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

by Kimberly Henry, Michael Strudler, and William Chen
An Empirical Evaluation of Various Direct, Synthetic, and Traditional Composite Small-Area Estimators

Kimberly Henry, Michael Strudler, and William Chen, Internal Revenue Service

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.

Data Description

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 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.
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 percent 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 noncertainty 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.

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 noncertainty 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 locations on Form 1040 and a brief description, in Table 1.

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 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.

Direct Estimators

Let \( y_k \) be the value of the characteristic of interest for the \( k \)th 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 Bernoulli 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 \( k \)th 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
AN EMPIRICAL EVALUATION OF VARIOUS DIRECT, SYNTHETIC, AND TRADITIONAL COMPOSITE SMALL-AREA ESTIMATORS

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

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

a: page numbers from IRS 2005.

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

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

different strata. Our domain cuts across the design strata, so that 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 \(d\)th domain (excluding the tax returns belonging to the certainty stratum), and \(y_k\) is the value of the study variable for the \(k\)th 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 \(Y\) 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 \(d\)th domain and \(g\)th group formed by a categorized size of
AGI. Let \( s_{d,c} \) 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:

\[
\hat{R}_{dg} = \frac{\hat{Y}_{dg,c}}{\hat{X}_{dg,c}},
\]

\[
\hat{R}_d = \frac{\hat{Y}_{d,c}}{\hat{X}_{d,c}},
\]

\[
\hat{R}_g = \frac{\hat{Y}_{g,c}}{\hat{X}_{g,c}},
\]

\[
\hat{R}_N = \frac{\hat{Y}_{N,c}}{\hat{X}_{N,c}},
\]

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

\[
\hat{Y}_{dg,c} = \sum_{k \in s_{dg,c}} w_k y_k, \quad \hat{X}_{dg,c} = \sum_{k \in s_{dg,c}} w_k x_k.
\]

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, Figure 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:

\[
\hat{Y}_{dD1} = \sum_{k \in s_d} w_k y_k, \tag{1}
\]

\[
\hat{Y}_{dD2} = N_d \times \left( \sum_{k \in s_d} w_k y_k / \sum_{k \in s_d} w_k \right), \tag{2}
\]

\[
\hat{Y}_{dD3} = \hat{R}_d X_d, \tag{3}
\]

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

\[
\hat{Y}_{dD5} = \sum_{g} \hat{R}_{dg} X_{dg}, \tag{5}
\]

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

\[
\hat{Y}_{dD7} = \hat{Y}_{dD1} + \sum_{g} \hat{R}_g (X_{dg} - \hat{X}_{dg}). \tag{7}
\]

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-area level. 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.

### Synthetic Estimators

We consider five synthetic estimators:

\[
\hat{Y}_{dS1} = \sum_{g} N_{dg} \times \left( \sum_{k \in s_{dg}} w_k y_k / \sum_{k \in s_{dg}} w_k \right), \tag{8}
\]

\[
\hat{Y}_{dS2} = \hat{R}_N X_d, \tag{9}
\]

\[
\hat{Y}_{dS3} = \sum_{g} \hat{R}_g X_{dg}, \tag{10}
\]

\[
\hat{Y}_{dS4} = \sum_{g} \hat{Y}_{g1} \times X_{dg} / X_g, \tag{11}
\]

\[
\hat{Y}_{dS5} = \sum_{g} \hat{Y}_{g1} \times N_{dg} / N_g. \tag{12}
\]
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.

**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:

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

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

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.

**Results**

Figure 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 on the horizontal axis are sorted by the size of the coefficient of variation CV of \( \hat{Y}_{dD} \) in (1). 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, respectively. 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 percent, 10 percent) were truncated.

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 non-sampling 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 \( \hat{Y}_{dD} \) was 18 percent, compared to 4 percent 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 the SOI Sample. 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 as the direct estimates from Figure A.1, for variables that are more susceptible to error. Relative errors outside (-100 percent, 100 percent) were truncated, and, again, the States were sorted by the CV of \( \hat{Y}_{dD} \) . 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 from 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.

