An Empirical Evaluation of Various Direct, Synthetic, and Traditional Composite Small Area Estimators

Kimberly Henry, Michael Strudler, William Chen
Internal Revenue Service, P.O. Box 2608, Washington DC 20013-2608

Abstract

Currently, the Statistics of Income (SOI) Division of the Internal Revenue Service uses the Individual Returns Transaction File, administrative data for the population of Form 1040 tax returns, to produce totals of various tax return variables at the state level. Previous research using SOI's Form 1040 sample, a large national sample of cleaned administrative tax records, suggests that the IRS data are subject to various kinds of errors that do not affect tax liability. For this reason, alternative approaches to state-level estimation using the edited sample file are desirable. This paper compares several variables' different direct, synthetic, and traditional composite estimates of state-level totals and compares the alternatives to the IRS-based totals.

Key words: Administrative Records, General Regression Estimator, Indirect Estimators, Survey sampling

1. Introduction: Small Area Estimation Using Administrative IRS Data and Its Associated Nonsampling Errors

The approximately 133 million tax records on the Internal Revenue Service's (IRS) Individual Returns Transaction File have several uses to multiple government agencies. In particular, these data serve as the sampling frame for the Statistics of Income (SOI) Division of IRS, as well as a source of population data for other tabulations. For example, SOI publishes tabulated monetary amounts and the associated number of returns by state and Adjusted Gross Income (AGI) categories using these data (Table 2 in each Spring issue of the *SOI Bulletin*).

These population data, based on administrative tax records for the U.S. tax filing population, are not error-free. While estimates from these data are free from sampling error, the data contain various nonsampling errors, as discovered in prior SOI research comparing return records in the transaction file to records for the same returns in SOI's augmented and edited Form 1040 sample. Only items necessary for computer processing of a tax return are retained on the transaction file, as opposed to items that might be needed for other purposes, such as producing statistical estimates. Measurement errors exist between the IRS and SOI data values due to different data editing rules. For revenue processing purposes, IRS does not spend scarce resources correcting errors that do not affect tax liability in the approximately 130 million tax return records it processes each year. Since tax liability is correct, this approach does no harm to IRS's tax collection mission or to taxpayers, but it can adversely affect the usability of the data for statistical purposes. SOI's transcription and editing staff receive extensive training, and the sample of approximately 230,000 returns is augmented with additional items from the return, and more closely monitored and checked for data consistency. Errors occur particularly for variables that are indirectly related to tax liability, such as State and Local Income Taxes deducted on Schedule A. They were also discovered for variables such as Taxable Interest and Business Income/Loss from Sole Proprietors (as reported on Schedule C) in the Tax Year 2003 IRS data. To correct these errors, SOI had to delay its publication of Table 2 for several months. Other limitations in the IRS data include a smaller amount of information being available, compared to SOI's sample, and data are often provided to SOI in tabular form, with monetary amounts rounded to thousands and certain high income taxpayers are omitted.

2. Data Description

2.1: The SOI Sample

SOI draws annual samples of the Form 1040 tax returns to produce richer and cleaner data for population estimation and tax modeling purposes. Stratification for the finite population of tax returns for SOI's Tax Year 2004 (i.e., income earned in 2004 and reported in 2005) Individual sample used the following categories:

1. Nontaxable returns with adjusted gross income or expanded income of $200,000 or more.

2. High combined business and farm total receipts of $50,000,000 or more.

3. Presence/absence of special forms or schedules (Form 2555, Form 1116, Form 1040 Schedule C, and Form 1040 Schedule F).

Stratum assignment priority was based on the order in which a return met one of these categories. For example, if a return met (1) and (2), it fell into strata based on (1). Within category (3), further stratification used size of total gross positive or negative income and an indicator of the return's "usefulness" for tax policy modeling purposes (Scali and Testa 2006). The positive/negative income values in strata boundaries were indexed for inflation between 1991 and the current tax year (Hostetter *et al.* 1990). This resulted in 216 strata. While the sample was designed for tax modeling and produces reliable national-level estimates, it is not large enough to produce state-level estimates.

Each tax return in the target population was assigned to a stratum based on these criteria, then subjected to sampling in a two-step procedure. Within each stratum, a .05 percent stratified simple random sample, called the Continuous Work History Sample (CWHS), was selected (Weber 2004). For returns not selected for this sample, a Bernoulli sample was independently selected from each stratum, with sampling rates from 0.05 to 100 percent.

SOI's data capture and cleaning procedures resulted in a sample of 200,778 (including 65,948 CWHS returns) returns from an estimated population of 133,189,982 returns. We placed the 34,484 tax returns that SOI sampled with certainty into one certainty stratum, since they represented a census of tax returns. Thus, without loss of generality, we exclude this stratum from the population and develop our estimation method to estimate totals from all other strata. In this way, all errors in the certainty units are isolated and accounted for; only the portion of the total produced from the non-certainty units needs to be estimated. To estimate the entire population total, we simply add the total from the certainty strata to our estimate for the remaining population.

