Do Imputed Earnings Earn Their Keep? Evaluating SIPP Earnings and Nonresponse with Administrative Records

Rebecca L. Chenevert
Mark A. Klee
Kelly R. Wilkin

Social, Economic, and Housing Statistics Division
U.S. Census Bureau

ASSA Annual Meetings
San Francisco, CA
1/3/2016

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

Disclaimer

This paper is released to inform interested parties of ongoing research and to encourage discussion of work in progress. The views expressed in this paper are those of the authors and not necessarily those of the U.S. Census Bureau. Any errors are our own.

How Do Survey & Admin Earnings Compare?

- Historically, survey data has been the main source of information about earnings.
- Previous work has evaluated the quality of survey earnings data by comparison to an alternative data source.
- “Measurement error” appears to be:
  - correlated negatively with earnings — Bound & Krueger (1991), Bollinger (1998), Gottschalk & Huynh (2010), and
  - correlated with earnings determinants — Pedace & Bates (2000), Cristia & Schwabish (2009).
- Abowd & Stinson (2013) suggest that reported survey earnings are similar to administrative earnings in reliability, but...
- imputed survey earnings are less reliable than administrative earnings.

Census Bureau Imputation

- Reported earnings might be missing because:
  - the person was not interviewed or declined to provide any labor force data ("unit nonresponse"), or
  - the person declined to answer earnings questions when asked ("item nonresponse").
- Imputation fills in by copying earnings reported by a "donor" with similar observables.
- Panel surveys can use previous wave data for imputation

To Keep or Not To Keep?

- Hirsch & Schumacher (2004) and Bollinger & Hirsch (2006) show that imputation introduces "match bias" if not all earnings determinants are incorporated.
- Analysts commonly exclude imputed earnings from regressions to mitigate match bias.
- This strategy assumes that earnings nonresponse is unrelated to true earnings ("ignorable").
- Although Bollinger & Hirsch (2013) argue that omitting imputed earners avoids major bias in estimated slope coefficients,...
- Bollinger et al. (2015a,b) show that earnings nonresponse is more likely in the tails of the administrative earnings distribution.

Our Contribution

- Compare labor earnings in the Survey of Income and Program Participation (SIPP) and in the Social Security Administration's Detailed Earnings Record (DER).
- Point to whose earnings would be affected most by replacing survey data with administrative data.
- Investigate the predictors of earnings nonresponse, aiming to understand:
  - who is likely to have noisier earnings data, and
  - whether earnings response bias is ignorable.
- Synthesize these findings by examining implications for estimates of the earnings structure.

Preview of results

- Reported earnings match administrative more closely than imputed earnings.
- There is heterogeneity in how well different methods of imputing survey data match administrative data on average.
- Individuals who are male, more educated, self-employed, and non-recipients of programs have larger average deviations of survey and admin earnings.
- Individuals who are male, self-employed, and non-recipients of programs are more likely to lack earnings data.
- We document a novel pattern of nonresponse along the adminsitrative earnings distribution.
- Ambiguous implications for estimates of earnings regression coefficients.

Road Map

Introduction

Data

Benchmarking

- Graphical Analysis
- Regression Analysis

Predictors of Survey Earnings Nonresponse

Implications for Estimates of the Earnings Structure

Conclusion

Survey of Income and Program Participation

- SIPP is a large, nationally representative, panel dataset.
- 2008 SIPP panel begins in May 2008.
- We use all waves through March 2013.
- At every interview (“wave”), respondents answer a core set of questions about the previous four months.
- Each SIPP wave offers monthly earnings from up to two jobs, up to two businesses, moonlighting, and severance pay.
- For person-year analysis, we aggregate nonresponse and earnings to the annual level.

Detailed Earnings Record

- Specialized extract of uncapped earnings from SSA's Master Earnings File.
- Record of all wages, tips, and other earnings reported on each W-2 received by workers for an employer.
- Record of all taxable income reported on Form 1040, Schedule SE for self-employed workers.
- We aggregate non-deferred earnings, deferred earnings, and self-employment earnings to the person-year level.
- We use 2009 through 2012.

