An Assessment of the Need for a Redesign of the Statistics of Income Individual Tax Sample

Final Report

March 31, 2014

John L. Czajka
Amang Sukasih
Brendan Kirwan

MATHEMATICA
Policy Research

Contract Number:
TIRNO-07-Z-00019 (GS-10F-0050L)

Mathematica Reference Number:
40133.001

Submitted to:
Internal Revenue Service
Statistics of Income Division
P.O. Box 2608
Washington, DC 20013-2608
Technical Representative: Michael Jung

Submitted by:
Mathematica Policy Research
1100 1st Street, NE
12th Floor
Washington, DC 20002-4221
Telephone: (202) 484-9220
Facsimile: (202) 863-1763
Project Director: John L. Czajka

An Assessment of the Need for a Redesign of the Statistics of Income Individual Tax Sample

Final Report

March 31, 2014

John L. Czajka
Amang Sukasih
Brendan Kirwan

MATHEMATICA
Policy Research

CONTENTS

EXECUTIVE SUMMARY

Each year the Statistics of Income (SOI) Division of the Internal Revenue Service (IRS) draws a sample of individual and sole proprietorship tax returns, abstracts and edits a large number of data items, and prepares a microdatabase that the Treasury Department, the Congress, and selected other agencies use for tax policy analysis. The SOI Division also produces a public use file so that academic and policy researchers as well as other federal agencies can have access to some of the same information for their own tax policy analyses.

The last formal redesign of the SOI Individual sample occurred in the late 1980s, and it was the result of a collaborative effort among staff in the SOI Division, the Office of Tax Analysis (OTA), and Mathematica Policy Research. The sample was designed with a target size of 95,000 returns. This grew to 126,000 returns (including a foreign supplement) by the time the design was implemented for the 1991 tax year. Since then the size of the Individual sample has increased to more than 330,000 returns. Part of the growth was due to a five-fold increase in the minimum sampling rate when the SOI Division decided to include returns that were being captured for an OTA panel and, therefore, were available at minimal additional cost. The residual increase was due primarily to upward movement in the income distribution, which was only partially offset by indexing the income stratifier, and to growth in the number of high-income nontaxable returns, which are selected with certainty under a provision in the Tax Reform Act of 1976.

Even in the absence of these substantial changes in the size and composition of the SOI Individual sample, a review of the sample design is overdue. After more than 25 years the needs of the sample's customers may have changed, and the characteristics of the filing population may have evolved in ways that diminish the effectiveness and efficiency of the design. Furthermore, new technology and other factors may have altered the cost of processing the sample. With the changes we have noted, the SOI Individual sample no longer conforms to the 1980s redesign, although stratum definitions and many of the sampling rates remain the same. It is appropriate to ask whether the current sample meets the needs of its users as fully as it could and, even if it does, whether in the current climate of reduced agency budgets but growing demands, a smaller sample, perhaps configured differently, could meet these needs.

In preparing this report, Mathematica reviewed the design of the current Individual sample and the principal uses of the Individual sample data. Mathematica also met with the major customers of the Individual tax data to discuss their uses of the sample data and to solicit their views on particular elements of the sample design and the products that are created from the data. In conducting this study our objective was not to develop a new design but, rather, to identify areas where improvements to the current sample may be possible and desirable. The report develops recommendations for improving the design of the sample, consistent with the current needs of major users, good statistical practice, and budgetary considerations.

We find that the SOI Individual sample continues to serve the needs of its principal customers exceedingly well. The sample currently provides substantially more precise estimates than it was originally designed to provide, and it supplies users with a very large case base for analyzing a wide range of tax policy options. The sample is larger than it needs to be, however, and while unit editing costs for Individual returns have declined markedly, a smaller, more efficient Individual sample would enable the SOI Division to reallocate some of its resources toward other Division needs.

With suggestions from SOI Division customers and staff, we conducted an empirical assessment of prospective changes to the income measure used to stratify the sample and to the index used to adjust income for inflation. To the current income measure we added one component, removed another, and replaced three others with alternatives. We also replaced the current index, a price index based on the Gross Domestic Product, with an alternative based on personal income, which has shown more rapid growth and, therefore, is more effective in dampening the effects of upward movement in the income distribution. With the alternative income definition and index, the size of the 2008 Individual sample would have been reduced by 23 percent and the editing costs by as much as 40 percent. Without a change in the sampling rates by stratum, however, the precision of the sample estimates of key income and tax variables would have declined significantly relative to the larger current sample. Because the sample size reduction was concentrated among higher income returns, sample sizes for income or tax items that are of particular interest for policy analysis would have fallen by as much as one-half. Coefficients of variation (CVs) for estimates of returns increased with the alternative design for all 34 items evaluated, and for all but 5 the increases were 10 percent or higher, with 6 exceeding 20 percent. For estimates of amounts 25 of the CVs increased by more than 20 percent, and 9 increased by more than 50 percent.

We would not expect the current sampling rates to be optimal with a new stratifier and index, and they may not be optimal for the current design either, given the changes to the sample composition that have ensued since the design was implemented. Using Individual sample data for 2008, we estimated the optimal allocation for each of 34 dollar amount fields with the new stratifier and index and their implied sample size of 252,588. In so doing we constrained the sampling rates in the two specialized strata and the highest positive and negative income strata to be 100 percent, but we did not extend that constraint to the next highest income strata as the current design does. We also constrained the minimum sampling rate to be 0.10 percent, but we did not require that this rate by used in any of the three income classes currently sampled at that rate. The alternative allocations are instructive in how the sampling rates depart from the current design. Only one of the 34 allocations assigned the minimum sampling rate to all of the strata that are currently sampled at this low rate, and only eight assigned a 100 percent rate to a stratum outside of those that we constrained to be sampled at that rate. None of the eight assigned these additional 100 percent rates to both positive and negative income strata, which means that none of the 34 allocations mirrored the current sample allocation with respect to where the minimum and 100 percent rates are used.

We calculated CVs for the 34 variables with their optimal allocations and with the allocations that were optimal for five important variables, including adjusted gross income less deficit and the alternative minimum tax, as only one allocation can be applied in practice. For more than a third of the variables the optimal allocation produced a CV smaller than the current design, indicating a significant potential to improve the precision of the alternative design with just a change in the sampling rates while retaining its smaller sample size. When the precision of all 34 variables was assessed with the optimal allocation for AGI less deficit, only two items had smaller CVs than with the current sample design while 11 additional items had CVs within 10 percent of those obtained with the current design. The CV for AGI less deficit, which is minimized with this allocation, was 24.6 percent higher than with the current design. Ten items had CVs that exceeded their current CVs by larger margins. Further work on an alternative sample design should focus on the income class boundaries as noted below.

Additional empirical findings have implications for proposals to oversample electronic returns and to enhance the panel aspects of the sample through a mechanism that would oversample joint returns. We also found that prior year returns may not be a good proxy for late returns and that the quality of advance estimates of many items has deteriorated markedly over a period of 15 years.

Our recommendations encompass several aspects of the Individual sample design including the retention of some features in their present form and the elimination or modification of others. We recommend:

- Continued use of gross positive and negative income in the income stratifier but replacement of some of their current components and addition of one or more other components
- Assessment of whether gross positive income should be replaced by adjusted gross income when the latter is larger
- Revision of the income stratum boundaries to reflect both inflation and real income growth
- Replacement of the current index, which is based on GDP, with one that is based on personal income
- Retention of form type as the second stratifier, with cross-sectional sampling rates by income class undifferentiated across form types except in foreign study years
- Elimination of sub-stratification by degree of interest
- Retention of certainty selection for high-income nontaxable returns if legally required; otherwise, sampling by stratum at enhanced rates sufficient to meet the annual reporting requirements
- Retention of the current minimum sampling rate of 1 in 1,000, which is very popular with the principal customers
- Continued selection of electronic and paper returns at the same rate
- No additional assessment of the merits of selecting sample returns based on both the primary and secondary SSN
- Retention of prior year returns as an integral part of the processing year sample rather than presenting them as representative of late returns
- Reallocation of the sample to maximize efficiency across a wide range of items in light of the increased minimum sampling rate and the substantial growth of income in the upper tail of the distribution
- Retention of certainty selection for returns with high business gross receipts
- Continuation of current procedures for handling misclassification error, which is likely to be reduced by recommended changes in the income stratifier
- Continuation of current procedures for handling missing returns, which are rare and becoming more so

With these recommendations the basic structure of the current design would be retained, but most of its elements would be modified to some degree.

An important decision in implementing a redesign based on these recommendations is a determination of the boundaries between income classes. We do not recommend simply applying the proposed index to the current stratum boundaries to define the new boundaries, as this

presumes that the income classes derived in this manner are more homogeneous than others that might be considered. Setting the initial boundaries is part of a broader research effort that includes determining a target sample size and an optimal allocation of the sample across strata. The recommended index may play a part in determining these boundaries, but its formal role as a new index would not begin until a revised design had been in place for a year.

With regard to other aspects of the Individual sample we recommend that the SOI Division:

- Maintain the current release schedule for the final file; there is no particular reason to accelerate delivery, but neither should it be delayed
- Develop as an annual product a person-level database of non-filers, using information returns with CWHS SSNs
- Follow up on customer comments about the declining usefulness of advance estimates, and if this decline is related to an actual deterioration in quality, investigate ways to improve the quality of these estimates
- Follow up on customer comments about the declining value of the SOCA study, due to the decreasing proportion of capital asset sales reported on Schedule D
- Explore whether SOI data show evidence of declining quality in SSNs
- Make available to CDW users the recent comparison of SOI and CDW aggregates
- Consider ways to assess the quality of CDW items that are too rare to estimate precisely with the Individual sample; the SOI Division can make an important contribution here
- Determine how any sample design changes might be reflected in the public use file and communicate this information to the major user of these data
- Ascertain what post-audit data might be available and whether it might be used to provide some sense of what the Individual sample data might look like if it were post-audit
- Develop comprehensive documentation of the sample design to supplement the description provided in the Complete Report

Of these the development of a database of non-filers, is the most significant undertaking but the one that will most enhance the value of SOI Individual data to its principal customers. Such an undertaking should build on the work these customers have already produced. A joint effort would further reduce the demand on SOI resources.

I. INTRODUCTION

Each year the Statistics of Income (SOI) Division of the Internal Revenue Service (IRS) draws a sample of individual and sole proprietorship tax returns, abstracts and edits a large number of data items, and prepares a microdatabase that the Treasury Department, the Congress, and selected other agencies use for tax policy analysis. The SOI Division also produces a public-use file (PUF) so that academic and policy researchers as well as other federal agencies can have access to some of the same information for their own tax policy analysis.

The last formal redesign of the SOI Individual sample occurred in the late 1980s, and it was the result of a collaborative effort among staff in the SOI Division, the Office of Tax Analysis (OTA), and Mathematica Policy Research. Before the design was implemented, meetings were held with staff in the Joint Committee on Taxation (JCT) of Congress and the Bureau of Economic Analysis (BEA) within the Department of Commerce to explain the new design. The sample was designed with a target size of 95,000 (Schirm and Czajka 1991), reflecting overall cost considerations and the limited computing resources available to staff in OTA at the time. With the size of the individual samples in the 1980s, OTA staff had to subsample the SOI microdata when running their tax microsimulation model.