2.2: Small Areas and Variables of Interest

The reduced dataset for this analysis was created by first separating SOI's Tax Year 2004 sample into the certainty and non-certainty units. For both, the weighted sample data were tabulated to the state by Adjusted Gross Income (AGI) category level, where "state" included the 50 U.S. states, Washington DC, and an "other" category that included returns filed by civilians and military individuals living abroad, such as U.S. possessions and territories, Puerto Rico, etc. We also considered eight categories of AGI: Negative; $0 under $20,000; $20,000 under $30,000; $30,000 under $50,000; $50,000 under $75,000; $75,000 under $100,000; $100,000 under $200,000; and $200,000 and higher. These 52 states combined with the AGI categories resulted in 416 small areas. We consider estimates for the 52 states in this paper, utilizing the fact that there are differences in our variables of interest at the AGI category-level data.

The IRS data, prior to cleaning by SOI staff, were also compiled to this level. The ten variables we selected for this study can be grouped into two categories: variables that are more or less susceptible to errors in the IRS data. They are as listed, with their location on the Form 1040 tax form and a brief description, in Table 1 below.

Table 1: Variable Names, Tax Form Location, and Description, by Variable of Interest

Susceptible to Error	*Variable*	*Location on 2004 Tax Form*	*Description* [a]
Less	Adjusted Gross Income	Line 36	Income reported from the calculation of total income (Line 22) (pp. 117-118).
	Salaries and Wages	Line 7	Amount of reported compensation primarily for personal services; includes salaries, wages, tips, bonuses, etc. (p. 138).
	Total Tax Liability	Line 62	Sum of tax-related line items on 1040 (p. 146).
	Earned Income Tax Credit	Line 65a	Taxpayer credit for lower-income working individuals (pp. 123-124).
More	Net Schedule C Business Profit/Loss	Line 12	Total of profits and losses from a taxpayer's business, reported on Sch C (p. 120).
	Net Schedule D Capital Gains/Loss	Line 13	Total of capital gains/loss, as reported on Sch D (p. 120).
	Total Contributions	Lines 15-16, Sch A	Total of cash and noncash charitable contributions itemized deductions (p. 122).
	Total Taxes Paid Deduction	Lines 5-9, Sch A	Total of State and Local Taxes, Real Estate Taxes, Personal Property Taxes, and Other Taxes (p. 144).
	Interest Paid Deduction	Line 14, Sch A	Total of Home Mortgage Interest and investment interest deductions, from lines 10-13 on Sch A (p. 130).
	Total Itemized Deductions	Line 39	Total of all itemized deductions reported on Sch A (pp. 144-145).

a: page numbers from IRS 2005.

Since SOI's sample does not use state in the stratification, the number of sample returns by state varies considerably. Six of the states we considered large enough, i.e., more than 5,000 noncertainty returns within each one, such that the associated direct estimates are reasonable. The remaining states were collapsed into groups based on whether or not the state had state income taxes, geographic region, and whether the state had a relatively large or small size of income. This resulted in 21 groups. They are listed, with the associated number of certainty and noncertainty sample units, in Table 2:

Table 2: States and Number of Certainty (c) and Noncertainty Sample Units (nc), by Collapsed Group

States Within Group	*c*	*nc*
Alaska, Washington	811	4,024
Arkansas, Alabama, Mississippi, Louisiana	620	5,927
Arizona, New Mexico, Utah, Colorado	1,432	7,415
California	6,539	23,990
Connecticut, Rhode Island, Massachusetts	2,211	7,952
Washington DC, Maryland, Delaware	777	4,180
Florida, Tennessee	4,052	14,566
Georgia, North Carolina, South Carolina	1,265	10,108
Hawaii, Other	790	1,815
Iowa, Nebraska, Kansas, Missouri, Oklahoma	997	8,061
Illinois	1,539	7,451
Indiana, Ohio, Kentucky	1,135	9,908
Maine, Vermont, New Hampshire	215	1,770
Michigan, Wisconsin, Minnesota	1,447	10,379
Montana, North Dakota, Idaho, Oregon	435	3,364
New Jersey	1,273	6,138
Nevada, Wyoming, South Dakota	934	2,450
New York	4,527	13,101
Pennsylvania	931	6,480
Texas	2,318	11,427
Virginia, West Virginia	731	4,798

3. Direct Estimators

Let y_k be the value of the characteristic of interest for the kth tax return, $k \in U$, the finite population of tax returns. We are interested in estimating the finite population total:

$$Y = \sum_{k \in U} y_k .$$

Let s denote the sample of tax returns drawn from the population of tax returns using the stratified Bernouli sampling design. Let $s_d \subset s$ denote the part of the sample that belongs to the domain d of interest. Let w_k denote the sampling weight for the kth sampled tax return, $k \in s$. The sampling weight represents a certain number of population units in the finite population. With Bernoulli sampling within each stratum, we have epsem sampling within each stratum, i.e., the sampling weights are the same for all the sampled units belonging to the same stratum. The weights vary across strata, due to disproportionate allocation of the sample into different strata. Our domain cuts across the design strata, so weights of sampled units inside a domain are generally different.