▸ Linking SIPP to DER

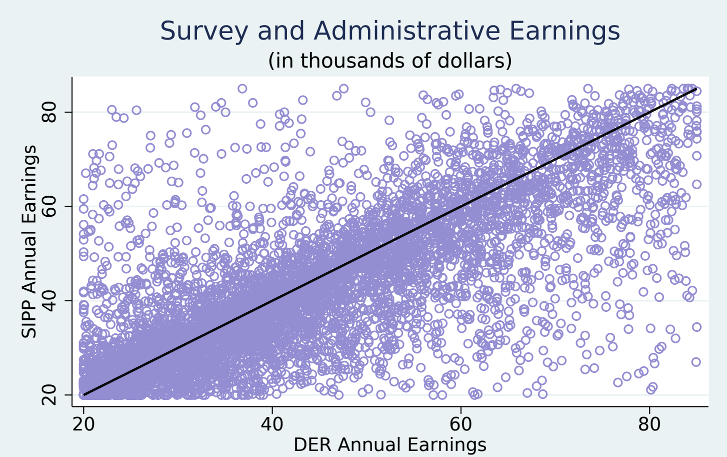
The Average Deviation of SIPP and DER Earnings

	(1)	(2) Absolute	(3)
	DER-SIPP	Difference	Obs
Including Zero Earners	$969	$6,277	158,168
Excluding Zero Earners	$1,928	$10,429	89,418

▸ Presence of Earnings

▸ Measurement Error?

DER vs SIPP: Earnings

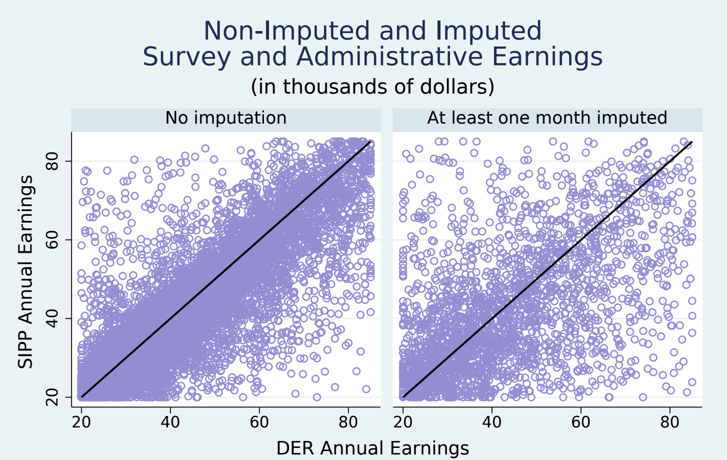


Source: Authors' calculation from the 2008 panel of the Survey of Income and Program Participation, Waves 1 through 14 and the Social Security Administration's Detailed Earnings Record, calendar years 2009 through 2012.

▸ Histogram

DER vs SIPP: Earnings

By Imputation Status



Source: Authors' calculation from the 2008 panel of the Survey of Income and Program Participation, Waves 1 through 14 and the Social Security Administration's Detailed Earnings Record, calendar years 2009 through 2012.

DER-SIPP Regressions

Any Nonresponse, Including Nonemployed

	(1)	(2)
VARIABLES	DER-SIPP	Absolute Difference
Any nonresponse	-1,919.228***	6,898.190***
	(150.799)	(115.027)
Female	-681.496***	-3,068.404***
	(98.211)	(80.129)
Black, non-Hispanic	633.243**	512.375**
	(266.341)	(221.789)
Asian, non-Hispanic	1,278.955***	791.004**
	(423.778)	(341.302)
White, non-Hispanic	367.802	626.719***
	(234.691)	(196.445)
Hispanic	1,344.658***	347.540
	(293.877)	(241.735)
Married, spouse absent	324.423	-83.473
	(486.621)	(413.664)
Never married	-405.975**	-1,326.117***
	(159.209)	(130.404)
Previously married	-104.780	-502.648***
	(121.522)	(100.496)
Any transfer income	-730.051***	-3,050.304***
	(92.119)	(94.792)
	(CONTINUED...)	

	(1)	(2)
VARIABLES	DER-SIPP	Absolute Difference
Elementary school	-275.560	-114.830
	(173.785)	(151.381)
Some high school	23.273	-483.810***
	(118.404)	(102.572)
Some college	327.244**	651.344***
	(138.261)	(113.620)
Associate's degree	291.111**	532.843***
	(130.985)	(106.704)
Bachelor's degree	975.960***	2,909.959***
	(172.324)	(139.714)
Master's degree	1,239.306***	4,055.376***
	(272.902)	(218.525)
Professional degree	2,851.460***	10,487.695***
	(869.993)	(693.289)
Doctorate degree	2,214.646***	6,993.012***
	(781.239)	(630.579)
Foreign-born, citizen	740.761***	431.731**
	(252.788)	(202.519)
Non-English speaker	-463.043**	343.148**
	(218.351)	(173.367)
Observations	158,168	158,168
R-squared	0.010	0.182