When implemented in 1992, for the 1991 tax year, the new design generated a little over 100,000 sample returns, but the size of the Individual sample has increased more than three-fold, exceeding 330,000 returns for tax year 2011 (SOI Division 2013). Part of the growth was due to a five-fold increase in the minimum sampling rate when the SOI Division decided to include returns that were being captured for an OTA panel and, therefore, were available at minimal additional cost. The residual increase was due to growth in the upper tail of the income distribution and in certain specialized strata.

Even in the absence of these substantial changes in the size and composition of the SOI Individual sample, a review of the sample design could be considered overdue. After more than 25 years the needs of the sample's customers may have changed, and the characteristics of the filing population may have evolved in ways that diminish the effectiveness and efficiency of the design. Furthermore, new technology and other factors may have altered the cost of processing the sample. With the changes we have noted, the SOI Individual sample no longer conforms to the 1980s redesign, although stratum definitions and many of the sampling rates remain the same. It is appropriate to ask whether the current sample meets the needs of its users as fully as it could and, even if it does, whether in the current climate of reduced agency budgets but growing demands, a smaller sample, perhaps configured differently, could meet these needs just as well or even better while saving resources that could be used more productively elsewhere.

In preparing this report, Mathematica reviewed the design of the current Individual sample and the principal uses of the Individual sample data. Mathematica also met with the major customers of the Individual tax data to discuss their uses of the sample data and provide them an opportunity to comment on particular elements of the sample design and the products that are created from the data. In conducting this study our objective was not to develop a new design but, rather, to identify areas where improvements to the current sample may be possible and desirable. The report develops recommendations for improving the design of the sample, consistent with the current needs of major users, good statistical practice, and budgetary considerations.

The report is organized as follows. Chapter II provides an overview of the current sample design and how it has evolved since implementation. Chapter III identifies the issues that we proposed or were asked to consider in reviewing the Individual sample. Chapter IV discusses our findings, including the results of empirical analyses, and Chapter V presents our recommendations.

II. OVERVIEW OF THE CURRENT SAMPLE DESIGN

This overview of the SOI Individual tax return sample design covers stratification, sample selection, and the evolution of the sample over time.

A. Stratification

The SOI Individual tax return sample is stratified by income and form type. Some of the income strata are substratified by relevance to tax policy modeling—one of if not *the* principal use of the Individual sample data. In addition, there are two specialized strata that are assigned separately from the rest. In describing the stratification we discuss, in turn, the definition of income, indexing, the degree of interest to policy analysts, form type, and the two specialized strata.

1. Definition of Income

The measure of income that represents the principal stratifier is the larger of gross positive income and gross negative income. The component amounts include items drawn from the 1040 and items drawn from the supplemental schedules. The income concept is broader than just the total income included in Adjusted Gross Income (AGI). Component items were selected to ensure that returns reporting large amounts—positive or negative—are selected at high rates, as such returns have more utility for tax policy analysis, potentially, than returns with the same AGI but no additional receipts, untaxed or with offsetting expenses.

Gross positive income is the sum of 12 items that can only assume positive values and 11 items that can be positive or negative and are included only if they are positive (Table II.1).

The first eight items come from the income section of Form 1040, but they may include amounts that are not counted in AGI. One of the items is entirely free from taxation (Tested Tax Exempt Interest), so it is not counted in AGI at all. Three of the items (Tested Pension Amount, Tested Unemployment Compensation, and Tested Social Security) may include untaxed portions. For these three items only the taxable portion is included in AGI. The taxable portions are reported in separate fields on the 1040. Fields that include untaxed portions are generally not subject to the

same level of scrutiny during master file processing as fields that capture amounts that are fully taxable. During sample selection, the SOI Division applies tests to these four fields to determine if their values are excessive or inappropriately zero, which could indicate a processing or reporting error. Amounts that fail are replaced so that they do not have an undue effect on gross income in either direction, which could result in an incorrect stratum assignment and too high or low a selection probability and, if selected, too low or high a weight.

The next four fields represent gains reported on Schedule E. Expenses and other deductions are applied to determine the taxable amount that is reported on Form 1040. The first two of these fields are not tested during master file processing, and neither are they tested during SOI sample selection.

The 11 business income and net income items that round out the positive amounts are all tested in master file processing. The first four of the net items appear in the income section of the 1040 as components of AGI. The remaining net items are taken from Schedule D, and the five business income items are taken from Schedules C and F.[1]

Gross negative income is the sum of 7 items that include only losses, 11 items that can be positive or negative but are included only if they are negative, and 2 deduction or expense items (Table II.2). The negatively signed items are included as absolute values, so that the overall sum of the 20 items is positive. A "negative income adjustment" that incorporates corrections to the profit/loss amounts reported on schedules C, E, and F is subtracted from the sum of the 20 items and can produce a negative result. If it does so, gross negative income is set to zero.

With one exception (Alimony Paid), the first seven loss items are taken from supplemental schedules. One of these items, Total Expenses All Property Amount, is not tested during master file

[1] A separate Schedule C is required for each non-farm business that the taxpayer owns. Similarly, a separate Schedule F is required for each farm business that the taxpayer owns.

processing. The next 11 items are counterparts to the business income and net income items that contribute to the positive amounts total when they are positive. They are counted in the negative amounts total—but as absolute values—if they are negative. The Total Deductions amount is summed across multiple Schedules C, and the Total Farm Expenses amount is summed across multiple Schedules F.

2. Indexing

Since 1996, the income stratifier has been indexed, using a price index for Gross Domestic Product. The index converts the dollar values of the income stratifier to a base of 1991 dollars, reflecting the initial implementation of the design. In this way the income stratum boundaries can remain fixed at their 1991 values, shown in the first column of Table II.3. Indexing the income stratifier helps to offset the effects of rising income over time on the distribution of the population by income class.

3. Degree of Interest

When the current sample design was developed, there was an explicit focus on the usefulness of returns for policy analyses. Within the same income class, returns were considered more useful (or "interesting") if less common income sources or deductions were prominent. The design assigns returns a degree of interest ranging from 1 to 4, where 1 is lowest and 4 is highest. The definition of interesting is complex, using more than 30 variables and a variety of ratios, so it is not recounted here. For example, returns are least interesting if a few common sources account for nearly all of their income. The degree of interest, as it is called, is used to sub-stratify returns with gross positive income less than $250,000 as shown in Table II.3.

4. Form Type

The second major dimension of stratification is form type. The sample design distinguishes among seven classes of returns based on their attachment of particular combinations of Form 2555

(Foreign Earned Income), Form 1116 (Computation of Foreign Tax Credit), Schedule C (Profit or Loss from Business or Profession), and Schedule F (Farm Income and Expenses):

1. Form 2555
2. No Form 2555 but Form 1116 and either Schedule C or Schedule F
3. Form 1116 but no Form 2555, Schedule C, or Schedule F
4. Schedule C and Schedule F but no Form 2555 or Form 1116
5. Schedule C but no Schedule F, Form 2555, or Form 1116
6. Schedule F but no Schedule C, Form 2555, or Form 1116
7. No Schedule C, Schedule F, Form 2555, or Form 1116

Returns filed on Form 1040A or 1040EZ, which cannot include any of the four forms or schedules, are placed in category 7.

Every five years (those ending in 1 or 6), the SOI Division uses its returns with Form 2555 and Form 1116 to support a foreign study. In the foreign study years returns with Form 2555 are oversampled. There is no need to oversample returns with Form 1116, which is far more common, because the annual sample collects sufficient numbers of such returns to meet the needs of the foreign study. The higher sampling rates for returns with Form 2555 add thousands of additional returns to the sample.

Each year the SOI Division draws a supplemental sample of 2,000 Schedule C returns (form types 4 and 5 and a subset of type 2) to prepare a set of tabulations of sole proprietorship returns for BEA. The tabulations use the full sample plus the supplement. Both OTA and JCT receive copies of the enhanced file.

Among returns with gross positive income less than $30,000, only those with form type 7 can be assigned a degree of interest equal to 1. By virtue of their supplemental schedules and forms, all returns with the other six form types qualify for degrees of interest of 2 or higher. All 7 form types can occur in the remaining 23 combinations of income class and degree of interest, as shown in Table II.3. Given how degrees of interest are grouped by income class, form types 1 through 6 have

23 levels of income by degree of interest while form type 7 has 24. Thus the total number of strata generated by the combination of income class and form type is equal to 23 times 6, plus 24, or 162.

For weighting purposes (that is, post-stratification to population totals), form type is collapsed into four categories by combining classes 1 through 3 (Form 2555 or Form 1116) and classes 4 and 5 (Schedule C with or without Schedule F and without Forms 2555 and 1116). The combination of income class, degree of interest, and form type yields 93 weighting classes (23 times 4, plus 1). In foreign study years, form class 1 is separated from classes 2 and 3 because it is sampled at a higher rate. The number of weighting classes is increased by 23 to 116.

5. Specialized Strata

In addition to the 162 strata generated by the cross-classification of form type and the combination of income class and degree of interest, there are two special strata that take precedence over all other strata. That is, all returns that qualify for these strata are assigned to them before they are considered for any other strata. Both of the specialized strata are sampled with certainty.

High-income nontaxable returns have AGI or expanded income of at least $200,000 in current dollars but an income tax liability—including the Alternative Minimum Tax (AMT)—that is zero after subtracting all credits.[2] The SOI Division collects such returns to comply with a legislative requirement, established by the Tax Reform Act of 1976, to report on such returns annually. The income threshold is not indexed, so the number of returns in this stratum has grown from 69 in 1976 when this stratum first appeared, to 2,757 in 1991 when the current sample design was implemented, to over 30,000 currently.

[2] Expanded income adds the following to AGI: tax-exempt interest, nontaxable Social Security benefits, the foreign-earned income exclusion, and tax preference items used to calculate the AMT. Expanded income subtracts unreimbursed employee business expenses, moving expenses, investment interest expense up to the value of investment income, and miscellaneous itemized deductions below the 2 percent of AGI floor.

Returns with gross business (Schedule C) receipts of $50 million or more are also selected with certainty. Historically, these returns have been few in number. They barely exceeded 300 in the 2011 sample, for instance.

B. Sample Selection

In summarizing how the SOI sample is selected, we discuss the universe from which the sample is selected, show recent sampling rates, and describe the method of selection.

1. The SOI Universe

The universe from which the SOI Individual tax sample is drawn consists of returns filed and processed during a given calendar year, excluding tentative and amended returns and returns from prior tax years too far removed. Most of the returns filed during a given calendar year, say 2012, have filing periods—or tax years—coinciding with the preceding calendar year, in this case 2011. The SOI sample from a given calendar year is identified by the prior tax year. A small percentage of the returns processed during a given calendar year have non-calendar year filing periods, most of which end in the current year or preceding year, or calendar year filing periods for tax years prior to the immediately preceding year (that is, prior to 2011 in this example). Returns from prior tax years are included in the sample to represent returns that were due to be filed during the year from which the sample was drawn but were not filed or processed in time (see below).

Tentative returns are excluded from the universe because revised, more complete returns may have been filed and sampled later. Amended returns are excluded because the original returns were already subject to sampling. Most tentative and amended returns can be identified prior to sampling so that they can be excluded from the possibility of selection, but some cannot be identified prior to sampling and are selected into the sample. These returns are retained in the sample because the return totals that are used to weight the sample include all tentative and amended returns that could not be identified prior to selection. The returns that ended up in the sample unintentionally are

included in the sample counts for weighting purposes, but they are assigned special "Reject" codes so that they can be excluded from tabulations and other analyses.[3]

2. Sampling Rates

Sampling rates vary from a minimum of about 1 in 1,000 (0.10 percent) to a maximum of 100 percent. That is, the highest sampling rate is 1,000 times the lowest. Table II.4 shows current sampling rates by income class and degree of interest for years that do not include a foreign study. In 2010, 92 percent of the filing population was sampled at the minimum 0.10 percent rate. In Section C we discuss how the sampling rates have changed since the design was implemented.

