	103.3	-0.62	0.029	-1,188.9	97.6
>200 %	0.56	0.006	-1,629.3	15.4	0.47	0.005	-1,299.1	15.4
Income Relative to Poverty Line Slopes								
≤ 50 %	-0.40	0.018	-190.6	34.5	-0.05	0.017	5.6	36.6
50-100 %	0.12	0.018	-1,264.8	35.6	-0.85	0.018	-934.1	37.1
100-150 %	0.29	0.016	-371.9	42.5	0.37	0.016	-359.7	43.6
150-200 %	0.42	0.018	-1,096.7	59.2	0.52	0.017	-82.8	56.0
>200 %	0.02	0.000	-22.8	2.2	0.04	0.000	-30.7	2.1
Age 18-29	-0.08	0.003	242.5	9.0	0.00	0.003	248.1	9.3
Age 30-39	-0.05	0.003	242.4	8.0	-0.14	0.003	38.6	8.2
Age 50-59	0.01	0.003	-42.4	8.7	0.04	0.003	-210.1	9.0
Age 60-69	0.12	0.004	-171.3	11.2	0.17	0.004	-132.7	11.6
Age 70 or Older	0.36	0.004	-258.5	11.9	0.34	0.004	-315.9	12.4
Any income from capital	0.33	0.003	-158.0	13.2	0.35	0.003	-594.7	13.9
# of Persons in HH	-0.20	0.001	438.2	3.5	-0.22	0.001	538.4	3.4
# of Children in HH	-0.03	0.002	879.3	4.8	-0.02	0.002	1,047.6	4.7
Not Married, no Children	-0.46	0.004	274.1	10.7	-0.55	0.000	475.5	10.6
Not Married, Children	-0.52	0.003	-89.2	8.7	-0.55	0.003	72.6	8.6
Married, no Children	-0.08	0.004	127.8	12.7	-0.22	0.004	481.7	12.6
Linguistic Isolation	-0.03	0.003	55.9	8.9	-0.05	0.003	-1.9	9.2
# of Persons Employed	0.09	0.002	-300.7	4.4	0.09	0.001	-391.2	4.3
Anyone in HH Employed	0.04	0.004	157.5	10.3	0.02	0.003	147.5	10.5
Elderly or Disabled in HH	-0.19	0.003	-141.4	7.6	-0.14	0.003	-405.9	7.8
Single Household	0.19	0.003	-1,487.1	11.0	0.23	0.003	-1,695.8	11.6
Unemployed	-0.05	0.004	-402.4	10.5	-0.10	0.004	-499.8	10.3
Not in Labor Force	-0.05	0.003	-115.4	8.3	-0.09	0.003	-5.7	8.4
Female	-0.06	0.002	-36.3	6.1	-0.06	0.002	56.2	6.2
Not a U.S. Citizen	0.14	0.003	-313.5	8.1	0.27	0.003	-539.7	8.4
White	0.42	0.002	-183.3	5.4	0.41	0.002	-201.5	5.6
Less Than High School	-0.17	0.003	182.0	7.2	-0.10	0.003	202.6	7.6
High School Degree	-0.13	0.002	72.4	7.1	-0.09	0.002	162.2	7.2
College or More	0.17	0.003	-76.1	10.2	0.20	0.003	46.7	10.3
Disabled	0.04	0.005	57.7	15.0	-0.07	0.005	329.4	16.3
Disabled, not Working	-0.17	0.006	-121.9	15.9	-0.07	0.006	-277.0	17.3
Speaks English Poorly	-0.06	0.004	301.8	9.3	-0.16	0.004	381.2	9.5
Speaks no English	-0.15	0.006	510.2	13.1	0.10	0.006	477.2	14.5
Constant	1.61	0.008	651.7	20.7	1.59	0.007	695.2	20.7
Conditional Variance (σ)	2,073	2.3			2,358	2.4		
Left Truncation Point	10	0.002			16	0.003		
Number of Observations	101,335				101,683			

Notes: Parameter estimates of truncated normal conditional distributions. The parameters in the columns labeled "Mass Point" determine the probability that the household does *not* receive SNAP and can be interpreted like probit coefficients. The parameters in the columns labeled "Amounts" determine amounts conditional on receipt. Individual characteristics refer to the household head. All analyses use household weights adjusted for incomplete linkage.