DER-SIPP Regressions

Detailed Nonresponse, Including Nonemployed

	(1)	(2)
VARIABLES	DER-SIPP	Absolute Difference
Any hot-deck imputation	-6,625.835***	3,681.865***
	(317.637)	(243.895)
Any Type-Z imputation	2,239.581***	162.567
	(193.760)	(170.977)
Any longitudinal labor force imputation	-3,051.713***	2,686.463***
	(448.855)	(343.988)
Any imputation based on last month — Reported	408.605	2,269.766***
	(309.375)	(232.990)
Any imputation based on last month — Imputed	3,515.126***	7,322.420***
	(431.305)	(313.817)
Any imputation based on last month — Logical	1,264.158	5,519.321***
	(828.437)	(624.038)
Any logical imputation	861.216***	387.869**
	(210.012)	(163.952)
Any proxy response	663.439***	192.747**
	(105.395)	(87.480)
Observations	158,168	158,168
R-squared	0.019	0.177

▸ Types of Imputation

DER-SIPP Regressions

Detailed Nonresponse, Positive Earners Only

VARIABLES	DER-SIPP	Absolute Difference
Any hot-deck imputation	-2,373.737***	-1,860.838***
	(454.256)	(335.743)
Any Type-Z imputation	3,218.760***	1,312.942***
	(655.354)	(504.035)
Any longitudinal labor force imputation	-2,671.681***	-493.946
	(776.459)	(589.965)
Any imputation based on last month — Reported	-603.211*	2,826.048***
	(361.068)	(272.946)
Any imputation based on last month — Imputed	1,844.083***	11,223.464***
	(569.619)	(410.897)
Any imputation based on last month — Logical	2,918.492***	6,352.998***
	(972.768)	(750.235)
Any logical imputation	267.916	247.797
	(239.249)	(185.570)
Self-employed	-11,713.350***	11,776.952***
	(503.736)	(375.654)
Any transfer income	-665.609	-900.633**
	(450.033)	(360.909)
Observations	82,936	82,936
R-squared	0.057	0.184

The Predictors of Survey Earnings Nonresponse

VARIABLES	(1) Any Non- response	(2) Unit Non- response	(3) Item Non- response
Number of household members	0.021***	0.026***	-0.005***
	(0.001)	(0.001)	(0.001)
Age	-0.023***	-0.021***	0.008**
	(0.001)	(0.001)	(0.003)
Age squared	0.001***	0.001***	-0.000***
	(0.000)	(0.000)	(0.000)
Female	-0.028***	-0.012***	-0.004**
	(0.001)	(0.001)	(0.001)
Elementary school	-0.012***	-0.008***	-0.012**
	(0.002)	(0.001)	(0.005)
Some high school	-0.018***	-0.009***	-0.016***
	(0.002)	(0.001)	(0.003)
Some college	-0.008***	-0.008***	-0.010***
	(0.001)	(0.001)	(0.002)
Associate's degree	-0.002*	-0.004***	-0.008***
	(0.001)	(0.001)	(0.002)
Bachelor's degree	0.000	-0.011***	-0.002
	(0.001)	(0.001)	(0.002)
Master's degree	-0.003	-0.013***	-0.002
	(0.002)	(0.001)	(0.003)

(CONTINUED...)

	(1)	(2)	(3)
	Any Non-	Unit Non-	Item Non-
VARIABLES	response	response	response
Professional degree	0.031***	-0.008***	-0.010*
	(0.005)	(0.002)	(0.006)
Doctorate degree	-0.005	-0.013***	-0.030***
	(0.005)	(0.002)	(0.006)
Self-employed	—	—	0.197***
			(0.003)
Foreign-born, citizen	0.015***	0.008***	0.008***
	(0.002)	(0.001)	(0.003)
Foreign-born, non-citizen	0.008***	0.009***	-0.001
	(0.003)	(0.002)	(0.003)
Proxy response	-0.050***	-0.086***	0.054***
	(0.001)	(0.001)	(0.001)
Any sample gaps	0.035***	0.019***	0.034***
	(0.001)	(0.001)	(0.002)
Attritor	0.039***	0.030***	0.023***
	(0.001)	(0.001)	(0.002)
Any children under 18	-0.042***	-0.035***	-0.021***
	(0.001)	(0.001)	(0.002)
Change in family composition	-0.023***	-0.017***	-0.014***
	(0.002)	(0.001)	(0.003)
Any transfer income	-0.027***	-0.006***	-0.008
	(0.002)	(0.001)	(0.006)

(*CONTINUED...*)