3. Method of Selection

Returns are selected into the SOI sample on the basis of the first or "primary" taxpayer's social security number (SSN), which is used in two different ways. The first method uses the last four digits of the SSN, which the Social Security Administration (SSA) assigns approximately randomly.[4] These digits have a range from 1 to 9999 (the sequence 0000 is not used), and so the primary SSNs with any one sequence of final four digits represent a random 1 in 9,999 sample of the entire universe of primary filers. The IRS has designated 10 sets of ending digits for selection into the SOI sample. These 10 sets of digits, which have their origin in SSA's Continuous Work History Sample (CWHS), yield a 10 in 9,999 sample of returns from the SOI universe. A filer with a CWHS SSN will always be selected into the SOI sample as long as he or she is the primary filer (as opposed to a "secondary" filer, or spouse, on a joint return).

The second method of selection uses a transformation of the primary SSN to generate a five-digit random number. Within each stratum the transformed value, or "transform," is compared to a

[3] Returns that contain no income information are also assigned Reject codes that differentiate them from other returns. They, too, are excluded from published tables.

[4] Beginning June 25, 2011 the Social Security Administration (SSA) has changed the structure of the SSN and the way that SSNs are assigned. All nine digits are now assigned randomly.

"sample number," which is a function of the sampling rate in that stratum. If the transform is less than or equal to the sample number, then the return is selected into the sample. The sample number takes into account the probability that a return might have been selected because of a CWHS SSN. The sample number for stratum *j* is calculated from the following:

$$S_j = 100000*(P_j - .001 + P_j*.001) - 1$$

where S_j is the sample number and P_j is the sampling rate, expressed as a proportion, for stratum *j*, and .001 is the probability that a return carries one of the 10 CWHS SSNs. For example, to select a 10 percent sample of a given stratum, the IRS would select all returns with transforms less than or equal to a sample number of 9,910.

The transformation is uniform from year to year so that a given SSN always yields the same transformed value. This implies that a particular taxpayer's return, if selected into the SOI sample in one year, will continue to be selected for as long as that taxpayer keeps filing in the primary position and remains subject to the same or higher selection probability. If a taxpayer is not already in a stratum with the lowest probability of selection, a reduction in income or a change in return type may result in a lower probability of selection (a reduction in sampling rates could produce the same result). In that case, the return will remain in the sample only if its SSN transform is below the new sample cut-off. If 100 returns fall from one stratum to another stratum with a selection probability only half as large as the first stratum, then we would expect that about half of the returns would still be selected from their new stratum while the other half would not be selected.

Finally, because of field size limitations on the Individual Master File, returns with exceedingly large amounts may be rejected and, therefore, not be available for sample selection through the mechanisms described above. Such returns have to be identified and added manually because they

bypass sample selection.[5] There were 32 such records in the 2008 file but only 4 in the 2009 file, implying that the IRS had expanded the fields sufficiently to prevent most of these omissions.

C. Evolution of the Sample over Time

The Individual sample has changed substantially in its size and composition since the implementation of the late 1980s redesign beginning with the 1991 tax year. These changes reflect a combination of several factors, including population growth, change in the composition of the filing population with respect to the sample strata, and changes in the sampling rates by strata. Figure II.1 plots the growth of the sample from 125,926 returns in 1991 to 333,106 returns in 2011 (both foreign study years). Over this period the number of returns filed grew only modestly, from 115.4 million in 1991 to 145.6 million in 2011, showing a mostly smooth trend except for an uptick in 2007 due to more than 10 million returns that were filed solely to collect an economic stimulus payment.

The growth in the Individual sample was anything but smooth over this period. Increases between 1997 and 1998 and between 2004 and 2005 were attributable in large part to expansions of the CWHS portion of the sample, from 2 to 5 endings in 1998 and from 5 to 10 endings in 2005. Together, these two additions raised the minimum sampling rate five-fold from .02 percent to .10 percent (Table II.4). The 1998 increase boosted sample sizes in just three strata, but these strata represented the largest populations. The 2005 change produced even larger sample size increases in these three strata while expanding the samples in six other strata below $250,000.[6]

[5] Because they bypass sample selection, the positive and negative gross income fields and the SSN transform are not calculated. A selection amount and an indicator as to whether it is positive or negative are assigned later and identified in the microdata file as computed fields.

[6] With the expansion to 10 CWHS endings, there would have been no differentiation in sampling rates by degree of interest at incomes below $120,000 (in 1991 dollars), but the sampling rates among more interesting returns in the three income classes were raised to 0.15 percent. In addition, the sampling rates in the two substrata between $120,000 and $250,000 were increased by 0.05 percentage points. In 2009 the rates that were raised to 0.15 percent in 2005 were reduced to 0.10 percent, and the sampling rate among less interesting returns between $120,000 and $250,000 was raised to match that of more interesting returns, eliminating all differential selection by degree of interest.

Most of the remaining sample growth was due to upward movement in the income distribution, which intensified during the economic booms of the late 1990s and mid 2000s. Rising income moved filers into strata with higher sampling rates. Increases in the top two positive income strata, from which returns are selected with certainty, were especially pronounced, but so were the declines that followed the busts in 2000 and 2007 (see Figure II.2).

In addition to the increase in the minimum sampling rate, which now applies to 92 percent of the filing population, there were other changes in sampling rates, although not all of the changes were in the same direction. Sampling rates were increased for the negative income strata between $0 and -$249,999. However, there were reductions in sampling rates among returns with losses between $500,000 and $5 million and among returns with positive income between $250,000 and $5 million.

Table II.5 decomposes the growth in the Individual sample from the implementation of the current design in tax year 1991 through 2011. Had there not been a foreign study in 1991, the Individual sample would have included 113,931 returns. Growth in the filing population between 1991 and 2011 would have added about 30,000 returns absent the changes in sampling rates and the composition of the filing population. Applying the rate changes to the expanded population but with fixed composition would have added another 77,000 returns. Additional increases in the strata sampled with certainty (their rates did not change over time) accounted for another 51,000 returns. The remaining shift in the distribution of returns by stratum added 46,000 returns to bring the 2011 total to 317,356 returns had there been no foreign study.

With the growth in the size of the sample over time the precision of sample estimates has improved considerably. Table II.6 presents coefficients of variation (CVs) for estimates of the aggregate number of returns for a selection of items, and Table II.7 presents the CVs for estimates

of aggregate amounts, at five-year intervals from 1996 through 2011.[7] Common items like AGI and salaries and wages show reductions around 40 percent. For example, the CV of AGI less deficit fell from 0.16 percent to 0.09 percent, a 44 percent reduction, while the CV for returns with salaries and wages declined from 0.18 to 0.11 percent (a 37 percent reduction), and the CV for the aggregate amount of salaries and wages dropped from 0.28 to 0.17 percent (a 38 percent reduction). Some of the capital gains items show reductions as high as two-thirds (see, for example, the Schedule D short-term loss carryover for both returns and amounts).

Some of the very large improvements in precision are explained in part by growth in the items' frequency in the population, which compounds the effects of a higher sampling rate. For example, the number of returns with Schedule D losses from other forms increased more than three times between 1996 and 2011, and its CV dropped from above 10 percent to less than 4 percent. The CV for the corresponding amount, however, showed a decline comparable to AGI.

Estimates of two items—returns with capital gain distributions on Form 1040 and the amount of net loss from other income—did become less precise over time. In both cases there were small reductions in the number of returns in the population with these items, which may have contributed to the declines in precision. For the rest of the items, however, a notable consequence of the growth in size and change in composition of the Individual sample has been a substantial improvement in the precision of sample estimates. In evaluating possible changes to the current design, an important question to answer is what will the design changes do to the precision of key sample estimates. While customers will focus on changes relative to the current precision of sample estimates, we need to keep in mind that the precision associated with the current sample is not entirely by design.

[7] The SOI Division did not publish CVs for 1991, the first year of the new design. However, the 1996 sample was only 500 returns larger than the 1991 sample, so the precision of sample estimates in the two years should be comparable. In addition, like 1991, all four years displayed in the table were foreign study years, so the samples are generally larger than those in surrounding years.