Table A6: Models of Employment, Probit Coefficients

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
		Reports		Cond. Distribution		SMC Method		
	Linked Data	ACS PUMS	CPS	current par.	lagged par. adj.	orig.	modified	TRIM CPS
SNAP Receipt	-0.773	-0.870	-0.893	-0.764	-0.759	-1.003	-0.854	-0.817
SNAP Amount ($1000)	-0.009	NA	-0.004	-0.010	-0.011	0.011	0.008	-0.041
Age 18-29	-0.120	-0.115	-0.296	-0.122	-0.111	-0.101	-0.114	-0.293
Age 30-39	0.051	0.046	-0.014	0.052	0.046	0.053	0.044	0.046
Age 50-59	-0.209	-0.196	-0.180	-0.206	-0.202	-0.187	-0.198	-0.116
Age 60-69	-1.045	-1.013	-0.895	-1.024	-1.018	-1.014	-1.016	-0.968
Age $\geq$ 70	-2.155	-2.119	-2.073	-2.137	-2.135	-2.139	-2.126	-2.083
Female	-0.223	-0.227	-0.207	-0.229	-0.230	-0.226	-0.232	-0.269
White	-0.099	-0.054	0.084	-0.085	-0.092	-0.095	-0.055	0.034
Less Than High School	-0.370	-0.354	-0.549	-0.357	-0.346	-0.258	-0.350	-0.410
High School Graduate	-0.120	-0.129	-0.214	-0.121	-0.117	-0.101	-0.127	-0.077
College Graduate and Beyond	0.237	0.236	0.135	0.226	0.229	0.209	0.227	0.136
Single, no children	-0.081	-0.094	-0.099	-0.098	-0.097	-0.091	-0.099	-0.043
Single, Children	0.107	0.101	0.091	0.104	0.106	0.189	0.094	0.326
Married, Children	-0.036	-0.045	-0.095	-0.046	-0.053	-0.078	-0.061	-0.061
Intercept	1.265	1.199	1.065	1.254	1.258	1.271	1.228	1.089

Notes: The dependent variable is an indicator of whether the household head is employed. See table 1 for notes on the methods. All analyses use household weights (adjusted for PIK probability in column 1).

Table A7: Parameter Estimates of the Conditional Distribution, Linked ACS NY Subsamples 2010

Sample	East New York				West New York			
	Mass Point		Amounts		Mass Point		Amounts	
	Coef.	SE	Coef.	SE	Coef.	SE	Coef.	SE
SNAP Receipt Reported	-2.49	0.003	1,145.8	9.7	-2.52	0.003	1,048.3	9.3
SNAP Receipt Imputed	-0.31	0.007	257.9	26.3	-0.55	0.007	516.0	25.0
Income Relative to Poverty Line Intercepts								
50-100 %	0.45	0.021	419.9	43.3	0.66	0.022	379.1	42.4
100-150 %	-0.37	0.028	573.7	80.0	-0.20	0.030	-1,584.3	73.4
150-200 %	-0.51	0.041	-909.5	137.7	-0.73	0.042	-1,000.9	135.1
>200 %	0.32	0.007	-1,177.6	21.9	0.57	0.008	-1364.8	21.3
Income Relative to Poverty Line Slopes								
$\leq$ 50 %	-0.12	0.023	-51.6	50.9	-0.01	0.027	51.8	51.9
50-100 %	-0.77	0.025	-881.3	52.4	-0.89	0.026	-1,003.6	51.5
100-150 %	0.36	0.022	-1,030.8	63.4	0.35	0.023	302.7	58.2
150-200 %	0.42	0.023	-240.4	78.7	0.60	0.024	-169.7	77.7
>200 %	0.05	0.000	-11.0	3.0	0.03	0.001	-40.1	2.9
Age 18-29	-0.05	0.004	-35.4	13.5	0.04	0.005	478.4	12.6
Age 30-39	-0.16	0.004	-24.3	12.1	-0.13	0.004	112.7	10.8
Age 50-59	0.02	0.004	-283.4	13.2	0.07	0.004	-123.6	11.9
Age 60-69	0.15	0.005	-284.7	16.8	0.20	0.005	1.5	15.7
Age 70 or Older	0.33	0.006	-489.7	18.0	0.37	0.006	-148.2	16.8
Any Income From Capital	0.31	0.004	-840.8	20.0	0.39	0.004	-339.3	18.9
# of Persons in HH	-0.23	0.002	409.7	4.9	-0.20	0.002	652.3	4.7
# of Children in HH	0.00	0.003	997.6	7.0	-0.04	0.002	1,056.8	6.3
Not Married, no Children	-0.51	0.005	415.1	15.4	-0.58	0.005	550.0	14.2
Not Married, Children	-0.56	0.005	206.9	12.6	-0.54	0.004	41.1	11.5
Married, no Children	-0.19	0.006	328.2	18.2	-0.24	0.005	621.1	17.1
Linguistic Isolation	-0.03	0.005	-220.8	13.5	-0.07	0.005	181.1	12.3
# of Persons Employed	0.09	0.002	-308.4	6.3	0.08	0.002	-447.1	5.8
Anyone in HH Employed	0.05	0.005	81.2	15.0	-0.02	0.005	189.3	14.5
Elderly or Disabled in HH	-0.15	0.004	-140.2	11.7	-0.13	0.004	-612.5	10.2
Single Household	0.16	0.004	-1,615.7	16.3	0.29	0.004	-1,731.4	16.0
Unemployed	-0.13	0.006	-273.8	15.5	-0.08	0.005	-677.0	13.3
Not in Labor Force	-0.08	0.004	-11.5	12.0	-0.10	0.004	15.3	11.5
Female	0.01	0.003	-66.9	9.0	-0.12	0.003	155.2	8.4
Not a U.S. Citizen	0.18	0.004	-488.4	12.4	0.35	0.004	-534.8	10.9
White	0.35	0.003	-349.1	8.0	0.46	0.003	-82.8	7.6
Less Than High School	-0.04	0.004	289.9	10.9	-0.16	0.004	130.0	10.2
High School Degree	-0.04	0.003	159.0	10.4	-0.13	0.003	178.5	9.7
College or More	0.20	0.004	160.8	14.6	0.20	0.004	-75.5	14.0
Disabled	-0.13	0.007	146.2	24.0	0.00	0.007	528.1	21.4
Disabled, not Working	0.05	0.008	-118.4	25.3	-0.21	0.008	-452.5	22.9
Speaks English Poorly	-0.13	0.005	306.1	14.1	-0.18	0.005	453.1	12.5
Speaks no English	0.23	0.009	495.4	20.7	-0.02	0.009	408.7	20.0
Constant	1.70	0.010	1,013.5	29.5	1.53	0.010	370.9	28.6
Sigma	2,221	3.4			2,442	3.1		
Left Truncation Point	16	0.004			16	0.004		
Number of Observations	49,577				52,106			