VARIABLES	(1) Any Non-response	(2) Unit Non-response	(3) Item Non-response
Any admin records	0.053***	0.010***	-0.078***
	(0.002)	(0.001)	(0.004)
Number of admin records	0.031***	0.002***	0.024***
	(0.001)	(0.001)	(0.001)
Bottom admin earnings quintile	-0.006***	-0.004***	0.045***
	(0.002)	(0.001)	(0.003)
Second admin earnings quintile	0.015***	-0.005***	0.018***
	(0.002)	(0.001)	(0.002)
Fourth admin earnings quintile	-0.010***	-0.002	-0.003
	(0.002)	(0.001)	(0.002)
Top admin earnings quintile	-0.002	-0.002	0.004**
	(0.002)	(0.001)	(0.002)
Observations	1,910,102	1,910,102	1,055,629
R-squared	0.054	0.076	0.080

▸ Person-Job-Month Regressions

Implications for Estimates of the Earnings Structure

Mincer Regression

VARIABLES	(1) SIPP	(2) DER	(3) Reported SIPP	(4) SIPP-DER Hybrid
Years of education	0.138***	0.137***	0.145***	0.140***
	(0.002)	(0.002)	(0.002)	(0.002)
Potential experience	0.091***	0.099***	0.093***	0.097***
	(0.001)	(0.001)	(0.001)	(0.001)
Potential experience^2	-0.002***	-0.002***	-0.002***	-0.002***
	(0.000)	(0.000)	(0.000)	(0.000)
Observations	88,971	88,971	60,994	88,971
R-squared	0.259	0.239	0.275	0.255

Future Work: 2014 Panel

- SIPP has been redesigned for the forthcoming 2014 panel.
- Data will be released after processing is complete.
- How are earnings nonresponse and the deviation of survey and admin earnings affected by the combination of:
    - changing interview frequency from every four months to annual,
    - relying less on prior months' earnings in the data collection and editing processes,
    - offering ranges when individuals initially decline to provide wage/salary data, and
    - more transparancy with users regarding reported earnings net of taxes.

Conclusion

- Imputed survey earnings differ from administrative earnings by more than reported survey earnings do on average.
- There is heterogeneity in how well different methods of imputing survey data match administrative data on average.
- Individuals who are male, more educated, self-employed, and non-recipients of programs have larger average deviations of survey and admin earnings.
- Individuals who are male, self-employed, and non-recipients of programs are more likely to lack earnings data.
- We document a novel pattern of nonresponse along the adminsitrative earnings distribution.
- Ambiguous implications for estimates of earnings regressions.

Thank You!

Name:

Becky Chenevert

Affiliation:

Social, Economic, and Housing Statistics Division

U.S. Census Bureau

Email:

rebecca.l.chenevert@census.gov

Phone:

(301) 763-8538

Linking SIPP to DER

- SIPP respondents are assigned a Protected Identification Key (PIK) based on self-reported SSN, name, sex, race, and age.
- DER records are also assigned a PIK, then linked to SIPP data.
- Bond et al. (2013) show that PIKs are less likely to be assigned to individuals who are:
  - mobile,
  - less educated,
  - worse English speakers,
  - non-employed,
  - minorities, and
  - non-participants in programs.
- We restrict our sample to person-years aged 15 and over,...
- who are assigned a PIK, and...
- who are present in the survey for all 12 months.

▸ Back

Interpretation

- Previous literature has often framed DER-SIPP comparisons as evaluations of measurement error in survey earnings.
- Abowd & Stinson (2013) and Bollinger et al. (2015) take a more agnostic approach, due to:
  - conceptual differences in earnings across datasets (*e.g.* health insurance, off-the-books earnings, some non-wage benefits),
  - reporting errors in administrative data as well as survey data, and
  - the possible misassignment of PIKs.
- We pursue the latter interpretation, characterizing the difference between SIPP and DER earnings.

▸ Back

DER vs SIPP: Presence of Earnings

	No DER earnings	Positive DER earnings	Total
No SIPP earnings	56,410	8,216	64,626
Positive SIPP earnings	4,745	90,394	95,139
Total	61,155	98,610	159,765