Let

$$Y_d = \sum_{k \in U_d} y_k$$

denote the population total for the dth domain (excluding the tax returns belonging to the centainty stratum) and y_k is the value of the study variable for the kth population unit. In order to understand the extent and cause of errors in the IRS file, we consider the estimation of

$$R = Y/X ,$$

where Y [X] denotes the AGI population total that the SOI [IRS] file corresponds to. We know X , but not Y. We estimate R for all the $D \times G$ cells [R_{dg}], D domains [R_d], G groups [R_g] and for the nation [R_N].

Let s_d, s_g and s_{dg} denote the set of sampled units belonging to domain d, group g and cell formed by dth domain and gth group formed by a categorized size of AGI. Let $s_{A;c}$ denote the set of sampled units in an arbitrary set of sampled units A that are common between the SOI and IRS files. For example, $s_{dg;c}$ denotes the set of samples in domain d and group g that are common between the SOI and IRS files. The notations $s_{d;c}$, $s_{g;c}$ and $s_{N;c}$ denote similar sets for the domain d, group g, and the nation. Note that we may not introduce the new symbols $s_{dg;c}$, $s_{d;c}$, $s_{g;c}$, and $s_{N;c}$ if there is a one-to-one correspondence between the SOI sample and IRTF. We estimate R_{dg} , R_d , R_g and R_N by:

$$\hat{R}_{dg} = \hat{Y}_{dg;c} \big/ \hat{X}_{dg;c} ,$$
$$\hat{R}_d = \hat{Y}_{d;c} \big/ \hat{X}_{d;c} ,$$
$$\hat{R}_g = \hat{Y}_{g;c} \big/ \hat{X}_{g;c} ,$$
$$\hat{R}_N = \hat{Y}_{N;c} \big/ \hat{X}_{N;c} ,$$

where the numerator and denominator components are the weighted sum of y_k and x_k over the appropriate summation, respectively. For example, with $\hat{R}_{dg}$, we have

$$\hat{Y}_{dg;c} = \sum_{k \in s_{dg;c}} w_k y_k \ , \ \hat{X}_{dg;c} = \sum_{k \in s_{dg;c}} w_k x_k .$$

If the IRS file is error free, we would expect the above ratios to be exactly 1. But, since there are errors in the IRS data, we expect them to vary around 1. For example, the figure in A.1 at the end of this paper contains $\hat{R}_d$ for each variable of interest. A vertical reference line of one is drawn and the states are sorted by their number of noncertainty units in the sample. The $\hat{R}_d$'s fluctuate around one for all variables, particularly when the state sample size decreases (and sampling variance increases). They also fluctuate more from one for variables that are more susceptible to the errors: that scale is 0.80 to 1.20 (compared to 0.99 to 1.04 for the "less" susceptible ones).

We consider seven direct estimators:

$$\hat{Y}_{dD1} = \sum_{k \in s_d} w_k y_k \ , \qquad (1)$$

$$\hat{Y}_{dD2} = N_d \times \sum_{k \in s_d} w_k y_k \Big/ \sum_{k \in s_d} w_k \ , \qquad (2)$$

$$\hat{Y}_{dD3} = \hat{R}_d X_d \ , \qquad (3)$$

$$\hat{Y}_{dD4} = \sum_g \left[N_{dg} \times \sum_{k \in s_{dg}} w_k y_k \Big/ \sum_{k \in s_{dg}} w_k \right], \qquad (4)$$

$$\hat{Y}_{dD5} = \sum_g \hat{R}_{dg} X_{dg} \ , \qquad (5)$$

$$\hat{Y}_{dD6} = \hat{Y}_{dD1} + \hat{R}_N (X_d - \hat{X}_d) \ , \qquad (6)$$

$$\hat{Y}_{dD7} = \hat{Y}_{dD1} + \sum_g \hat{R}_g (X_{dg} - \hat{X}_{dg}) \ . \qquad (7)$$

These are equal to or are various forms of the expansion estimator, weighted survey mean estimator, combined ratio estimator, poststratification estimator, separate ratio estimator, combined survey regression estimator, and separate regression estimator, respectively. They are "direct" estimators since all involve sample-based components at the small areas levels. The benefit of direct estimators is that they are completely or nearly design-unbiased estimators for the population total. However, they are subject to higher sampling variability, since they are based on the number of returns within each state (or state crossed with AGI group), which can be small.