Notes: East NY combines the eastern counties of the NYC metro area and the eastern counties of upstate NY. West NY contains the remainder of the state. Parameter estimates of truncated normal conditional distributions. The parameters in the columns labeled "Mass Point" determine the probability that the household does *not* receive SNAP and can be interpreted like probit coefficients. The parameters in the columns labeled "Amounts" determine amounts conditional on receipt. Individual characteristics refer to the household head. All analyses use household weights adjusted for incomplete linkage.

Table A8: Extrapolating SNAP by Income in % of the Poverty Line From East to West NY

Income in % of HH Poverty Line	≤50%	50-100%	100-150%	150-200%	>200%
Total Amount Received (in Million $)					
Linked Data	379	575	372	191	496
Reporting Rates					
Conditional Distribution	103%	98%	98%	100%	97%
Cond. Distribution (adjusted)	105%	99%	99%	101%	97%
SMC Method (modified)	101%	98%	111%	134%	103%
Percentage of Households Receiving SNAP					
Linked Data	55.8%	61.1%	39.4%	23.1%	7.1%
Reporting Rates					
ACS Reports	88%	89%	83%	74%	61%
Conditional Distribution	98%	99%	97%	99%	97%
Cond. Distribution (adjusted)	101%	100%	99%	101%	99%
SMC Method (modified)	105%	108%	109%	115%	100%
Share of Households					
Linked Data	4.4%	7.8%	8.3%	8.2%	71.4%
ACS (all other methods)	4.6%	7.6%	8.1%	8.2%	71.0%

Notes: Estimates based on the 2010 ACS. East NY combines the eastern counties of the NYC metro area and the eastern counties of upstate NY. West NY contains the remainder of the state. Columns defined based on annual reported household cash income divided by the household poverty line. The rows for each measure contain the same methods as the columns in table 4, see the notes there. The first row of each panel contains the estimate from the linked data. The remaining rows contain estimates from the respective method divided by the same statistic from the linked data. All analyses are conducted using household weights (adjusted for incomplete linkage in the linked data).

Appendix B: Data Sources and Linkage

I use three types of data: survey data, administrative records and aggregate statistics. The main source of survey data is the 2010 ACS, which covers calendar years 2009 and 2010. U.S. Census Bureau (2008) provides detailed information on the ACS. All years refer to ACS survey years. To extrapolate across time, I also use the 2009 ACS. There are two versions of the ACS, the PUMS and a restricted internal version. The PUMS data have been edited for confidentiality and contain about 75% of households from the restricted internal ACS. The linked data were created from the internal ACS, but the PUMS data would be used in practice. Results using the internal ACS are very similar. The ACS asks about SNAP receipt in the 12 months before the interview, but does not ask about amounts received. To evaluate survey error in analyses that require benefit amounts, I also report results using the CPS Annual Economic and Social Supplement. U.S Department of Labor (2002) provides detailed information on the CPS. Comparability of the ACS and CPS may be a concern, but they are samples from the same population and variable definitions are similar. To match the ACS reference period, I pool the two corresponding years of the CPS.

All analyses restrict the sample to households, i.e. they exclude group quarters. The analyses in section IV use the NY subsample of the ACS and CPS. Extrapolation within NY state in section V splits the NY ACS sample into east and west NY. To avoid extrapolating from NYC to upstate NY or vice versa, I split both NYC and upstate NY into eastern and western PUMAs. See Appendix Figure A1 for a map. The CPS and TRIM are not representative below the state level, so I do not provide estimates for substate areas.