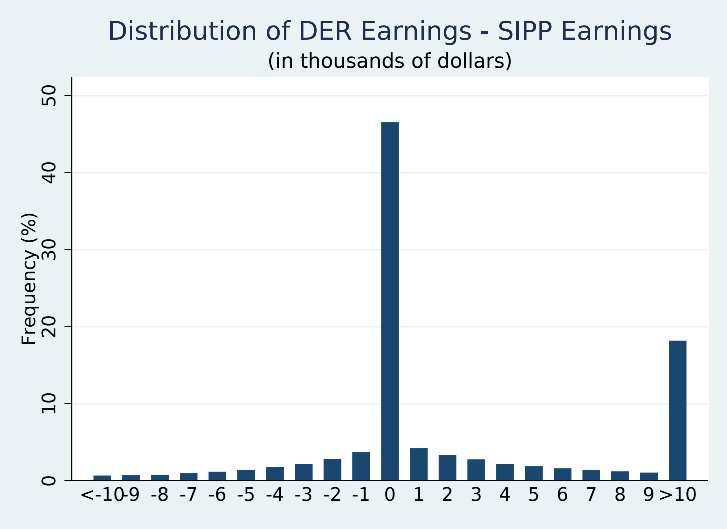
▸ Back

Different Types of Imputation

- Unit nonresponse:
    - If no known employment at the start of the reference period based on last wave's data, impute full record from a donor.
    - If known employment at the start of the reference period, project last wave's labor force data through current wave.
- Item nonresponse:
    - If no known earnings from last month, impute only earnings from a donor based on observable characteristics.
    - If known earnings from a job/business last month, impute only earnings from a donor based on last month's earnings.
    - If strong reason to believe earnings were reported incorrectly, impute logically using hourly pay rate and hours worked, reported annual pay rate, or weeks away without pay.

▸ Back

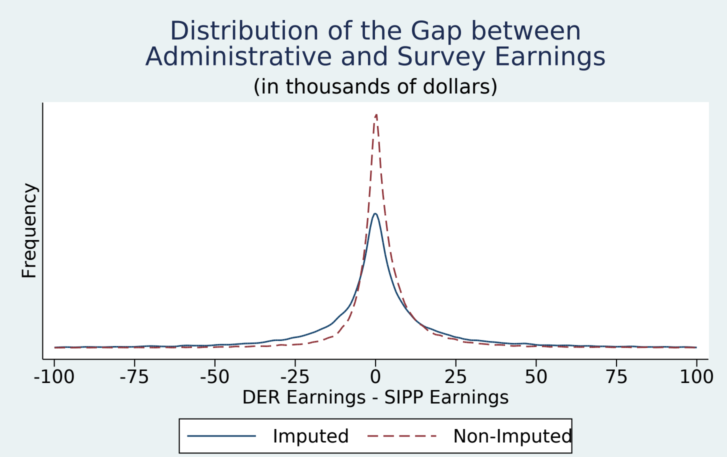
DER vs SIPP: Earnings



▸ Back

DER vs SIPP: Earnings

By Imputation Status

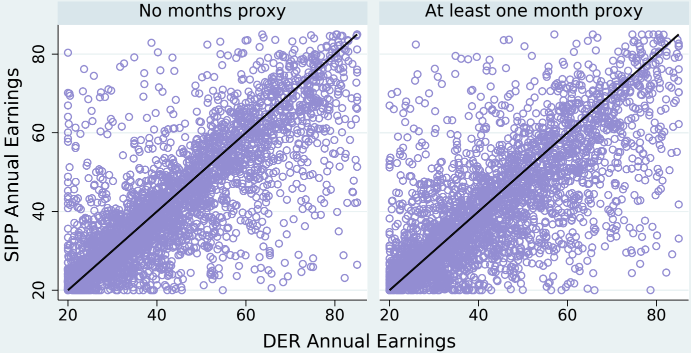


Source: Authors' calculation from the 2008 panel of the Survey of Income and Program Participation, Waves 1 through 14 and the Social Security Administration's Detailed Earnings Record, calendar years 2009 through 2012.

DER vs SIPP: Earnings

By Proxy Response Status

Non-Proxy and Proxy
Survey and Administrative Earnings
(in thousands of dollars)



Source: Authors' calculation from the 2008 panel of the Survey of Income and Program Participation, Waves 1 through 14 and the Social Security Administration's Detailed Earnings Record, calendar years 2009 through 2012.

The Predictors of Survey Earnings Item Nonresponse

	(1)	(2)	(3)
VARIABLES	Jobs	Jobs	Businesses
Weeks worked	-0.003***	—	0.013***
	(0.001)		(0.004)
Hours worked	-0.001***	-0.001***	0.001***
	(0.000)	(0.000)	(0.000)
Stopped work	0.045***	—	0.003
	(0.002)		(0.018)
Paid hourly	-0.019***	-0.018***	—
	(0.001)	(0.001)	
Contingent worker	—	0.058***	0.205***
		(0.007)	(0.058)
Salaried	—	—	-0.025***
			(0.004)
Other income	—	—	0.218***
			(0.014)
Observations	1,703,276	1,756,622	201,191
R-squared	0.027	0.026	0.039

Back