**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 10-percent stratified simple random sample. This means that approximately 65,000 noncertainty 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 estimators from (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, and 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**


Figure A.1. Estimated $\hat{R}_d$'s by Variable Type, Variable of Interest, and State (Sorted by Ascending State Noncertainty Sample Size)

Variables less susceptible to error

Variables more susceptible to error
Figure A.2. Table of IRS Totals \( X_d \) (in Thousands of Dollars), by Collapsed State Group

<table>
<thead>
<tr>
<th>State Group</th>
<th>Adjusted Gross Income</th>
<th>Salaries and Wages</th>
<th>Total Tax Liability</th>
<th>Earned Income Tax Credit</th>
<th>Business Profit / Loss</th>
<th>Schedule D Capital Gains/Loss</th>
<th>Total Contributions</th>
<th>Taxes Paid</th>
<th>Interest Paid Deduction</th>
<th>Total Itemized Deductions</th>
</tr>
</thead>
<tbody>
<tr>
<td>AK, WA</td>
<td>169,450,168</td>
<td>121,048,669</td>
<td>22,235,015</td>
<td>663,339</td>
<td>6,219,819</td>
<td>11,796,324</td>
<td>3,695,776</td>
<td>6,195,165</td>
<td>11,176,349</td>
<td>3,695,776</td>
</tr>
<tr>
<td>AR, AL, MS, LA</td>
<td>240,278,162</td>
<td>181,097,537</td>
<td>26,373,363</td>
<td>3,522,206</td>
<td>8,066,558</td>
<td>9,685,873</td>
<td>6,321,664</td>
<td>9,344,287</td>
<td>10,274,109</td>
<td>6,321,664</td>
</tr>
<tr>
<td>AZ, NM, UT, CO</td>
<td>309,176,619</td>
<td>222,883,909</td>
<td>37,248,137</td>
<td>1,792,852</td>
<td>10,050,014</td>
<td>22,747,510</td>
<td>8,439,119</td>
<td>14,321,997</td>
<td>21,963,900</td>
<td>8,439,119</td>
</tr>
<tr>
<td>CA</td>
<td>881,752,963</td>
<td>624,514,425</td>
<td>121,399,747</td>
<td>4,449,344</td>
<td>44,266,526</td>
<td>73,195,955</td>
<td>21,867,927</td>
<td>67,399,138</td>
<td>71,443,413</td>
<td>21,867,927</td>
</tr>
<tr>
<td>CT, RI, MA</td>
<td>339,301,588</td>
<td>238,252,628</td>
<td>51,940,323</td>
<td>879,239</td>
<td>14,452,095</td>
<td>27,862,461</td>
<td>7,083,498</td>
<td>23,445,637</td>
<td>17,701,780</td>
<td>7,083,498</td>
</tr>
<tr>
<td>DC, MD, DE</td>
<td>197,805,664</td>
<td>145,614,770</td>
<td>26,519,027</td>
<td>785,261</td>
<td>6,133,845</td>
<td>11,962,072</td>
<td>6,034,002</td>
<td>13,795,287</td>
<td>12,607,868</td>
<td>6,034,002</td>
</tr>
<tr>
<td>FL, TN</td>
<td>529,878,356</td>
<td>349,964,046</td>
<td>72,055,545</td>
<td>3,955,877</td>
<td>18,799,220</td>
<td>53,912,196</td>
<td>12,509,213</td>
<td>16,996,688</td>
<td>21,598,213</td>
<td>12,509,213</td>
</tr>
<tr>
<td>GA, NC, SC</td>
<td>428,830,827</td>
<td>322,440,428</td>
<td>49,915,635</td>
<td>3,990,066</td>
<td>13,042,714</td>
<td>22,351,530</td>
<td>13,201,226</td>
<td>24,522,310</td>
<td>28,768,405</td>
<td>13,201,226</td>
</tr>
<tr>
<td>HI, Other</td>
<td>77,696,909</td>
<td>73,436,533</td>
<td>9,644,685</td>
<td>196,176</td>
<td>2,709,878</td>
<td>7,738,281</td>
<td>1,019,693</td>
<td>2,824,833</td>
<td>2,926,349</td>
<td>1,019,693</td>
</tr>
<tr>
<td>IA, NE, KS, MO, OK</td>
<td>325,128,782</td>
<td>240,695,352</td>
<td>37,419,546</td>
<td>2,140,592</td>
<td>9,668,713</td>
<td>14,985,751</td>
<td>7,800,656</td>
<td>16,992,219</td>
<td>15,056,976</td>
<td>7,800,656</td>
</tr>
<tr>
<td>IL</td>
<td>312,951,784</td>
<td>228,115,769</td>
<td>42,656,588</td>
<td>1,576,538</td>
<td>9,333,379</td>