The second data source is administrative microdata. I use administrative records of all SNAP payments by NY OTDA. They contain payment dates, amounts, basic demographics, addresses and individual identifiers. The records include every individual on a SNAP case in NY. Linkable individual IDs for both the survey data and the administrative records are obtained using the PVS as discussed in section II and in Wagner and Layne (2014). Records are linked to the survey at the individual level and then aggregated to the household level.

I use the ACS household definition, which differs from the definition of a SNAP household. I consider a household to receive SNAP if any member received SNAP according to the administrative data in the reference period of the ACS (12 months before the interview). The administrative records contain payment dates, so I match this reference period exactly.

Working at the household level allows me to correctly classify households as long as I do not fail to match *all* true recipient members to the administrative records.[1] I cannot link administrative records to survey households in which no member has a PIK, so all analyses using administrative receipt variables are based on the sample of households in which at least one member has a PIK. 94 perrcent of households in the NY ACS contain at least one member with a PIK. Despite the low rate of unlinkable records, PIKs are not missing completely at random in the survey data. To restore representativeness of the linked data for the NY survey population, I use inverse probability weighting (Wooldridge, 2007). I estimate a probit model to predict the probability that a household has a PIK. I then multiply the ACS survey weights by the inverse of this predicted probability. The probit estimates are in Appendix Table A1. Inverse probability weighting assumes that, conditional on the covariates in the probit model, whether a household has a PIK or not does not predict receipt or reporting. See Meyer and Mittag (2018) for further discussion.

The third data source is official aggregates of total SNAP recipients and amounts received. I use these numbers for two purposes: First, to improve estimates by constraining the corrections to match official aggregates. For example, both the SMC method and the adjusted conditional distribution method match state totals in the extrapolation to the entire U.S. Second, I use the official aggregates for smaller geographic areas as a benchmark to evaluate extrapolation. Annual statistics on the number of recipients are available from the U.S. Department of Agriculture (USDA) and amounts received are available from the BEA for the entire U.S., by state and by county. The USDA uses a different household defini-

[1] I cannot link households in which all members with a PIK are true non-recipients, but there are true recipients among those without a PIK. Usually only few PIKs are missing per household, as 89 percent of individuals are PIKed, and only few non-recipients cohabit with recipients. See Meyer and Mittag (forthcoming) and Meyer, Mittag and Goerge (2018) for arguments why these exceptions should be uncommon.

tion and publishes average monthly participation instead of annual participation. For NY, the number of participating households is 5-8 percent lower in the linked data for average monthly participation than for annual participation. I use this factor to make the official aggregates for other states comparable to the survey estimates.

To provide a benchmark for the survey, these numbers need to match both the time period and the geography of the survey. The USDA numbers refer to calendar years, so pooling two years makes them comparable to the survey. The BEA numbers refer to fiscal years, while the surveys combine two calendar years. To make the time periods comparable, I take the weighted average of the annual official numbers where the weights correspond to the fraction of the year included in the survey period as in Meyer, Mok and Sullivan (2015). The BEA numbers are available at the county and state level, but neither the ACS PUMS nor the CPS contains county identifiers. The smallest level of geography in the ACS are public use microdata areas (PUMAs), which sometimes split or combine counties. To find the smallest level of geography for which one can obtain numbers from both the ACS and the county-level BEA statistics, I sort counties into groups of counties that minimize the population size of each county group subject to the constraint that the boundaries of the county groups do not intersect with PUMA boundaries. I then aggregate the BEA numbers to the county group level. The CPS is not representative of sub-state areas except for large MSAs. To evaluate the performance of the CPS and TRIM in the extrapolation to the entire U.S., I also aggregate the BEA numbers for large MSAs that can be identified in both data sources. Being identifiable in the data requires an MSA to be large and well aligned with both PUMA and county borders. I only use MSAs with more than half a million inhabitants and exclude MSAs with more than 1 percent of the population in rural PUMAs to ensure comparability to the ACS.

I use the aggregates from the BEA and USDA in the extrapolation to other states in section V, but use aggregates calculated from the linked data for the extrapolation within NY. Aggregates from the linked data are available for smaller geographic areas (PUMAs instead

of counties) and better isolate the difference in the results that is due to extrapolation.

Appendix C: Methods and Implementation

This appendix provides instructions to implement the SMC method and the conditional distribution method. Programs for both methods are available from my website. TRIM is described in Zedlewski and Giannarelli (2015).

SMC Method

I use the term "SMC method" to refer to the method of imputing additional program recipients and amounts received used by Scholz, Moffitt and Cowan (2009) adapted to the ACS survey data. Moffitt and Scholz (2010), Ben-Shalom, Moffitt and Scholz (2012) and Moffitt and Pauley (2017) use similar approaches to impute recipients based on models of reported receipt until survey totals match official statistics for the number of recipients and amounts received. Scholz, Moffitt and Cowan (2009) provide a detailed description of their implementation on page 218-219. Applying the method to the ACS data requires some modifications as described below.