Figure II.2. Returns with Positive Total Income Sampled with Certainty

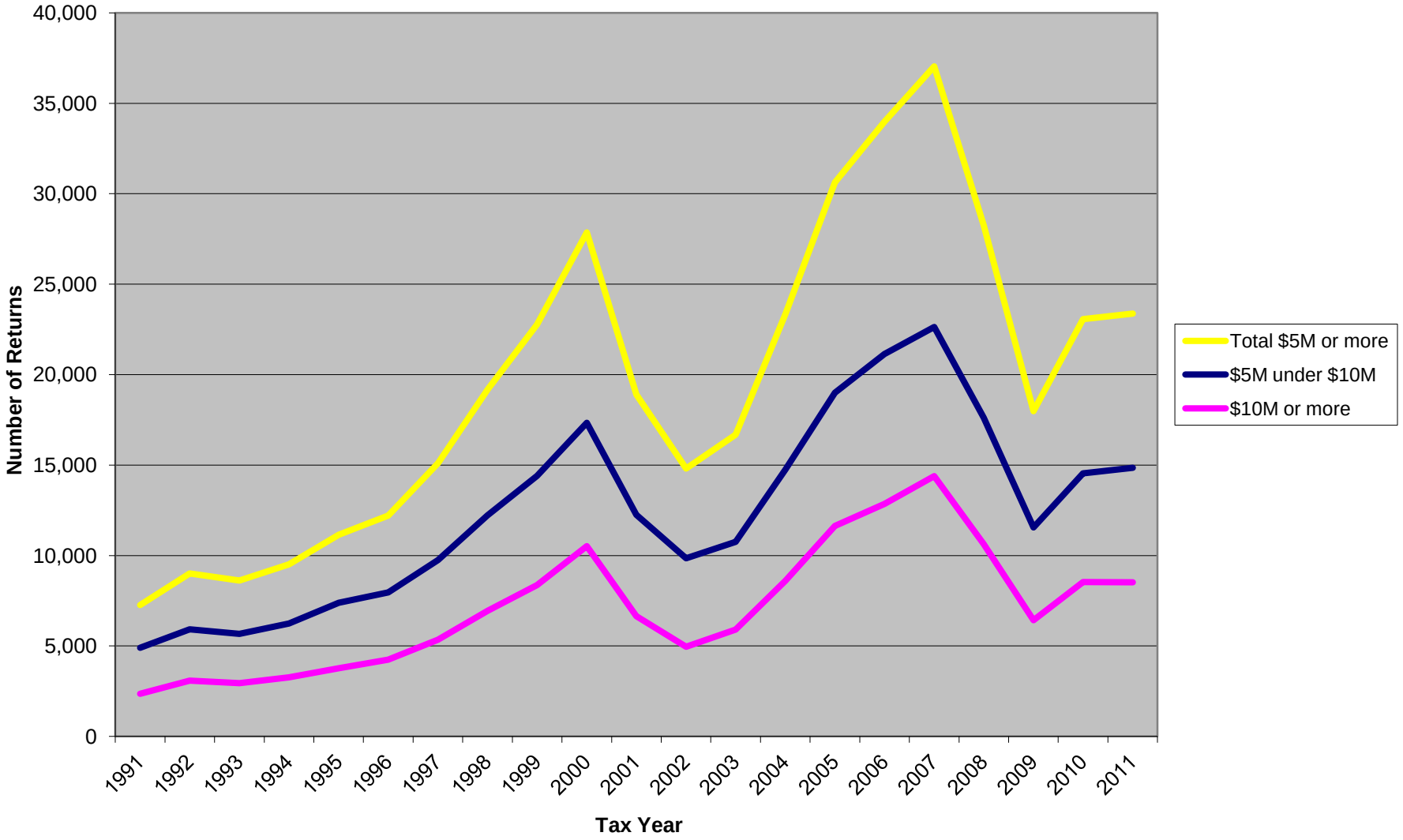

Figure II.1. Number of Returns Sampled and 1,000s of Returns Filed

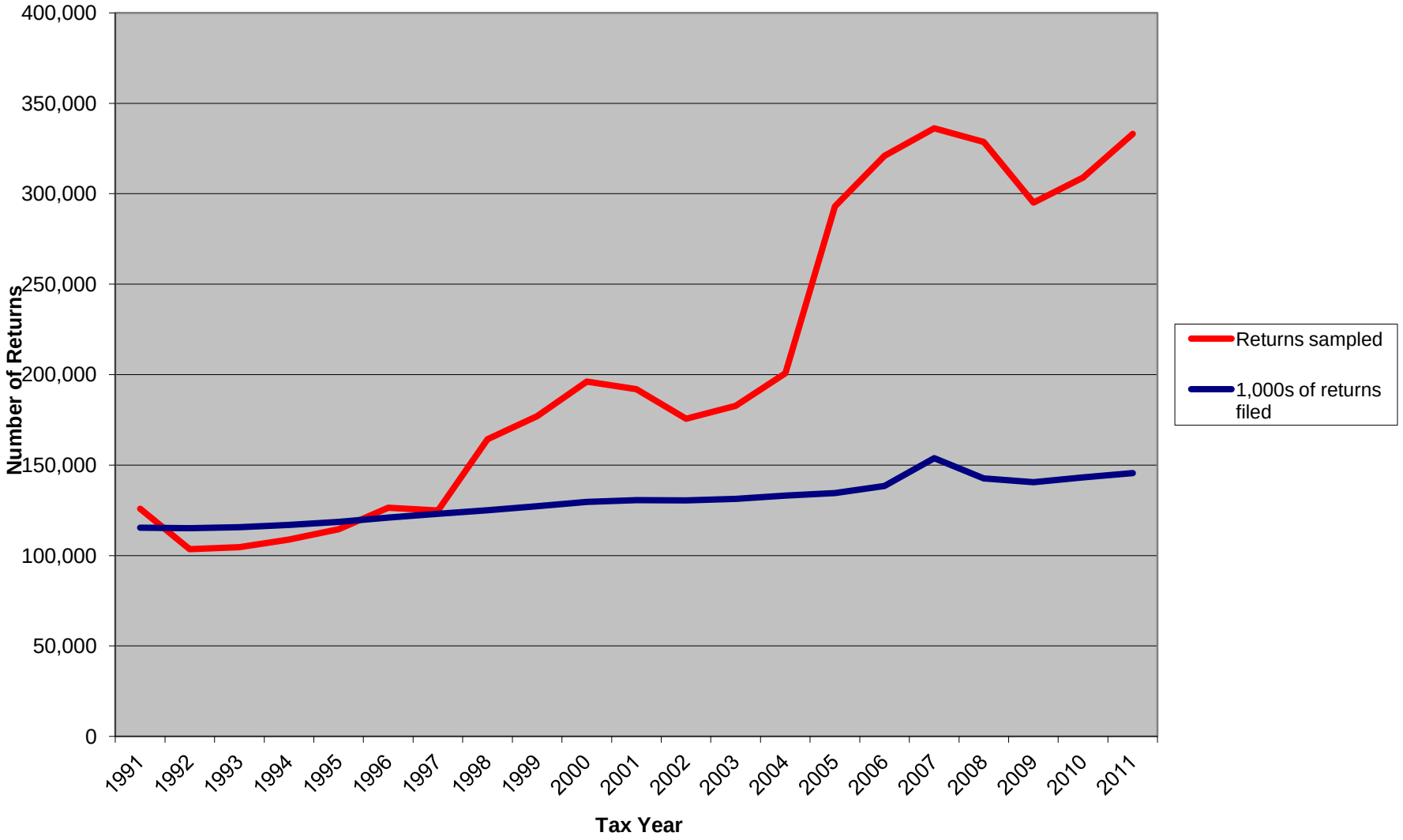

Table II.1. Items Contributing to Gross Positive Income

Item	Source
Strictly Positive Items	
1. Wage Amount	Form 1040
2. Tested Tax Exempt Interest	Form 1040
3. Taxable Dividends	Form 1040
4. Alimony Received	Form 1040
5. Tested Pension Amount	Form 1040
6. Taxable IRA Distribution	Form 1040
7. Tested Unemployment Compensation	Form 1040
8. Tested Social Security	Form 1040
Strictly Gain Items	
1. Total Rental Payments Amount	Schedule E
2. Total Royalty Payments Amount	Schedule E
3. Partnership, S-Corporation Income	Schedule E
4. Estate and Trust Income	Schedule E
Business Income Items (if positive)	
1. Schedule C-1 Gross Profit/Loss	Schedule C
2. Schedule C-2 Gross Profit/Loss	Schedule C
3. Schedule C-3 Gross Profit/Loss	Schedule C
4. Schedule F-1 Gross Profit/Loss	Schedule F
5. Schedule F-2 Gross Profit/Loss	Schedule F
Net Items	
1. Supplemental Gains/Losses	Form 1040
2. Other Income Amount	Form 1040
3. Rarm/Rent Income/Loss	Form 1040
4. Taxable Interest Income	Form 1040
5. Net Short Term Gain/Loss Amount	Schedule D
6. Net Long Term Gain/Loss Amount	Schedule D

Source: Statistics of Income Division.

Table II.2. Items Contributing to Gross Negative Income

Item	Source
Loss Items	
1. Partnership, S Corporation Loss	Schedule E
2. Estate and Trust Loss	Schedule E
3. Total Expenses All Property Amount	Schedule E
4. Total Depreciation All Property Amount	Schedule E
5. Alimony Paid	Form 1040
6. Form 3903 Moving Expense Amount	Form 3903
7. Business at Home Expense	Schedule C
Business Loss Items (include absolute value if negative)	
1. Schedule C-1 Gross Profit/Loss	Schedule C
2. Schedule C-2 Gross Profit/Loss	Schedule C
3. Schedule C-3 Gross Profit/Loss	Schedule C
4. Schedule F-1 Gross Profit/Loss	Schedule F
5. Schedule F-2 Gross Profit/Loss	Schedule F
Net Items (include absolute value if negative)	
1. Supplemental Gains/Losses	Form 1040
2. Other Income Amount	Form 1040
3. Rarm/Rent Income/Loss	Form 1040
4. Taxable Interest Income	Form 1040
5. Net Short Term Gain/Loss Amount	Schedule D
6. Net Long Term Gain/Loss Amount	Schedule D
Deduction Items	
1. Total Deductions	Schedule C
2. Total Farm Expenses	Schedule F
Adjustment Item (subtract from total of above items)	
1. Negative Income Adjustment	

Source: Statistics of Income Division.

Table II.3. Current Sampling Rates for the Individual Tax Sample by Income Class and Degree of Interest

Income Class	Degree of Interest	Form Type Strata
High-income nontaxable	All	N/A
High Schedule C receipts	All	N/A
-$10,000,000 or less	All	1-7
-$9,999,999 to -$5,000,000	All	1-7
-$4,999,999 to -$2,000,000	All	1-7
-$1,999,999 to -$1,000,000	All	1-7
-$999,999 to -$500,000	All	1-7
-$499,999 to -$250,000	All	1-7
-$249,999 to -$120,000	All	1-7
-$119,999 to -$60,000	All	1-7
-$59,999 to -$1	All	1-7
$0 to under $30,000	1	7 only
$0 to under $30,000	2	1-7
$0 to under $30,000	3-4	1-7
$30,000 to under $60,000	1-2	1-7
$30,000 to under $60,000	3-4	1-7
$60,000 to under $120,000	1-3	1-7
$60,000 to under $120,000	4	1-7
$120,000 to under $250,000	1-3	1-7
$120,000 to under $250,000	4	1-7
$250,000 to under $500,000	All	1-7
$500,000 to under $1,000,000	All	1-7
$1,000,000 to under $2,000,000	All	1-7
$2,000,000 to under $5,000,000	All	1-7
$5,000,000 to under $10,000,000	All	1-7
$10,000,000 or more	All	1-7

Table II.4. Sampling Rates for the Individual Tax Sample by Income Class and Degree of Interest: 1991 and Currently

Income Class	Degree of Interest	1991	Current
High-income nontaxable		100.00	100.00
High Schedule C receipts		100.00	100.00
-$10,000,000 or less	All	100.00	100.00
-$9,999,999 to -$5,000,000	All	100.00	100.00
-$4,999,999 to -$2,000,000	All	50.09	34.07
-$1,999,999 to -$1,000,000	All	15.36	16.08
-$999,999 to -$500,000	All	4.00	3.41
-$499,999 to -$250,000	All	1.02	0.99
-$249,999 to -$120,000	All	0.42	0.51
-$119,999 to -$60,000	All	0.16	0.31
-$59,999 to -$1	All	0.10	0.19
$0 to under $30,000	1	0.02	0.10
$0 to under $30,000	2	0.02	0.10
$0 to under $30,000	3-4	0.08	0.10
$30,000 to under $60,000	1-2	0.03	0.10
$30,000 to under $60,000	3-4	0.10	0.10
$60,000 to under $120,000	1-3	0.06	0.10
$60,000 to under $120,000	4	0.15	0.10
$120,000 to under $250,000	1-3	0.20	0.33
$120,000 to under $250,000	4	0.40	0.33
$250,000 to under $500,000	All	1.01	0.72
$500,000 to under $1,000,000	All	4.03	2.48
$1,000,000 to under $2,000,000	All	15.98	12.19
$2,000,000 to under $5,000,000	All	49.70	32.47
$5,000,000 to under $10,000,000	All	100.00	100.00
$10,000,000 or more	All	100.00	100.00

Note: The 1991 rates were calculated from population and sample counts. As 1991 was a foreign study year, the rates shown above were obtained from all non-foreign returns (that is, those with no Form 2555 or Form 1116). The current rates were derived from sampling specifications.