4. Synthetic Estimators

We consider five synthetic estimators:

$$\hat{Y}_{dS1} = \sum_g N_{dg} \times \sum_{k \in s_g} w_k y_k \Big/ \sum_{k \in s_g} w_k \ , \qquad (8)$$

$$\hat{Y}_{dS2} = \hat{R}_N X_d \ , \qquad (9)$$

$$\hat{Y}_{dS3} = \sum_g \hat{R}_g X_{dg} \ , \qquad (10)$$

$$\hat{Y}_{dS4} = \sum_g \hat{Y}_{g1} \times X_{dg} \big/ X_g \ , \qquad (11)$$

$$\hat{Y}_{dS5} = \sum_g \hat{Y}_{g1} \times N_{dg} \big/ N_g \ . \qquad (12)$$

These estimators involve combining information across states and/or AGI groups to estimate the state-level totals. Estimators (8) and (12) are a form of (4), (9) of (3), and (10) and (11) are a form of (5). Due to implicit assumptions with each (see, e.g., section 4.2.1 in Rao 2003), they may not necessarily be design-unbiased. However, they may have lower variances, resulting in overall lower total error.

5. Composite Estimators

To overcome the problems separately associated with the direct and synthetic estimators, we also examine composite estimators. They have the following general form:

$$\hat{Y}_{dC} = \hat{\phi}_d \hat{Y}_{dD} + (1 - \hat{\phi}_d) \hat{Y}_{dS} \ ,$$

where $\hat{Y}_{dD}$ is a direct estimator for the state- total, $\hat{Y}_{dS}$ is a synthetic estimator, and $\hat{\phi}_d$ is a "suitably chosen weight" on the direct estimator (expression 4.3.1 in Rao 2003). We present results using two composite estimator weights:

$$\hat{\phi}_d = \begin{cases} 1 & \text{if } \hat{N}_d / N_d \geq 1 \\ \left[\hat{N}_d / N_d \right]^2 & \text{if } \hat{N}_d / N_d < 1 \end{cases} \qquad (13)$$

$$\hat{\phi} = \begin{cases} 0 & \text{if } \sum_d var(\hat{Y}_{dD}) \Big/ \sum_d (\hat{Y}_{dD} - \hat{Y}_{dS})^2 \geq 1 \\ 1 - \sum_d var(\hat{Y}_{dD}) \Big/ \sum_d (\hat{Y}_{dD} - \hat{Y}_{dS})^2 & \text{otherwise} \end{cases}, \qquad (14)$$

with various combinations of direct estimators (1)-(7) combined with synthetic estimators (8)-(12). The weight in (13) was proposed by Sarndal and Hidiroglou (1989), while (14) is a form of the James-Stein estimator (expression 4.4.3 in Rao 2003) with a common weight. They are different in that (13) depends on the state but not variable of interest, while (14) depends on the variable but not state.

6. Results

Table A.2 at the end of this paper contains the IRS totals X_d for each variable and collapsed state group. This allows for a useful comparison between estimates similar to those published by SOI and our alternatives.

Figure A.3 contains plots of the relative differences of the direct estimates in (1), shown in A.2, to various alternatives, for variables that are less susceptible to error. The estimates are referenced with the subscript in each plot. Three combinations of direct and synthetic estimators are considered: (1) and (8); (1) and (11); and (2) and (12). Combined with the two weight choices (13) and (14), we have six composite estimators. These are labeled "C" and "JS," with the direct and synthetic number, resp. For example, "C 1,1" refers to a composite estimator with (1) as the direct, (8) as the synthetic, and weight (13). Relative differences outside (-10%, 10%)

were truncated and the states are sorted on each horizontal axis by ascending size of the CV of $\hat{Y}_{dD1}$ in (1).

For AGI, the relative differences to the IRS totals were within 2 percent for all groups except NV/WY/SD and HI/Other, as this variable had lower amounts of both sampling error in the direct estimates and nonsampling error in the IRS totals. Salaries and Wages and Total Tax Liability had similar patterns as noted in AGI, but somewhat larger relative differences. The Earned Income Tax Credit plot showed even larger relative differences. This was caused by larger sampling errors (e.g., the highest CV of $\hat{Y}_{dD1}$ was 18%, compared to 4% for AGI). Since this credit was also claimed only by lower income taxpayers, there were several zero values in both the SOI and IRS data given the sample design described in section 2.1. Differences between the AGI-category level ratios and one resulted in poorer synthetic estimates (with extremely high relative differences) using (9) and (12) for both larger and smaller states. Direct estimates (3) and (4) looked stable.

Figure A.4 contains the same ratio plots the direct estimates from Table A.1, for variables that are more susceptible to error. Relative errors outside (-100%, 100%) were truncated and again the states were sorted by the CV of $\hat{Y}_{dD1}$. These variables had much different results for the different estimators, particularly for the smaller state groups - there was not a clear pattern due to the sampling error, as noted in Figure A.3. The relative differences were most often highest for the HI/Other group, where the SOI sample estimates are very far from the IRS totals. This also caused differences in the direct or synthetic estimates that used the estimated group population size form the SOI sample (about 790,000 returns) and the IRS total (about 1.5 million). The same instability with estimators (9) and (12) occurred with all of these variables. However, the IRS totals here are considered less reliable due to nonsampling error, resulting in larger relative differences.