Specifically, the SMC method as applied in this paper first uses the survey data to estimate a probit model of program receipt. I include the following covariates: income (9 indicators based on quantiles), education, marital status, number of children, race and ethnicity, age of the family reference person (indicators for 5 year intervals), age of children and participation in other programs. I then use the estimated probit coefficients to predict the probability of program receipt for each non-recipient household in the survey. I then assign transfer receipt to the households with the highest probability of receipt that do not report receipt in the survey until the number of recipient individuals matches administrative aggregates. In the analysis of NY, the administrative aggregates are calculated from the administrative microdata. In the analysis of the entire U.S., I use state-level aggregates

based on numbers provided by the BEA and USDA.

The modified SMC method imputes receipt probabilistically (rather than assigning it to the most likely recipients) until the number of recipients matches the administrative aggregates. That is, I draw a sample of households that do not report receipt with probabilities proportional to the predicted probabilities of receipt, such that the survey weights of these new recipients add up to the difference between the number of recipient individuals according to the official statistics and the survey reports. In order to match this number exactly, I sample individual households according to their predicted probabilities until the number of recipient individuals matches the official totals. A simple way to match this number in expectation is to draw a random sample with probabilities proportional to the predicted probabilities where the factor of proportionality is the ratio of survey to official totals.

The ACS does not contain benefit amounts, so I impute benefit amounts to both reported and imputed recipients based on the CPS. The ACS and the CPS are representative of the same populations and the covariates are comparable in the two surveys. I first regress amounts received per household member on basic demographic characteristics (education, marital status, race, Hispanic, number of children, participation in other programs) among those reporting receipt in the CPS. I use the regression coefficients to predict the expected amount received per household member in the ACS and then add a randomly drawn residual. To obtain the household amount, I multiply this prediction by the household size. Finally, I scale up amounts for all recipients to match total spending from official statistics.

Scholz, Moffitt and Cowan (2009) develop this method using the SIPP, but I apply it to the ACS. Key differences are that I impute amounts for all recipients (rather than just the imputed ones) and do so based on a different survey, the CPS. In the extrapolation to the entire U.S., I impute recipients and amounts until they match official aggregates for each state, while Scholz, Moffitt and Cowan (2009) adjust to national statistics.

Conditional Distribution Method

The first step in implementing the conditional distribution method is to estimate the conditional distribution $f_{X^A|X^R,Z}$. This distribution depends on X^A, which is only observed in the linked data. Thus, it needs to be estimated by a researcher with access to the confidential data. The linked data contain X^A, X^R and Z, so estimating the conditional distribution by maximum likelihood is a standard exercise. For example, the log likelihood function of the conditional distribution I use in this paper is simple to derive from equation (3):

$$\ell(X^A, X^R, Z; \alpha, \beta, \gamma, \delta, \sigma, \tau) = \sum_{i=1}^{N} \mathbb{1}[x_i^A = 0] \cdot \log\left(\Phi(x_i^R\alpha + z_i\beta; 0, 1)\right) + \left(1 - \mathbb{1}[x_i^A = 0]\right) \cdot$$
$$\left(\log\left(1 - \Phi(x_i^R\alpha + z_i\beta; 0, 1)\right) + \log\left(\phi(x_i^A; x_i^R\gamma + z_i\delta, \sigma)\right) - \log\left(1 - \Phi(\tau; x_i^R\gamma + z_i\delta, \sigma)\right)\right)$$

Maximizing this likelihood function with respect to the parameters $(\alpha, \beta, \gamma, \delta, \sigma, \tau)$ over the linked sample is simple using standard software packages. To correct for multiple mismeasured variables, one could replace the normal distributions in the equation above by multivariate normal distribution functions. As long as one is willing to specify a fully parametric model of the conditional distribution, this remains a standard estimation problem.

Specifying a fully parametric model introduces the risk of misspecification bias, so it is important to test model specification. As pointed out above, it is crucial to condition on variables that are likely to be included in the models to which the conditional distribution method is applied, unless they do not affect the conditional distribution. One can use standard significance tests to examine whether the variables in the model matter and whether their functional form is correct. The parametric model I use assumes the conditional density of non-zero amounts to be a truncated normal at every value of (X^R, Z). One could choose another parametric model, such as a truncated t-distribution or a Weibull distribution instead. Programs to estimate common models are available from my website. Standard specification tests, such as likelihood ratio tests can be used to choose between these models. The main purpose of estimating the conditional distribution is to reproduce the joint

distribution of the linked data. Researchers with access to the linked data can test whether simulating values of X^A from the estimated conditional distribution reproduces the actual distribution of X^A. A Kolmogorov-Smirnov test can show whether draws of X^A from the conditional distribution come from the same distribution as X^A in the linked data. This procedure only tests the marginal distribution of X^A and not its relation to other variables. The tests in Andrews (1997) and Rothe and Wied (2013) can test the specification of the entire conditional distribution.