Table II.5. Decomposition of Growth in the Individual Sample, 1991 to 2011, by Income Class and Degree of Interest

Income Class	Degree of Interest	1991 Sample Absent a Foreign Study	2011 Sample with Population Growth Alone	2011 Sample Adding Changes in Rates	2011 Sample Adding Growth in Certainty Strata	2011 Sample Absent a Foreign Study
Total returns		113,931	143,700	220,439	271,555	317,356
High-income nontaxable		2,757	3,477	3,477	34,663	34,663
High Schedule C receipts		46	58	58	305	305
-$10,000,000 or less	All	1,126	1,420	1,420	3,237	3,237
-$9,999,999 to -$5,000,000	All	1,480	1,867	1,867	5,512	5,512
-$4,999,999 to -$2,000,000	All	2,682	3,383	2,301	2,301	7,398
-$1,999,999 to -$1,000,000	All	1,591	2,007	2,101	2,101	7,111
-$999,999 to -$500,000	All	1,028	1,297	1,106	1,106	3,536
-$499,999 to -$250,000	All	630	795	772	772	2,254
-$249,999 to -$120,000	All	553	697	846	846	2,352
-$119,999 to -$60,000	All	308	388	752	752	1,787
-$59,999 to -$1	All	644	812	1,543	1,543	2,554
$0 to under $30,000	1	5,289	6,671	33,355	33,355	32,074
$0 to under $30,000	2	8,167	10,301	51,505	51,505	33,528
$0 to under $30,000	3-4	7,617	9,607	12,009	12,009	12,790
$30,000 to under $60,000	1-2	6,295	7,940	26,467	26,467	23,933
$30,000 to under $60,000	3-4	7,780	9,813	9,813	9,813	11,358
$60,000 to under $120,000	1-3	5,434	6,854	11,423	11,423	14,106
$60,000 to under $120,000	4	5,080	6,407	4,271	4,271	6,435
$120,000 to under $250,000	1-3	2,873	3,624	5,980	5,980	6,133
$120,000 to under $250,000	4	5,954	7,510	6,196	6,196	14,531
$250,000 to under $500,000	All	7,889	9,950	7,093	7,093	12,097
$500,000 to under $1,000,000	All	9,219	11,628	7,156	7,156	13,477
$1,000,000 to under $2,000,000	All	10,529	13,280	10,130	10,130	21,185
$2,000,000 to under $5,000,000	All	11,701	14,758	9,642	9,642	21,623
$5,000,000 to under $10,000,000	All	4,907	6,189	6,189	14,857	14,857
$10,000,000 or more	All	2,352	2,967	2,967	8,520	8,520

Table II.6. Estimated Coefficients of Variation (Percent) for Aggregate Numbers of Returns: Selected Items, 1996 to 2011

Item	Description	1996	2001	2006	2011
N1	Number of returns	0.04 [a]	0.02	0.01	0.01
n00200	Salaries and wages	0.18	0.13	0.10	0.11
n00300	Taxable interest	0.39	0.31	0.24	0.30
n00600	Ordinary dividends	0.73	0.53	0.39	0.44
n00700	State income tax refunds	0.95	0.70	0.50	0.53
n00900p	Business or profession net income	0.60	0.48	0.35	0.36
n00900n	Business or profession net loss	1.74	1.42	0.96	1.06
n01100	Capital gain distributions on Form 1040	2.08	2.54	1.31	2.14
n01000p	D Taxable net gain	1.08	0.92	0.62	0.90
n01000n	D Taxable net loss	1.88	0.99	0.82	0.70
n21800	D Short-term loss carryover	3.86	2.19	1.55	1.24
n21620p	D Net short-term gain from other forms	8.96	7.77	5.03	3.79
n21620n	D Net short-term loss from other forms	10.08	9.51	4.60	3.77
n22390	D Net long-term loss carryover	2.66	2.11	0.99	0.85
n22320p	D Net long-term gain from other forms	2.04	1.91	1.38	1.71
n22320n	D Net long-term loss from other forms	10.48	10.54	5.53	4.04
n01200p	Sale of property other than capital assets, gain	3.50	3.06	2.21	2.44
n01200n	Sale of property other than capital assets, loss	3.48	3.17	2.36	2.25
n01400	Taxable IRA distributions	1.84	1.22	0.83	0.76
n01700	Pensions and annuities, taxable	0.93	0.70	0.49	0.49
n27310p	Total rental and royalty, net income	1.47	1.29	1.00	1.07
n27310n	Total rental and royalty, net loss	1.75	1.59	1.14	1.23
n26270p	Partnership and S-corp, net income	1.71	1.35	0.94	1.06
n26270n	Partnership and S-corp, net loss	2.56	2.16	1.46	1.51
n26500p	Estate and trust, net income	5.18	4.35	3.16	3.53
n26500n	Estate and trust, net loss	14.72	15.08	10.57	9.89
n02100p	Farm net income	3.79	3.46	2.90	2.93
n02100n	Farm net loss	2.31	1.96	1.45	1.68
n02300	Unemployment compensation	1.74	1.33	1.05	0.79
n02500	Social Security benefits, taxable	1.46	1.02	0.65	0.63
n02600p	Other income, net income	2.01	1.56	1.08	1.13
n02600n	Other income, net loss	7.42	5.49	4.32	4.35
n02900	Statutory adjustments, total	0.74	0.58	0.36	0.37
n03260	Deduction for one-half of self-employment tax	0.70	0.55	0.39	0.41
n03270	Self-employed health insurance deduction	1.78	1.45	1.10	1.25
n04470	Total itemized deductions	0.54	0.38	0.26	0.30
N2	Exemptions	0.28	0.22	0.15	0.16
n04800	Taxable income	0.24	0.17	0.12	0.13
n05800	Income tax before credits	0.24	0.17	0.12	0.13

Source: Statistics of Income Division, Individual Returns Complete Report, various years.

[a] We are not sure what accounts for the high estimated CV for the overall number of returns relative to even five years later, when the sample size was only 52 percent larger, but we have confirmed it with estimates from surrounding years.

Table II.7. Estimated Coefficients of Variation (Percent) for Aggregate Amounts: Selected Items, 1996 to 2011

Item	Description	1996	2001	2006	2011
E00100	Adjusted gross income less deficit	0.16	0.11	0.09	0.09
E00200	Salaries and wages	0.28	0.21	0.16	0.17
E00300	Taxable interest	1.18	0.97	0.62	0.85
E00600	Ordinary dividends	1.31	1.08	0.65	0.75
E00700	State income tax refunds	1.47	0.92	0.70	0.74
e00900p	Business or profession net income	1.13	0.97	1.18	0.75
e00900n	Business or profession net loss	2.26	1.91	1.42	1.49
E01100	Capital gain distributions on Form 1040	5.44	21.63 [a]	3.63	6.12
e01000p	D Taxable net gain	0.83	0.59	0.36	0.53
e01000n	D Taxable net loss	2.02	1.06	0.88	0.74
E21800	D Short-term loss carryover	2.63	1.57	1.23	0.84
e21620p	D Net short-term gain from other forms	7.51	4.24	2.51	2.70
e21620n	D Net short-term loss from other forms	6.99	6.31	4.75	4.09
E22390	D Net long-term loss carryover	1.91	1.75	0.90	0.65
e22320p	D Net long-term gain from other forms	2.37	1.39	0.81	1.11
e22320n	D Net long-term loss from other forms	6.91	7.00	5.98	4.56
e01200p	Sale of property other than capital assets, gain	4.57	4.03	2.50	2.19
e01200n	Sale of property other than capital assets, loss	4.52	3.43	3.37	2.45
E01400	Taxable IRA distributions	3.06	1.22	1.37	1.11
E01700	Pensions and annuities, taxable	1.36	1.06	0.74	0.73
e27310p	Total rental and royalty, net income	1.63	1.47	1.27	1.40
e27310n	Total rental and royalty, net loss	2.08	1.96	1.41	1.59
e26270p	Partnership and S-corp, net income	1.04	0.82	0.55	0.63
e26270n	Partnership and S-corp, net loss	1.84	1.30	1.02	0.97
e26500p	Estate and trust, net income	4.55	3.36	2.64	2.92
e26500n	Estate and trust, net loss	7.66	3.12	4.35	3.77
e02100p	Farm net income	4.54	4.16	3.80	2.77
e02100n	Farm net loss	3.03	2.73	2.09	2.30
E02300	Unemployment compensation	2.41	1.81	1.42	1.11
E02500	Social Security benefits, taxable	1.75	1.23	0.79	0.77
e02600p	Other income, net income	2.74	2.50	2.09	2.27
e02600n	Other income, net loss	5.59	5.70	5.10	6.53
E02900	Statutory adjustments, total	1.21	0.98	0.65	0.67
E03260	Deduction for one-half of self-employment tax	1.10	0.93	0.71	0.72
E03270	Self-employed health insurance deduction	2.04	1.61	1.21	1.33
E04470	Total itemized deductions	0.55	0.38	0.27	0.30
E04600	Exemptions	0.28	0.23	0.16	0.16
E04800	Taxable income	0.21	0.15	0.12	0.12
E05800	Income tax before credits	0.23	0.17	0.16	0.14

Source: Statistics of Income Division, Individual Returns Complete Report, various years.

[a] Individual records with exceedingly large dollar amounts and large weights can produce inflated CVs. We speculate that this is what may have happened here.

III. ISSUES ADDRESSED

Our review of the Individual sample design encompasses not only the major elements of the design itself but the products and documentation that are generated for the customers.

A. Design Elements

The following elements of the Individual sample design are examined in this review:

- Use of positive and negative income to define the primary stratifying variable
- Sub-stratification by a measure of usefulness for policy analysis
- Indexing of the income stratifier
- Designation of "special focus returns" as distinct strata or groups of strata
- Secondary stratification by form type
- Sampling rates
- Possible sub-stratification by filing mode (electronic versus paper)
- Panel aspects of the design
- Use of prior year returns to represent late filers
- Handling of returns sampled from an incorrect income class
- Handling of "missing" returns—that is, returns that could not be accessed for editing

Issues to consider in reviewing each of these design elements are discussed below.

1. Stratification by Income

Positive and negative total income are calculated separately as the sum of several components—some of them taxable and some of them not. Using gross rather than net income as the basis for stratification has a long history at SOI. The rationale is that returns with large gross incomes make important contributions to the aggregates of positive and negative amounts. Selecting the sample on net income would weaken the estimates by pushing returns with large gains or losses into strata with lower sampling rates.

A drawback of the present strategy is that it uses a number of nontaxable amounts to calculate gross positive or negative income. Historically, nontaxable items are at greater risk of being recorded incorrectly in the master file data used for sample selection. The trade-offs need to be weighed.

2. Sub-stratification by Interestingness

The interesting return indicator (IRI) was originally designed to allow more complex, lower income returns to be sampled at higher rates than simpler returns, but the expansion of the CWHS sample has boosted the minimum selection rate beyond the rates used for the most interesting returns below $60,000, and SOI has eliminated differential sampling based on the IRI at higher income levels. The IRI continues to be used to define separate strata within the same income class, and these are used for post-stratification, but this has a much smaller impact than the use of differential sampling rates.

The principal issue regarding the use of the IRI is whether to retain any type of stratification of this type. As we noted above, the current measure of interestingness is no longer used to sample returns at differential rates. But while the current measure has outlived its usefulness, this does not mean, necessarily, that there is no longer a need for some such measure to differentiate among returns at the same income level that may have different utility for tax policy modeling. This is entirely a question for the principal customers, however, rather than one we can address with empirical analysis.

3. Indexing of Income

As explained in Chapter II, a major source of growth in the SOI sample over time is a substantial increase in the number of returns qualifying to be selected with certainty. Indexing the income stratifier, initiated in 1996, reduces upward movement across strata. (The stratum boundaries are expressed in 1991 dollars, in effect.) At issue is whether the chosen index—a price index based on the GDP—is the most suitable index for income data and whether an alternative index based on income would do more to limit the growth of the sample.