7. Conclusions, Limitations, and Future Considerations

In general, the direct estimates are further from the IRS totals (particularly for smaller states), while the synthetic are closer, and the composite are a compromise between the two. Starting in Tax Year 2005, the CWHS will become a ten percent stratified simple random sample. This means that approximately 65,000 non-certainty units will be added to SOI's sample, which will increase the reliability of the direct estimates.

Our comparisons were between various estimated totals and the corresponding IRS ones. We should also compare the direct estimates' sampling error to the mean square error in the synthetic and composite ones. These are more difficult to compute, particularly since more sample units are required for reliable estimates. Another alternative to consider is using composite estimatorsfrom (1), (3) and (4) as the direct estimates and the IRS totals as the synthetic ones. A natural extension of the composite estimates is small area modeling, which is also currently under consideration. Ultimately we are also interested in the state-level estimates, but the collapsing of states into groups allowed for a useful comparison between the alternatives, it also demonstrated that the direct estimates were affected by sampling error in smaller states. Thus, adjustments are needed when applying them simply at the state level.

REFERENCES

Hostetter, S., Czajka, J. L., Schirm, A. L., and O'Conor, K. (1990), "Choosing the Appropriate Income Classifier for Economic Tax Modeling," *Proceedings of the Section on Survey Research Methods Section*, American Statistical Association, pp. 419-424.

Internal Revenue Service, (2005), "Explanation of Terms," *Statistics of Income – 2004 Individual Income Tax Returns, Internal Revenue Service, Publication 1304*, pp. 117-147.

Statistics of Income Bulletin, Spring 2006, Washington, DC, 2006, pp. 261-313.

Rao, J.N.K. (2003). *Small Area Estimation*, John Wiley & Sons: New York.

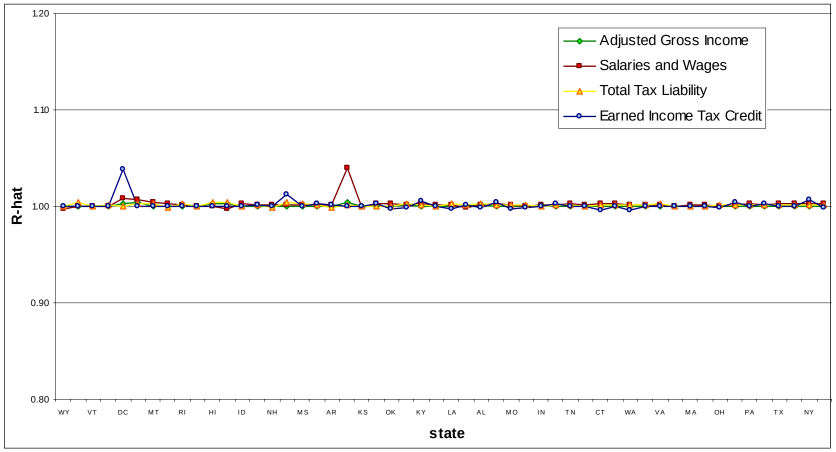
Sarndal, C.E., and Hidiroglou, M.A. (1989), Small Domain Estimation: A Conditional Analysis, Journal of the American Statistical Association, 84, pp. 266-275.

Scali, J, and Testa, V. (2006), *Statistics of Income – 2004 Individual Income Tax Returns, Internal Revenue Service, Publication 1304*, pp. 23-27.

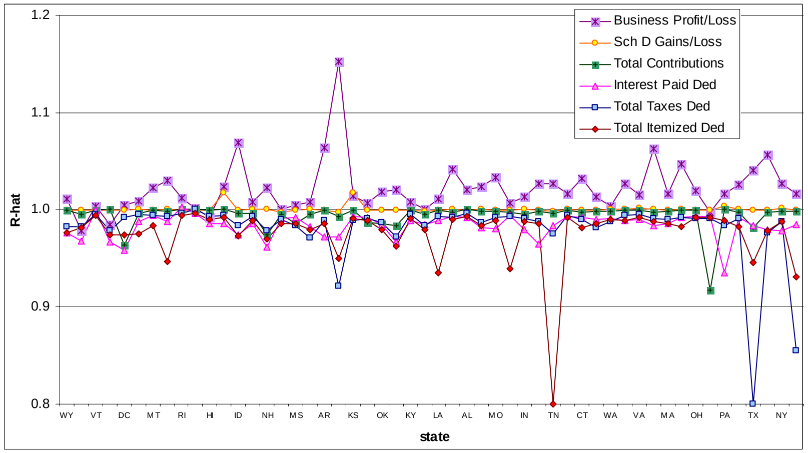
Weber, M., (2004), "The Statistics of Income 1979-2002 Continuous Work History Sample Individual Tax Return Panel," *Proceedings of the Section on Survey Research Methods*, American Statistical Association.