If the tests indicate that the estimated conditional distribution does not reproduce the data well, one can relax the constraints the models impose. For example, one can allow the parameters of the distribution to depend on covariates in a more flexible way by including higher order terms in Z. One can also allow additional parameters such as σ to vary with covariates. If the tests reject even for flexible parametric models, one may be able to use a semi-parametric estimator, such as a sieve estimator (Chen, 2007).[2] Once a good specification has been found, the estimated parameters of $f_{X^A|X^R,Z}$ can be disclosed to the public. Researchers who want to use the conditional distribution method to correct their estimates can use these parameter estimates in step two and will usually not need to estimate a conditional distribution themselves.

An advantage of a parametric conditional distribution is that it is simple to incorporate additional information as restrictions even after the distribution has been estimated and disclosed. Thereby, one can relax the assumption that the conditional distribution does not change when extrapolating. For example, in the extrapolation to the U.S., I add state-specific intercepts to the conditional distribution to make total SNAP recipients and spending match

[2]Both the extension to the multivariate case and the extension to semi-parametric distributions are straightforward. Still, combining both extensions by semi-parametrically estimating a joint distribution is likely impractical for more than a few variables. An alternative is to estimate a conditional distribution for each mismeasured variable, consecutively adding the true values of the previously estimated distributions to the conditioning set. One can then simulate the mismeasured variables one variable at a time, always conditioning on the draws of the previously simulated variables. Both approaches may be feasible for a small number of variables, but are likely to suffer from slow convergence rates due to a curse of dimensionality. Both semi-parametric joint densities and chained conditional distributions require integration over the product of *estimated* conditional distributions, which amplifies estimation noise.

official statistics in each state. This addition avoids the assumption that misreporting levels are the same across states and only requires the slope parameters to be identical. The model is non-linear, so I calculate the intercepts using a Newton-Rhapson procedure that adjusts each intercept in β iteratively until the expected number of recipients match. The intercepts in δ are calculated the same way, but take the adjustment of β into account.

In the second step, researchers use the estimated conditional distribution to integrate X^A out of the objective function of their estimator. This step does not require access to the linked data and is simple to do by simulation. For every observation i in the public use data, the researcher first takes D draws from the estimated conditional distribution that conditions on the reports x_i^S, z_i of this observation. When the cumulative distribution function is invertible for every (x_i^S, z_i), this is simple to do by drawing D draws from a uniform distribution for every observation and applying the inverse of the cumulative distribution function to these draws. In case the cumulative distribution function is not invertible, one can use importance sampling to generate these draws. For the case of multiple mismeasured variables, one simply simulates D draws from the joint distribution of the mismeasured variables.[3]

The resulting data set includes $D \cdot N$ observations. As discussed in section III, solving the original estimation problem on this expanded data set amounts to integrating over X^A and thereby yields consistent estimates of θ. For example, to estimate receipt rates by income category, one computes the average receipt rates in the relevant income category from the expanded data set. To estimate a probit model, one maximizes the standard probit likelihood over the expanded data set. Note that this requires solving the estimation problem on the stacked draws, i.e. a data set of $D \cdot N$ observations, which is a key difference to multiple imputation. Multiple imputation solves the estimation problem D times for simulated data sets of size N each. The two approaches are equivalent for the example of the

[3]If a joint conditional distribution was estimated, this does not differ from the univariate case. If multiple "chained" univariate conditional distributions were estimated, start by taking D draws from the first conditional distribution for each observation. Then, for each of the $D \cdot N$ lines of the resulting data, take one draw from the second conditional distribution that conditions on the simulated value from the first distribution. Then, simulate $D \cdot N$ values from the third distribution, including the first two simulated variables in Z, which yields a sample from the joint distribution of X^A, X^S, Z.

(conditional) mean, but they differ for many other estimators including the probit example. Many estimators require *both* N and D to go to infinity, so estimates based on one draw and hence the multiple imputation estimator are inconsistent. The conditional distribution method remains consistent by simulating the integral, but solving the estimation problem on the stacked data also restores the consistency of multiple imputation estimators.

Solving the estimation problem on the expanded data set yields consistent parameter estimates, but the SEs need to be corrected for several reasons. The SEs should use N rather than $N \cdot D$ as the sample size (before any degree of freedom adjustment), which is simple to undo in standard software output. Even after this correction, SEs according to the standard formulas are "naïve" in that they do not take error due to simulation and estimation error in the parameters of the conditional distribution into account. If feasible, one should choose D to be large enough that simulation does not affect the estimates.[4] Unfortunately, little can be said about the required number of draws in general, as it depends on the curvature of the objective function. More non-linear objective functions tend to require larger numbers of draws. For a specific case, it is simple to determine whether D is sufficiently large by increasing D until the estimates stabilize.