4. Certainty Selection of Special Focus Returns

As noted in Chapter II, high-income nontaxable returns have become a very large stratum. An important question to be resolved is whether the legal mandate to prepare an annual report on these

returns requires that they be included in the SOI sample with certainty. A secondary question is whether a markedly smaller sample of such returns would be sufficient to support the annual report

Returns with total receipts in excess of $50 million are sampled with certainty because they form the upper tail of business returns. In contrast to the high-income nontaxable returns, however, the returns with high total receipts are very few in number. Reducing their number would save very little in resources while potentially weakening the sample.

5. Stratification by Form Type

The form type classification that is used as the secondary stratifier has seven categories, but except in foreign study years each income class is sampled at the same rate across all seven form types. The supplemental BEA sample draws returns from a subset of form types but does not require a form type stratifier to do so. Form type is used to form post-strata for calculating the final sample weights, but even here the seven form types are collapsed to four. The principal issue with regard to form type is whether the full classification continues to be necessary as a stratifying variable.

6. Sampling Rates

Recurring sample surveys generally fix their sample sizes in order to control costs. Sampling rates are adjusted to maintain fixed sample sizes, and since variances are a function of sample size rather than sampling rates, except when the latter exceed 5 or 10 percent, restricting sample growth in this way does not lead to declines in precision over time. The SOI sample is an exception in that the rates in a number of strata are high enough that fixing the sample sizes would reduce precision in these strata. Fixed rates in other strata, however, could be replaced by fixed sample sizes except where the only returns selected are from the CWHS subsample. Assuming that CWHS selection continues, sampling rates in these strata can be adjusted only in increments of 0.01 percent, representing one sequence of final four digits in the SSN. In the highest income strata, from which returns are selected with certainty, the only option for fixing sample size is to revise the stratum

boundaries. To what extent would it be possible to reduce the year-to-year sample growth through a combination of reduced sampling rates and revised stratum boundaries, and could this be done without a year-to-year reduction in precision?

7. Sub-stratification by Filing Mode

Returns filed electronically are much less expensive to process and edit than returns filed on paper. Sampling electronic returns at a higher rate than paper returns could reduce the cost of editing the Individual sample but would make the sample more complex, doubling most of the strata. If electronic returns cannot be substituted for paper returns because of differences in their characteristics, the impact of differential weighting on precision has to be considered as well.

8. Panel Aspects of the Design

The panel aspects of the SOI sample design include the use of 10 CWHS endings and a selection mechanism that utilizes a fixed transform of the primary SSN. Together these ensure that a substantial proportion of the sample in consecutive years will consist of the same filers. As long as a filer with a CWHS SSN continues to file in the primary position, his or her return will continue to be selected into the sample. Persons without CWHS SSNs will continue to be selected as long as they remain in a stratum with the same or higher sampling rate—again, conditional on remaining the primary filer. Those who change from primary to secondary filer will be retained with a probability that corresponds to the unconditional probability of selection in the stratum in which they fall. This will be no lower than 0.1 percent and as high as 100 percent.

The loss of returns because a filer switches from primary to secondary SSN could be addressed in part by extending CWHS selection to the secondary SSN, as is done for the special panel studies. Applying selection with the SSN transform to the secondary SSN would expand this to higher income returns. The loss of returns with a decline in income could be addressed in different ways, but they would involve making the SOI sample more like a true panel. The issue for users is whether the panel aspects of the design as they exist currently are sufficient.

9. Use of Prior Year Returns to Represent Late Filers

A few percent of the returns for a given tax year are filed after the end of the processing year in which they were due to be filed. Most of these late returns are filed during the next calendar year. Rather than delay the Individual sample data by as much as a year to capture these late returns, the SOI Division uses returns from prior years, in effect, as a substitute for the late current year returns. The recent swings in the economy highlight the potential risks in this strategy, but a good solution has not presented itself. Demonstrating whether there is indeed a persistent problem with the long-standing strategy will shed light on whether a better approach is truly in need.

10. Returns Sampled from an Incorrect Income Class

Reporting or transcription errors in the fields used to calculate the income stratifier can result in returns being assigned to and sampled from incorrect strata. If a return is assigned to too high (in absolute dollars) an income class, this does not present a serious problem unless the error is widespread and produces a discernible increase in sample size and editing cost. The latter has not occurred in recent memory, so the returns that are selected from too high an income class are allowed to remain in their assigned strata, where they receive smaller weights than if they had been selected from their correct strata. On the other hand, if a return is assigned to too low an income class and selected into the sample from that stratum, it will receive too high a weight, and this can adversely affect sample estimates if the return carries large dollar amounts. Current SOI practice is to reassign such returns to their correct strata when the displacement is two or more income classes. The population totals used for post-stratification are adjusted to reflect such changes using the weights associated with the new strata. That is, if a return was selected from a stratum with a weight of 1,000 and given a weight of 10 instead, then 10 returns will be moved from the original stratum population total to the new stratum population total. This approach is practical but not especially appealing from a theoretical standpoint, yet a better approach has not been presented. If such errors continue to be rare, there may be little incentive to seek a more satisfying solution.

11. Handling of Missing Returns

Returns that are selected into the SOI sample but cannot be located for editing present a challenge, which is handled in different ways depending on the size of the returns. The missing fields on the more important returns are imputed manually—often by reviewing prior returns by the same filer. Smaller returns may be excluded from the sample. Whether a better approach is needed depends, first, on the magnitude of the problem. If such returns are few in number, an automated and generally more complex approach may not be warranted. On the other hand, if such returns are numerous or growing, then a more automated method of handling such returns may be merited.

B. Individual Sample Products

The SOI Division generates a number of products from the Individual sample. We considered the following products:

- Advance (or preliminary) data
- Final data
- The public use file

We also considered the possible merits of developing products from the population of returns maintained in the Compliance Data Warehouse (CDW). Issues related to each of these are discussed below.

1. Advance Data

Advance estimates from the Individual sample are delivered to OTA in mid-November of each year based on an extrapolation from returns processed through late September. The tabular estimates of key items are followed by a microdata file in December. A 2005 change in the length of an extension that is automatically granted to taxpayers who request it had the effect of pushing these returns beyond the cut-off date for the advance data, thereby increasing the fraction of the complete sample that is excluded from the advance estimates. The methodology for producing the advance estimates has remained unchanged for decades. Returns in the advance sample are weighted to

estimates of the final population, by stratum. This approach entails an implicit assumption that the returns omitted from the advance estimates share the characteristics of those that are excluded. Whether this assumption became less viable in light of the decline in the coverage of the advance sample is a question that needs to be addressed.

2. Final Data

The principal questions to be asked about the final data concern their timeliness and quality, although there do not appear to be concerns about either. Our assessment will be based entirely on our discussions with the principal customers of the Individual sample data.

3. The Public Use File

For customers outside of select federal agencies, the public use file remains the only source of microdata from the Individual sample. While this review is not concerned with the public use file, potential changes to the Individual sample design may affect the size and composition of the public use sample and possibly the quality of the public use data. For this reason some attention to the possible impact of sample changes on the public use file is warranted.

4. Using Returns from the CDW

The CDW is a repository of population-level data on both the Individual and Corporate filing populations. These data include the same return transaction file data that are used to create the SOI Individual and Corporate sample files. The CDW provides a mechanism through which electronic data on the returns filed each year are more readily accessible to Treasury Department and JCT staff than the underlying source data. This has created demand for these data that in some cases replaces reliance on the Individual sample. The CDW data are unedited, however, so they do not incorporate the corrections and enhancements that the SOI Division applies to its annual sample files. Their overall quality, therefore, is demonstrably below that of the Individual sample data. The task for the SOI Division would seem to be one of educating potential users about the strengths and limitations of the CDW data and considering whether there is a potential SOI use for these data.

C. Documentation of the Individual Sample

The principal vehicle for releasing tabulations from the annual SOI Individual sample is a printed volume—also available electronically—called the Complete Report. The Complete Report includes a brief description of the sample, including a table showing population totals and sample counts by stratum. The text on the sample, which runs less than two pages, describes the universe from which the sample is drawn, lists the stratifiers, summarizes the method of selection, discusses data capture and editing, mentions missing returns, and indicates how the weights are calculated. The chapter also explains standard errors and confidence intervals (tables of coefficients of variation are provided in the report). We evaluated the adequacy of the sample documentation in accurately describing the key elements of the design and the rationale for particular features.

IV. FINDINGS

The findings presented in this chapter are based on several sources. These sources include the views gathered from our meetings with the principal customers of the Individual sample data, the results of empirical analyses that we conducted with IRS data, materials provided by SOI or presented in SOI publications, and documentation published annually by SOI, supplemented by conference papers and our own prior work on the current sample design. Findings from these distinct sources are discussed in turn.

A. Views of SOI Customers

To solicit the views of the SOI Division's principal customers for Individual tax data, we interviewed staff in OTA, JCT, BEA, the Congressional Budget Office (CBO), the Brookings/Urban Institute Tax Policy Center (TPC), and the National Bureau of Economic Research (NBER). Both OTA and JCT use the Individual sample microdata heavily in their work, and both have current staff members who participated in the development of the current Individual sample design. CBO can use the restricted microdata file under particular circumstances but relies more heavily on the public use file. The remaining organizations work solely with the public use file and with tabulations. In addition to the published tabulations, BEA receives special tabulations on sole proprietors.

We sought customer input on each of the following elements of the Individual sample design.

- Income stratification
- High-income nontaxable returns
- Indexing
- Income growth—specifically, ways to address this besides indexing
- Interesting returns
- CWHS subsample
- Late filers—specifically how they are handled

We also requested views on each of the following, which are not directly related to the design of the Individual sample:

- Timeliness of the Complete Report data
- Usefulness of the advance data
- Usefulness of CDW data and the value of potential enhancements
- Design of the public use file
- Other needs that are not met adequately by current Individual sample products

All of these topics were communicated prior to meeting with each customer.

Customer views on these topics are discussed below. Some of the comments that were provided in response to these two sets of issues are also reflected in our empirical research, discussed in Section B.

Income stratification. Staff confirmed that the current stratifier is not an income concept that was ever used by OTA or JCT for other purposes. Rather, it was developed solely for the sample in order to ensure that returns with large amounts of income or large losses would be adequately represented for policy analytic purposes. AGI or other income concepts used by these agencies are not suitable for selecting the Individual sample because they make too much use of net amounts. As explained in Chapter II, the income concept that was developed includes some nontaxable or not fully taxable items of income and counts losses separately from income. Both OTA and JCT reaffirmed the importance of having separate positive and negative income totals. Only JCT staff had sufficient familiarity with the income stratifier to recognize that there may be issues with some of the fields. They noted that non-tested fields are an issue and wondered if tax exempt interest is included (it is). OTA was interested in learning more about what is in the stratifier.

High-income nontaxable returns. Both OTA and JCT have consulted internally with lawyers and other staff members with respect to the question of whether the law mandating the annual reporting of information on high-income nontaxable returns requires the collection of the entire universe of returns or whether a sample would suffice. Neither had received a definitive answer.

Regardless of whether the returns are sampled, there is no flexibility on the use of $200,000 in current dollars to determine which returns are included in this group.

Indexing. JCT staff had not realized that SOI used a chained index, which provides even less adjustment for real growth than the Consumer Price Index or an index of wage growth. They would prefer that income be indexed using a measure of national income.