A.1. Estimated $\hat{R}_d$'s, by Variable Type, Variable of Interest, and State (sorted by ascending state noncertainty sample size)

Less Susceptible to Error



More Susceptible to Error

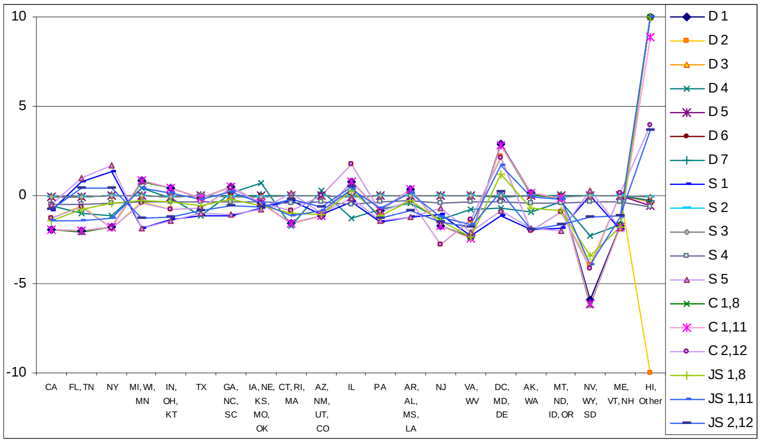


A.2. Table of IRS Totals X_d (in Thousands of Dollars), by Collapsed State Group

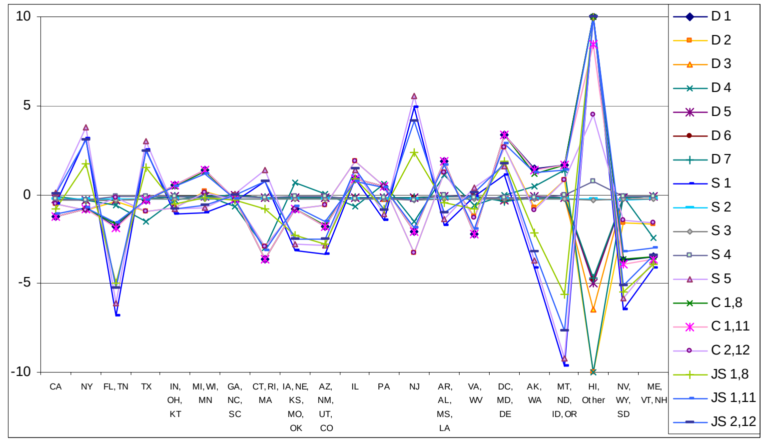
	Less Susceptible to Error				*More Susceptible to Error*					
State Group	Adjusted Gross Income	Salaries and Wages	Total Tax Liability	Earned Income Tax Credit	Business Profit / Loss	Schedule D Capital Gains/Loss	Total Contributions	Taxes Paid	Interest Paid Deduction	Total Itemized Deductions
AK, WA	169,450,168	121,048,669	22,235,015	663,339	6,219,819	11,796,324	3,695,776	6,195,165	11,176,349	3,695,776
AR, AL, MS, LA	240,278,162	181,097,537	26,373,363	3,522,206	8,066,558	9,685,873	6,321,664	9,344,287	10,274,109	6,321,664
AZ, NM, UT, CO	309,176,619	222,883,909	37,248,137	1,792,852	10,050,014	22,747,510	8,439,119	14,321,997	21,963,900	8,439,119
CA	881,752,963	624,514,425	121,339,747	4,449,344	44,266,526	73,195,955	21,867,927	67,399,138	71,443,413	21,867,927
CT, RI, MA	339,501,588	238,252,628	51,940,323	879,239	14,452,095	27,862,641	7,083,498	23,445,637	17,701,780	7,083,498
DC, MD, DE	197,805,604	145,614,770	26,519,027	785,361	6,133,845	11,962,072	6,034,002	13,795,287	12,607,868	6,034,002
FL, TN	529,878,356	349,964,046	72,055,545	3,955,877	18,799,220	53,912,196	12,509,213	16,999,688	26,594,987	12,509,213
GA, NC, SC	428,830,827	322,440,428	49,915,635	3,990,066	13,042,714	22,351,530	13,201,226	24,522,210	28,768,405	13,201,226
HI, Other	77,696,909	73,436,533	9,644,685	196,176	2,709,878	7,738,281	1,019,693	2,824,833	2,926,349	1,019,693
IA, NE, KS, MO, OK	325,128,782	240,695,352	37,419,546	2,140,592	9,668,713	14,985,751	7,800,656	16,992,219	15,056,976	7,800,656
IL	312,951,784	228,115,769	42,656,588	1,576,538	9,333,379	21,421,047	7,054,523	15,938,756	16,575,098	7,054,523
IN, OH, KT	441,711,523	336,208,255	50,887,852	2,768,805	13,651,937	16,774,468	9,202,408	23,364,434	23,091,120	9,202,408
ME, VT, NH	74,966,026	54,891,225	9,186,096	292,609	3,739,824	4,999,477	1,181,216	3,904,904	3,482,376	1,181,216
MI, WI, MN	471,487,557	354,296,826	57,030,989	2,067,922	13,387,287	20,990,791	10,832,989	27,737,408	25,623,815	10,832,989
MT, ND, ID, OR	127,237,758	89,502,052	14,287,382	742,205	5,008,009	8,444,990	3,070,032	8,569,985	7,509,054	3,070,032
NJ	264,917,673	199,028,894	39,188,251	857,954	9,598,198	14,729,732	5,533,706	22,336,098	13,915,865	5,533,706
NV, WY, SD	90,163,667	57,805,634	12,298,955	423,539	2,938,308	12,274,045	2,085,551	2,929,571	5,025,388	2,085,551
NY	509,011,438	359,825,754	75,885,191	2,672,975	18,993,061	45,111,257	14,454,792	44,903,606	21,255,412	14,454,792
PA	278,531,309	207,054,076	35,026,827	1,304,085	9,707,338	13,124,940	5,687,268	14,473,936	11,985,528	5,687,268
TX	448,956,879	338,710,156	59,941,678	4,509,906	18,836,804	26,138,888	9,927,578	15,421,149	17,658,406	9,927,578
VA, WV	225,665,995	168,339,288	28,857,291	1,122,083	7,402,708	11,953,120	5,195,825	11,621,403	13,660,405	5,195,825