<td>21,421,047</td>
<td>7,054,523</td>
<td>15,938,756</td>
<td>16,759,098</td>
<td>7,054,523</td>
</tr>
<tr>
<td>IN, OH, KT</td>
<td>441,711,523</td>
<td>336,208,255</td>
<td>50,887,852</td>
<td>2,768,805</td>
<td>13,651,937</td>
<td>16,774,468</td>
<td>9,202,408</td>
<td>23,364,434</td>
<td>27,091,120</td>
<td>9,202,408</td>
</tr>
<tr>
<td>ME, VT, NH</td>
<td>74,966,026</td>
<td>54,891,235</td>
<td>9,186,996</td>
<td>292,609</td>
<td>3,739,824</td>
<td>4,999,477</td>
<td>1,181,216</td>
<td>3,904,904</td>
<td>3,482,376</td>
<td>1,181,216</td>
</tr>
<tr>
<td>MI, WI, MN</td>
<td>471,847,557</td>
<td>354,296,826</td>
<td>57,030,989</td>
<td>2,067,922</td>
<td>13,387,287</td>
<td>20,990,791</td>
<td>10,832,989</td>
<td>27,373,408</td>
<td>25,623,815</td>
<td>10,832,989</td>
</tr>
<tr>
<td>MT, ND, ID, OR</td>
<td>127,237,758</td>
<td>89,502,052</td>
<td>14,287,382</td>
<td>742,205</td>
<td>5,008,009</td>
<td>8,444,990</td>
<td>3,070,032</td>
<td>8,569,985</td>
<td>7,509,054</td>
<td>3,070,032</td>
</tr>
<tr>
<td>NJ</td>
<td>264,917,673</td>
<td>199,028,894</td>
<td>39,188,251</td>
<td>857,954</td>
<td>9,598,198</td>
<td>14,729,732</td>
<td>5,533,706</td>
<td>22,336,098</td>
<td>13,915,865</td>
<td>5,533,706</td>
</tr>
<tr>
<td>NV, WY, SD</td>
<td>90,163,667</td>
<td>57,805,634</td>
<td>12,298,955</td>
<td>423,539</td>
<td>2,938,308</td>
<td>12,274,045</td>
<td>2,085,551</td>
<td>2,929,571</td>
<td>5,025,388</td>
<td>2,085,551</td>
</tr>
<tr>
<td>NY</td>
<td>509,011,438</td>
<td>359,825,754</td>
<td>75,885,191</td>
<td>2,672,975</td>
<td>18,993,061</td>
<td>45,111,257</td>
<td>14,454,792</td>
<td>44,903,606</td>
<td>21,255,412</td>
<td>14,454,792</td>
</tr>
<tr>
<td>PA</td>
<td>278,531,309</td>
<td>207,054,076</td>
<td>35,026,827</td>
<td>1,304,085</td>
<td>9,707,338</td>
<td>13,124,940</td>
<td>5,687,268</td>
<td>14,473,936</td>
<td>11,985,528</td>
<td>5,687,268</td>
</tr>
<tr>
<td>TX</td>
<td>448,956,879</td>
<td>338,710,156</td>
<td>59,941,678</td>
<td>4,508,906</td>
<td>18,836,804</td>
<td>26,138,888</td>
<td>9,875,578</td>
<td>15,421,149</td>
<td>17,658,406</td>
<td>9,927,578</td>
</tr>
</tbody>
</table>
Figure A.3. Comparison Plots of Percentage Relative Difference of Alternative Estimators to IRS Totals, by Variable, Collapsed State Group, and Estimator for Variables Less Susceptible to Error

Adjusted gross income

Salaries and wages
Figure A.3. Comparison Plots of Percentage Relative Difference of Alternative Estimators to IRS Totals, by Variable, Collapsed State Group, and Estimator for Variables Less Susceptible to Error—Continued

Total tax liability

Earned income tax credit
Figure 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
Figure 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)—Continued

**Total contributions**

![Graph showing relative differences in total contributions across states and variables.]

**Total taxes paid deduction**

![Graph showing relative differences in total taxes paid deduction across states and variables.]

- **D1**
- **D2**
- **D3**
- **D4**
- **D5**
- **D6**
- **D7**
- **S1**
- **S2**
- **S3**
- **S4**
- **S5**
- **C1,8**
- **C1,11**
- **C2,12**
- **JS 1,8**
- **JS 1,11**
- **JS 2,12**

---

HENRY, STRUDLER, AND CHEN   2007 SOI PAPER SERIES
Figure 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)—Continued

Interest paid deduction

![Interest paid deduction graph]

Total itemized deductions

![Total itemized deductions graph]