One can correct SEs for the estimation error in the parameters of the conditional distribution using the GMM approach described in Newey and McFadden (1994). The key idea is that one can stack the moment conditions of the outcome model from equation (2) and the first order conditions of the likelihood function of the conditional distribution, so that they form a joint system of moment equations. Standard GMM formulas provide the asymptotic variance matrix of the entire coefficient vector. The asymptotic variance matrix of the second stage parameters that takes estimation of the first stage parameters into account is the block of this matrix that corresponds to the second stage coefficients. As shown in Newey and McFadden (1994), under the assumption that the moment conditions of the outcome model and those of the conditional distribution are not correlated,[5] the variance matrix of

[4] If this is not feasible, one can correct SEs for simulation error using the formulas in McFadden (1989).

[5] If the moment conditions are correlated, the correction requires access to the microdata. See Newey

the second step estimates, $\hat{V}_\theta$, can be estimated by

$$\hat{V}_\theta = \hat{V}_\theta^* + \hat{G}_\theta^{-1}\hat{G}_\Gamma\hat{V}_\Gamma\hat{G}_\Gamma'\hat{G}_\theta^{-1\prime}$$

Where $\hat{V}_\theta^*$ is the "naïve" variance estimate that ignores the estimation of the parameters of the conditional distribution. $\hat{G}_\theta$ is the sample Jacobian of the second step, i.e. the average of the partial derivatives of $\tilde{m}$ with respect to the second stage parameters θ. Analogously, $\hat{G}_\Gamma$ is the sample average of partial derivatives of $\tilde{m}$ with respect to the first stage parameters, $\Gamma = (\alpha, \beta, \gamma, \delta, \sigma, \tau)$. $\hat{V}_\Gamma$ is the estimated variance of the first stage parameters.

This formula provides some intuition on the likely effects of the correction. First, the correction vanishes if and only if $\hat{G}_\Gamma = 0$, i.e if at the true parameter values, the limit of the objective function of the second stage estimator does not depend on the estimated first stage parameters. This assumption is strong, as it requires the second stage estimates to be consistent even when the parameters estimated in the first step are not consistent (Newey and McFadden, 1994). Second, the correction always increases the variance, which may not be the case when the moment conditions are correlated; see Newey and McFadden (1994) for a discussion. Third, in addition to the usual determinants of the variance, the magnitude of the adjustment increases in the variance of the estimated first stage parameters ($\hat{V}_\Gamma$, which is small in this application due to the large linked sample) and the sensitivity of the objective function of the second stage to the estimated first stage parameters, $\hat{G}_\Gamma$.

Implementing this correction requires the variance matrix of the first stage parameters, $\hat{V}_\Gamma$, to be disclosed along with the parameters. $\hat{V}_\theta^*$ and $\hat{G}_\theta$ can be obtained from the second stage estimates. They are part of the standard output of most estimation routines. $\hat{G}_\Gamma$ is model specific and needs to be calculated manually from the public use data and the simulated values. One can do so analytically by deriving the partial derivatives of $\tilde{m}$ with respect to the first stage parameters and computing their sample average. If one can make simulation error negligible, one can also take numerical derivatives. This correction makes

and McFadden (1994) for discussion.

estimation more complicated, but most other corrections do not take the uncertainty from imputation into account, because their asymptotic properties are unknown.

Appendix D: Abbreviations and Definitions

Adjusted conditional distribution method: Corrects for misreporting as the conditional distribution method does, but adjusts the parameters of the estimated conditional distribution to make survey totals match known control totals.

ACS: American Community Survey

ACS PUMS: American Community Survey Public Use Microdata Sample

BEA: Bureau of Economic Analysis

Conditional distribution method: The conditional distribution method corrects for misreporting by integrating the true variable out of the objective function of the estimator. To do so, it uses an estimate of the conditional distribution of the true variable given survey reports and other covariates obtained from validation data.

County groups: The smallest geographic area (in terms of population count) that can be identified in both the ACS PUMS data and from county-level official statistics. To find county groups, I sort counties into groups of counties that minimize the population size of each county group subject to the constraint that the boundaries of the county groups do not intersect with PUMA boundaries.

CPS: Current Population Survey

East NY: East NY combines the eastern counties of the NYC metro area and the eastern counties of upstate NY. See Appendix Figure A1 for a map.

Group Quarters: Group quarters are grouped living arrangements owned or managed by an entity or organization providing housing and/or services for the residents. Group quarters are excluded from the analysis throughout.

Modified SMC method: The modified SMC method imputes transfer receipt and amounts

using models based on survey reports as the original SMC method (see below). Contrary to the original SMC method, it imputes receipt probabilistically.

Poverty rate: Poverty rates are calculated using poverty status according to reported annual household income. I use the unadjusted federal poverty cutoffs for households throughout (also when including SNAP in the income definition).