A.3. Comparison Plots of Relative Differences of Alternative Estimator to IRS Totals, by Variable, Collapsed State Group, and Estimator (Variables Less Susceptible to Error, sorted by size of the CV of the Direct)

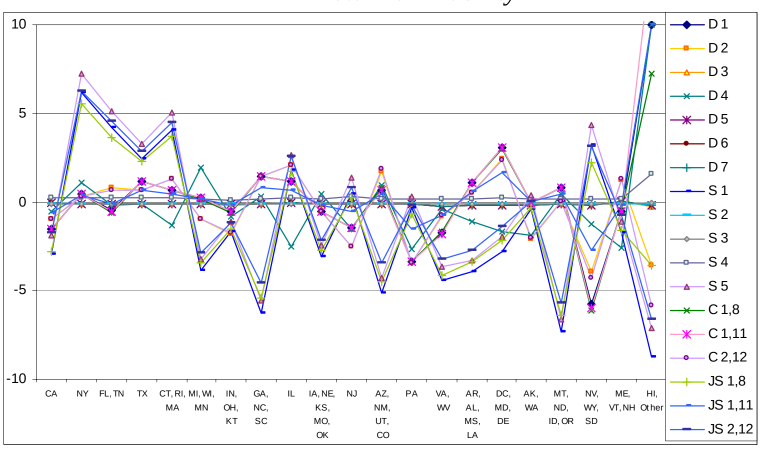
Adjusted Gross Income



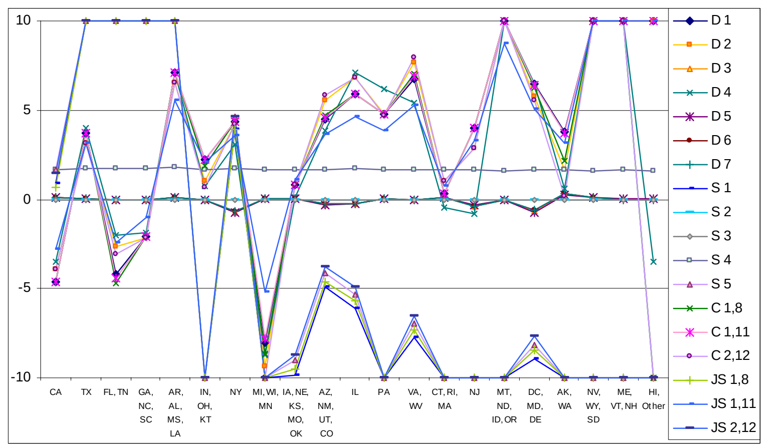
Salaries and Wages



Total Tax Liability

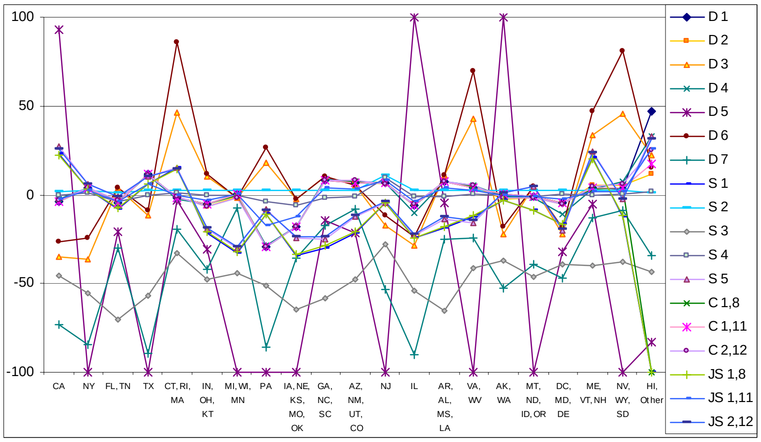


Earned Income Tax Credit

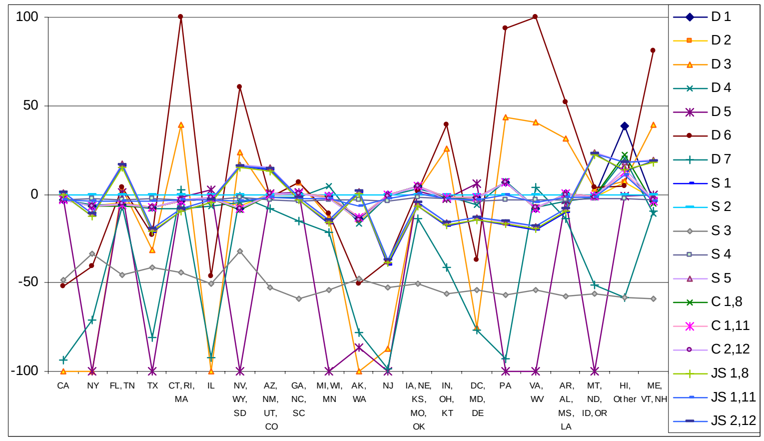