Income growth. Staff from JCT felt that the growth in the top income category was probably the biggest problem posed by real income growth, although it did not present an issue, per see, for JCT. They suspected that such growth was a big cost driver for SOI, however.

Interesting returns. OTA staff observed that the IRI had become moot by the time that it was implemented for the 1991 tax year and should be dropped. They did not suggest a replacement. They noted that OTA's "green book" provides a sense of current policy issues. As an example they mentioned surtaxes on large incomes (for example, $5 million and up), but they did not suggest that these issues should be translated into elements of the sample selection. JCT staff did not see value in using something like the IRI in sample selection—particularly given what the CDW can provide (see below).

CWHS subsample. OTA, JCT, and CBO were strongly supportive of continuing to include 10 CWHS endings in the Individual sample. JCT staff viewed these returns as a permanent addition as long as they are relatively inexpensive to edit. Staff in all three offices recognize the value of having a large, simple random sample of the filing population with a substantial panel component, and the CWHS subsample provides this. Staff in OTA suggested considering the addition of the secondary SSN to CWHS selection. CBO would give up some CWHS endings on the primary SSN to have CWHS selection applied to the secondary SSN.

Late filers. JCT staff had the impression that the fraction of returns that are late is growing. They have noticed that a lot of the prior year returns are posted to the master file in January and wondered if SOI could shift or extend its processing cycle to encompass these returns. Six-month

extensions are common among very high-income returns, and they noted that IRS processing slows down in the latter part of the year, which they thought might contribute to pushing late-filed returns into the next processing year. OTA did not see late filers as a particular issue but would be interested in our following up on the JCT staff's suggestion to consider including in the Individual sample the prior year returns that post in January.

Timeliness of the Complete Report. The timeliness of delivery of the Complete Report data is fine for JCT and OTA. JCT staff remarked that getting the Complete Report a little earlier would be more helpful than getting the advance data any earlier, but they qualified this by clarifying that timeliness for the Complete Report matters only in the years that they rebuild their tax model, which is every few years. For CBO the Complete Report arrives too late to be used in their baseline projections. They incorporate the new Complete Report data later.

Usefulness of advance data. OTA uses the advance data less and less. November 15 remains a critical date. JCT staff thought that advance data were more important to OTA than to them because of differences in their relevant business cycles. CBO does not use the advance data because they are concerned about deficiencies among high-income returns, which are disproportionately missing from advance data because of taxpayers obtaining the automatic extension. BEA was not even aware of the advance estimates.

Usefulness of CDW data. OTA is using the CDW data increasingly to conduct analyses that cannot be performed with a sample. These analyses focus on rare tax issues and items. Improvements in the quality of CDW data would be helpful, for sure, but to be of much value to OTA such improvements would have to address the specific kinds of rare items for which OTA finds the CDW data useful. JCT also has a lot of interest in the CDW, although they have had access issues. They feel that the CDW can take the place of trying to identify potentially interesting returns for the Individual sample. The value of the CDW data is limited by the fields that are not included (which SOI adds to the Individual sample file through editing) and the data entry errors that are too

prevalent. CBO indicated that the CDW would be of greater interest to them if it were better edited. They mentioned in particular that the 2006 earned income tax credit indicator was "way off."

Design of the public use file. For those users who had access to only the public use sample, or only limited access to the non-public use data, the potential impact of any Individual sample design changes was of great interest. Anything that might affect the precision of estimates from the public use file was of particular interest. CBO would like to obtain the public use file sooner. They are currently using the 2006 data in their model. They noted as well that they do not need a state code.

Other needs. Most of the customers expressed a strong interest in obtaining a database of non-filers organized at the person level (that is, combining W-2s and 1099s that represent the same taxpayer). OTA noted that in efforts to work with the information documents as individual entities they cannot produce totals that are consistent with other estimates. CBO asked if the independent work of SOI and its customers on such data could be better coordinated—and CBO included. They would also like to have non-filer data added to the public use file.

OTA noted that they recognize the high cost of putting together the Sales of Capital Assets (SOCA) data and wonder if the usefulness of the data continues to justify these expenditures. Increasingly, sales of capital assets are carried out through pass-throughs, which are not reflected on Schedule D, which is capturing less and less of the total sales. They wondered if there is a way to capture a larger share of these sales in the SOCA database. BEA expressed concerns about the accuracy of capital gains estimates in general.

OTA is also interested in obtaining better state estimates. They are working on a statistical approach to improving the quality of state estimates from the Individual sample microdata but would be interested in other approaches.

OTA also noted that, increasingly, they are pulling together data across different SOI files in contrast to what they characterized as the SOI Division's "silo approach."

JCT staff commented that the fact that SOI data are pre-audit is a limitation for some uses. Revenue, they noted, is really a post-audit concept. Any information at all on post-audit data would be helpful. They also indicated that being able to look at the next return filed by someone who was audited would be useful, given that their behavior might change.

We did not specifically ask about stratification by form type, but we note that no one volunteered that increasing the representation of any particular form type was a priority or even particularly helpful. Both OTA and JCT benefit from the BEA supplemental sample of sole proprietorship returns, and this may reduce the need for additional returns of that type. This does suggest, however, that the impact of an overall sample reduction on the number of Schedule C returns selected may be an issue that needs to be addressed as part of any redesign. Maintaining the stratification by form type provides a ready means to oversample such returns should this be necessary or at least desirable.

We raised the issue of SSN quality, which appears to be declining, and how this might affect Individual sample selection and, perhaps more importantly, data from the CDW and efforts to do more with information documents. As a group the customers were not aware of such concerns; nor did they have any direct evidence of their own. But they recognized the importance of learning more about this phenomenon and its implications.

B. Empirical Findings

We conducted empirical analysis using Individual sample data for the years 2007 through 2009, with most of the analysis focusing on 2008. Our empirical analysis covered many of the sample design topics discussed with the SOI customers, namely the definition of income used for stratification, indexing, the specialized strata, stratification by filing mode, using the secondary SSN in sample selection, and late filing. Our examination of the income stratifier and indexing included assessments of the impact of prospective changes on editing time and the precision of the estimates of key variables.

1. Definition of Income Used for Stratification

Based on suggestions from Michael Strudler of the SOI Division, we examined the implications of five specific changes in the definition of positive and negative income:

- Adding Schedule C other income to positive or negative income
- Substituting taxable for tested total pension income
- Replacing total rent and royalties received with a net amount
- Replacing tested total social security benefits with taxable social security benefits
- Removing tax exempt interest from positive income

For each of these five, we determined how many records in the population would be reclassified into a different income class and then estimated the effect on the sample distribution. Table IV.1 shows the impact of each of these changes on the distribution of the population of returns in 2008 by income class. We review the results for each of the changes individually and then examine their combined impact.

Adding Schedule C Other Income. The income reported on Schedule C as other income is not included in the income stratifier despite being included in AGI. SOI staff have observed increasing amounts being reported on this line of Schedule C. Adding this additional income source to the income stratifier would appear to be a needed improvement.

Including this additional source of income produces an upward shift in the income distribution. Population sizes among the negative income classes are reduced by 1 to 4 percent, with the change being inversely related to the amount of negative income (the smaller the income, the greater the reduction). The impact on the positive income classes is smaller and opposite, with the number of returns changing not at all at incomes under $30,000 and increasing by only .06 percent in the next higher income class but as high as 0.77 percent at incomes between $2 million and $10 million.

Substituting Taxable for Tested Total Pension Income. In addition to including untaxed amounts, total pension income includes rollovers—that is, shifts of funds from one retirement account to another. Since rollovers do not represent any realization of income, whether taxable or

not, replacing total pension income with taxable pension income would appear to provide a clear improvement that would shift some persons from inappropriately high income strata to lower income strata.

The biggest impact of this change is to shift enough returns from positive to negative income so as to increase the size of the lowest negative income class by nearly 36 percent. The next higher negative income class is increased by nearly 6 percent, and each of the remaining negative income classes is increased by at least some amount, with the impact declining progressively as negative income rises. The top class is increased by just 0.20 percent. Among positive incomes, both of the two smallest classes are increased by less than 1 percent while higher income classes are reduced by growing fractions up to the class representing returns between $250,000 and $500,000, where the reduction is 8.71 percent. Reductions continue but by ever smaller fractions as income increases beyond that level.

Replacing Rent and Royalties with a Net Amount. The total rent and royalty payment amounts used to calculate the positive amounts total are untested fields in master file processing, so replacing them with alternative, tested fields would reduce classification errors. SOI staff recommended replacing the positive totals with net amounts, which can be positive or negative. This could produce a downward shift in the income distribution and reduce the size of the sample. A possible drawback is that it would reduce the number of sample returns with large rents or royalties.

This change produces very large increases among the negative income classes, with the four lowest negative income classes showing increases ranging from 21.68 percent to 47.14 percent. The impact diminishes as negative income increases, but even the highest negative income class shows an increase of 5 percent. All but the lowest positive income class shows a reduction in size, with declines ranging from 2.28 percent to 5.93 percent among classes with positive incomes of $60,000 or higher.

Replacing Tested Total Social Security with Taxable. Social security benefits do not get particularly large. The rationale for replacing tested total social security benefits with taxable benefits is to eliminate the effects of potential errors in the total benefits.

Most of the impact of this substitution of a taxable amount for a total amount is confined to low income levels. Returns with negative amounts below $60,000 are increased by nearly 30 percent while returns in the neighboring income classes (the lowest positive income class and the second smallest negative income class) are increased by 4 percent.. All of the positive income classes are reduced, with the effects diminishing as income grows. Between $30,000 and $60,000 the reduction is 7.73 percent, but above that level the reduction declines progressively from 1.91 percent to 0.02 percent. Conversely, the number of negative returns in all but the highest income classes grows, with the rate of increase generally declining with the absolute value of negative income.

Removing Tax Exempt Interest. Unlike the previous three changes, eliminating Tax Exempt Interest does not involve replacing one source with another. The impact is to reduce the income of every return reporting this source, but no income class is affected by very much. Positive income classes above $250,000 are reduced by 1.76 to 3.53 percent, and all negative income classes are increased, but by amounts ranging between 1.18 and 2.14 percent—a very narrow range. Positive income classes below $60,000 are increased as well, but by just over 0.1 percent.

Combined Impact. Table IV.2 reports the collective impact of the five changes to the definition of positive and negative total income. Every negative income class grows in size while every positive income class except the lowest decreases in size. The lowest negative income class more than doubles in size, growing from 1.4 million to 2.9 million returns—an increase of 115.50 percent. Returns with negative incomes between $60,000 and $120,000 increase by more than one-half. The increases grow progressively smaller as the absolute value of negative income rises, but even the highest income class grows by more than 5 percent. Returns with positive income below $30,000 grow by 3.2 million, but this represents only a 4.33 percent increase. All positive income

classes above $30,000 decline in size, with the classes between $120,000 and $2 million decreasing by nearly 10 percent to 16 percent. The top class declines by 5 percent. Overall, there is a net shift of 2 million returns from positive to negative income.