Poverty reduction: The poverty reduction is calculated as the difference in the poverty rate when using pre-tax cash income and when adding SNAP, i.e. it only has a causal interpretation if one assumes that there are no behavioral effects.

Public Use Microdata Areas (PUMAs): The smallest geographic area that can be identified in the ACS public use data. The U.S. Census Bureau defines PUMAs as statistical geographic areas containing at least 100,000 people.

OTDA: New York Office of Temporary and Disability Assistance

SIPP: Survey of Income and Program Participation

SMC method: I use the term "SMC method" to refer to the method of imputing additional program recipients and amounts received used by Scholz, Moffitt and Cowan (2009) adapted to the ACS survey data. The key idea of this correction for underreporting is to impute additional transfer receipt based on models estimated from the survey reports.

SNAP: Supplemental Nutrition Assistance Program, formerly the Food Stamp Program

TRIM: TRIM is the Transfer Income Model, version 3, developed by the Urban Institute. TRIM is a microsimulation model that simulates the major governmental tax, transfer, and health programs that affect the U.S. population.

USDA: U.S. Department of Agriculture

West NY: West NY combines the western counties of the NYC metro area and the western counties of upstate NY. See Appendix Figure A1 for a map.

References

Andrews, Donald W. K. 1997. "A Conditional Kolmogorov Test." *Econometrica*, 65(5): 1097–1128.

Ben-Shalom, Yonatan, Robert A. Moffitt, and John K. Scholz. 2012. "An Assessment of the Effectiveness of Anti-Poverty Programs in the United States." In *Oxford Handbook of the Economics of Poverty.* , ed. Philip Jefferson, Chapter 22, 709–749. Oxford:Oxford University Press.

Chen, Xiaohong. 2007. "Large Sample Sieve Estimation of Semi-Nonparametric Models." In *Handbook of Econometrics.* Vol. 6b, , ed. James J. Heckman and Edward Leamer, Chapter 76, 5549–5632. Amsterdam:Elsevier.

McFadden, Daniel. 1989. "A method of simulated moments for estimation of discrete response models without numerical integration." *Econometrica*, 995–1026.

Meyer, Bruce D., and Nikolas Mittag. 2018. "An Empirical Total Survey Error Decomposition Using Data Combination." Unpublished Manuscript.

Meyer, Bruce D., and Nikolas Mittag. forthcoming. "Using Linked Survey and Administrative Data to Better Measure Income: Implications for Poverty, Program Effectiveness and Holes in the Safety Net." *American Economic Journal: Applied Economics.*

Meyer, Bruce D., Nikolas Mittag, and Robert M. Goerge. 2018. "Errors in Survey Reporting and Imputation and Their Effects on Estimates of Food Stamp Program Participation." Unpublished Manuscript.

Meyer, Bruce D., Wallace K. C. Mok, and James X. Sullivan. 2015. "Household Surveys in Crisis." *Journal of Economic Perspectives*, 29(4): 199–226.

Moffitt, Robert A., and Gwyn Pauley. 2017. "Trends in the Distribution of Social Safety Net Support After the Great Recession." The Stanford Center on Poverty and Inequality Policy Brief, Stanford, CA.

Moffitt, Robert A., and John K. Scholz. 2010. "Trends in the Level and Distribution of Income Support." *Tax Policy and the Economy*, 24: 111–152.

Newey, Whitney K., and Daniel L. McFadden. 1994. "Large Sample Estimation and Hypothesis Testing." In *Handbook of Econometrics.* Vol. 4, , ed. Robert F. Engle and Daniel L. McFadden, Chapter 36, 2111–2245. Amsterdam:Elsevier.

Rothe, Christoph, and Dominik Wied. 2013. "Misspecification Testing in a Class of Conditional Distributional Models." *Journal of the American Statistical Association*, 108(501): 314–324.

Scholz, John K., Robert A. Moffitt, and Benjamin Cowan. 2009. "Trends in income support." In *Changing Poverty, Changing Policies.* , ed. Maria Cancian and Sheldon Danziger, Chapter 8, 203–241. New York, NY:Russell Sage Foundation.

U.S. Census Bureau. 2008. "A Compass for Understanding and Using American Community Survey Data: What General Data Users Need to Know." U.S. Government Printing Office, Washington, D.C.

U.S Department of Labor, Bureau of Labor Statistics. 2002. "Current Population Survey: Design and methodology." Bureau of Labor Statistics Technical Paper 63RV, Washington, D.C.

Wagner, Deborah, and Mary Layne. 2014. "The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications'(CARRA) Record Linkage Software." U.S. Census Bureau Center for Administrative Records Research and Applications Working Paper.

Wooldridge, Jeffrey M. 2007. "Inverse probability weighted estimation for general missing data problems." *Journal of Econometrics*, 141(2): 1281–1301.

Zedlewski, Sheila, and Linda Giannarelli. 2015. "TRIM: A Tool for Social Policy Analysis." The Urban Institute Research Report.