A.4. Comparison Plots of Relative Differences of Alternative Estimator to Direct Estimates, by Variable, Collapsed State Group, and Estimator (Variables More Susceptible to Error, sorted by size of the CV of the Direct)

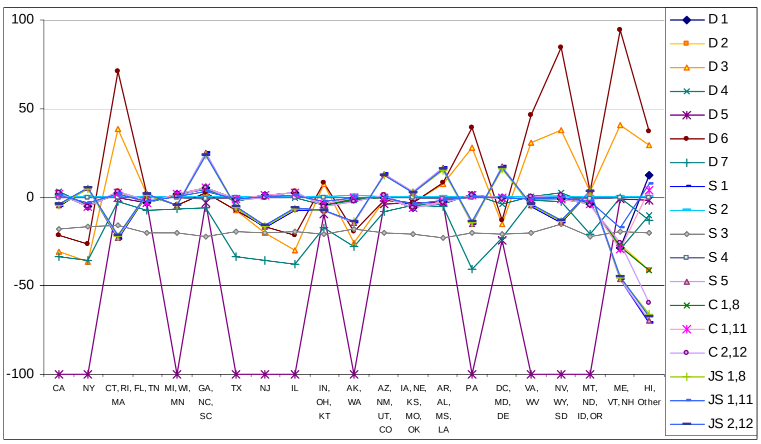
Net Schedule C Business Profit/Loss



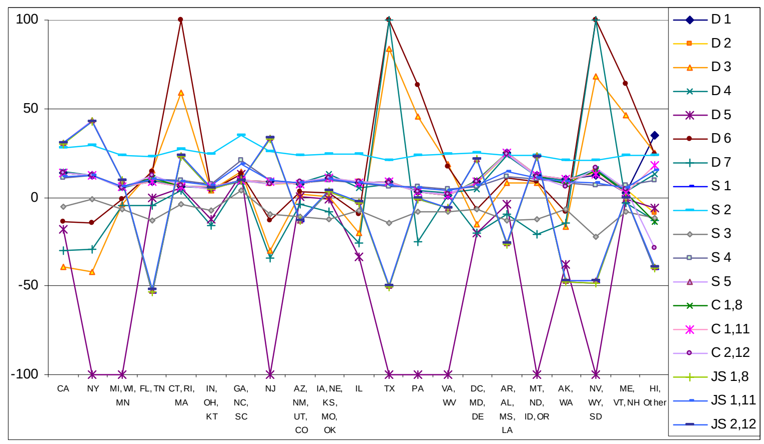
Net Schedule D Capital Gains/Loss



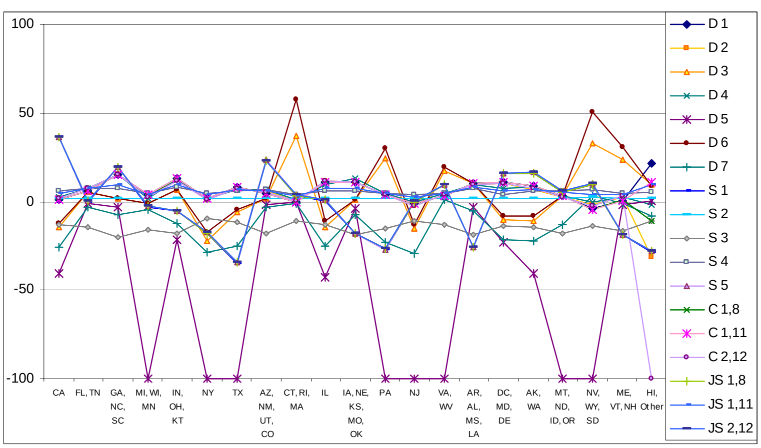
Total Contributions



Total Taxes Paid Deduction



Interest Paid Deduction



Total Itemized Deductions