We did not explore additional changes to the income concept, but there is evidence that there may be good reason to do so. Currently, more than a dozen returns per year are assigned to different strata after editing than the ones to which they were assigned at selection. This is an exceedingly small number, but in the 2008 file one such case had an exceptionally high value on a component of AGI that is not included in the income stratifier. Because of this item, the AGI on that return was about 10 times as large as gross positive income. Had this return retained its original stratum, it would have been assigned a weight that greatly magnified its contribution to aggregate estimates and greatly increased the estimated variance of the income item with the excessive value. If the income stratifier does not include all components of AGI, perhaps it should at least include some option for dealing with cases where gross positive income as measured in the stratifier is significantly less than AGI. In the 2008 file we identified 799 returns with AGI that exceeded gross positive income by more than $100,000. Are there some common patterns that suggest additional variables to include in the stratifier? Further exploration of such cases would be merited even if a full redesign were not undertaken.

2. Indexing

When the 1987 sample redesign was implemented in 1991, indexing the income stratifier in order to keep the stratum boundaries in constant dollars was not part of the design. In 1996, however, the SOI Division introduced indexing based on the "Gross Domestic Product Implicit Price Deflator." Using fourth quarter values of this quarterly index, the income stratifier was converted to 1991 dollars in each year. This was accomplished by dividing the value of the income stratifier by the ratio of the current value of the index to its 1991 value. The indexed income amount could then be compared to fixed boundaries, which were expressed in 1991 dollars.

Published index values are revised with subsequent publication. Consequently, it is generally not possible to reconstruct earlier index ratios from more current published series. Table IV.3 shows the value of the index ratio used by SOI in each year from 1991 through 2010 and an index constructed from fourth quarter, seasonally adjusted values of a gross domestic product (GDP) price index published by the Bureau of Economic Analysis (BEA) on January 30, 2013. In 2008 the SOI index was 1.4181 while the GDP price index, converted to a fourth quarter 1991 base, was 1.4459 (Table IV.3).

We explored alternative indices based on personal income, as suggested by JCT staff. With an index based on personal income, the factor needed to convert 2008 dollars to 1991 dollars would be 2.4073.

The implications of these alternative indices for the current dollar equivalents of the 1991 stratum boundaries are shown in Table IV.4. With the SOI index, the boundary between the first two income classes is $42,543 instead of $30,000, and the boundary between the top two income classes is $14.181 million instead of $10.0 million. With the alternative index, based on personal income, the boundary between the first two income classes is raised to $72,219, and the boundary between the top two income classes is increased to $24.073 million.

Table IV.5 compares the allocation of the filing population, excluding HINT and high total receipt returns, with four different approaches to dealing with income growth over time. The first approach uses no indexing. The second uses the SOI index. The third uses an alternative index based on personal income, and the fourth maintains the 1991 proportionate distribution—that is, fixed shares of the population. The differences among the alternatives are striking. With no indexing the lowest positive income stratum would include 55.8 million returns, and the top income stratum would include 17,457 returns. With the SOI index the bottom positive income class grows to 74.5 million while the top income class falls to 10,595 returns. With the alternative index based on personal income the bottom positive income class would grow even further to 102.4 million returns

while the top class would decline to 4,936. If the 1991 shares were maintained, the bottom income class would fall between those with the SOI index and the alternative index, with 84.9 million returns, but the top income class would fall to 2,902 returns.

3. Impact of Changing Both the Income Definition and the Index

The effects of combining the change in the income stratifier with the substitution of an alternative index are of interest because any redesign of the Individual sample will almost certainly involve both. We examine the effects of these dual changes on the distribution of the population by income class and both the size and distribution of the sample. Using recent data on unit editing time by income class we then compare estimates of total editing time between the current design and the alternative design. Then we examine the impact of the two design changes, both separately and together, on the precision of estimates of a selection of items.

a. Population Distribution and Sample Size

Table IV.6 shows the distribution of the 2008 return population by income class for the current design and the three alternatives produced by changing the income stratifier to one that incorporates the five changes examined above, changing the index to one based on personal income, or changing both. As we have seen, changing the stratifier has a comparatively small effect among returns classified by positive gross income, but it more than doubles the number of returns with small losses exceeding small positive gross income. Changing the index redistributes returns toward smaller absolute income levels. Combining the two changes nearly triples the number of returns in the smallest negative income class while increasing the number of returns in the smallest positive income class by more than a third. All positive income classes above $250,000 are reduced by more than one half while returns with $60,000 to $250,000 are reduced by about two-thirds.

If we assume the same sampling rates by income class across the four designs, the changes in sample size by income class mirror the changes in population size (Table IV.7). But unlike the total population, which is constant across the four designs, the total sample size declines as returns in the

population are moved from income classes with higher sampling rates to those with lower sampling rates. Combining the two design changes produces only a slightly smaller sample size then replacement of the current index with a new index—the sample declines by fewer than 2,000 additional returns. This additional reduction is less than a third as large as the reduction achieved by implementing the new income definition alone, which implies considerable overlap in the effects of changing the stratifier and the index on the size and composition of the sample by income class. With the two changes the sample is 23 percent smaller than the actual 2008 sample.

b. Editing Cost

The SOI Division maintains data on average editing time for its sample of returns by stratum and filing mode (electronic versus paper). We used these data to estimate the total editing time for the alternative sample design and compare this estimate to one based on the current design. The estimates for the alternative design represent minimum editing times. While the underlying average editing times are differentiated by income class, the alternative design will redistribute returns among the income classes, generally shifting returns to lower income classes. It is highly likely that a redistribution of the magnitude that we have documented will increase the average editing time in every income class, as every class will gain returns that were in a higher income class under the current design (while the top income class will experience a rise in average income as the returns near the bottom of the income class are redistributed to lower income classes). We do not have a reliable way to estimate the impact on average editing time by income class, but it is important to recognize that the estimates of total editing time are almost certainly understated.

For 2008 the alternative design would reduce the total editing time from 57,000 hours with the current design to 32,000 hours, a reduction of 44 percent. Similarly, for 2009 the alternative design

would reduce the total editing time from 48,000 hours to 29,000 hours, a reduction of 40 percent.[8] While the reduction is overstated, it nevertheless indicates how a sample size reduction achieved by decreasing the size of all but the lowest positive income class can affect the time required to edit the sample.

c. Precision

Significant sample size reductions cannot be achieved without reducing the precision of the estimates obtained from the sample unless there is an extraordinarily effective reallocation of the sample. Using a program supplied by SOI, we calculated CVs by AGI class for estimates of the number of returns and aggregate amounts for 84 variables in 2008 and 2009.

The effects of a sample size reduction are not distributed uniformly across items because items differ in their distribution by income class. This is evident in Table IV.9, which displays sample sizes for the 84 variables across the four designs.[9] Items that are relatively more common among higher income than lower income returns have larger reductions than the total number of returns. We note in particular the Schedule D items, which have reductions between 30 and 40 percent, and the alternative minimum tax, which has a reduction of 52 percent. The smallest reductions are for the foreign earned income exclusion and the student loan interest deduction, both of which are between 11 and 12 percent.

[8] The lower aggregate editing time for both designs in 2009 versus 2008 reflect the impact of the Great Recession in producing a downward shift in the income distribution and, to a lesser extent, reducing the total number of returns filed.

[9] The sample sizes reported in Table IV.9 for number of returns under the three alternative designs are smaller than those reported in Table IV.7. The sample sizes in Table IV.9 are based on a simulation in which returns were subsampled based on their primary SSN transforms to obtain the set of sample records in each stratum that would be used to calculate CVs. This mechanism cannot add returns to reflect sample increases attributable to returns that the alternative designs would shift into strata with higher sampling rates than their current strata. Such increases occur with the new stratifier—primarily because of returns that shift between the low positive and low negative income classes. All of the changes in sampling rates produced by the new index are reductions, as every return's gross income is reduced, and no return is moved from positive to negative income.

Table IV.10 reports CVs for estimates of the number of returns for 34 items selected from the 84 in Table IV.9, and Table IV.11 reports the CVs for estimates of aggregate amounts. Arguably, the alternative designs should also be evaluated against the precision of the current design when it was implemented (see Chapter II). Toward this end, Tables IV.10 and IV.11 include CVs from 1996 where available. With one exception (for the amount of total rental and royalty net income), all of the alternative models have smaller CVs—and generally markedly so—than the 1996 sample. Given sample sizes more than twice as large, this is hardly surprising. Nevertheless, the SOI customers will judge any changes to the design relative to its current manifestation and the associated precision, so we focus our comparison on the current design in 2008.

Table IV.12 displays the percentage increase in CVs relative to the current design for the alternative design that combines a change in the stratifier with a change in the index. In general, the CV for the aggregate amount increases by more than the CV for the estimated number of returns. In fact, there are just two instances where the proportionate increase in the CV for an estimate of the number of returns matches or exceeds the increase for an estimate of the aggregate amount: exemptions and alternative minimum tax. Overall, the CVs for estimates of returns are increased by 10 to 20 percent while the CVs for estimates of amounts are increased by 30 to 40 percent, although some of the increases to the CVs for amounts are substantially lower than this and some are quite a bit higher. Among the former, the CVs for net loss from rent and royalties and for unemployment compensation are increased by just 7 to 8 percent while the CVs for net income from rent and royalties and alternative minimum tax are increased by around 60 percent, and the CVs for net short-term loss from sales of capital assets (SOCA) and Schedule D capital gain distributions are increased by 75 to 80 percent.

The increases in the CVs (loss of precision) with the alternative sample design are much larger than we would want to see with a new design. Since one of the major goals of a redesign would be to reduce the size of the sample, some loss of precision is expected, but the losses observed here are

excessive. It should be noted, however, that the alternative sample designs that we have explored were evaluated using the income classes (adjusted for inflation) and sampling rates of the current design. Under a full redesign, both would be subject to revision. In particular, the sampling rates would be optimized over a range of items as they were when the current design was developed (see Schirm and Czajka 1991).

To demonstrate the potential for improving the precision of the estimates obtained with the alternative design by changing the sample allocation across the new income classes, we did the following. For each of the variables listed in Table IV.11 we computed a Neyman allocation to determine the distribution of sampling rates that minimized the variance of the sample estimate for that variable. We did so using the strata obtained with the new stratifier and index and a fixed sample size of 252,588 (from Table IV.7). We constrained the sampling rates to equal 100% in the two specialized strata and the highest positive and negative income strata. We also constrained the minimum sampling rate to equal 0.10 percent although we did not require that this rate be used in any income class much less the three income classes in which it is used currently. Neither did we constrain the rates in the second highest positive and negative income strata to be 100% but allowed these strata to take whatever values minimized the variance of the variable for which the sample allocation was being optimized.

Only one allocation can be used for the sample, so we must choose an overall best allocation. We would like to select an allocation that minimizes sampling error over a wide range of variables. If we were engaged in an actual redesign of the Individual sample, this is what we would do. For the purposes of this report, we chose a more limited approach, selecting five variables and using the Neyman allocations for these five variables to generate five sets of CVs for all 34 variables. By comparing the CVs across the different sets of sampling rates, we can determine how much improvement in precision is possible with alternative allocations that reflect the range of what might be considered with a full redesign. We note, however, that none of these alternative allocations

involved changing the boundaries of the income classes themselves, which is another aspect of the sample design that would be explored in an actual redesign.

Table IV.13 shows the sampling rates by income class under the current design and the alternative sampling rates that would be obtained by optimizing the sample allocation for each of five variables. Interestingly, only one of the five alternative allocations would assign a sampling rate of 100 percent to the second highest positive or negative income stratum (although one would assign a rate of 100 percent to the third highest negative income stratum),[10] and none would assign the minimum sampling rate of 0.10 percent to all three of the lowest positive income classes. Note, however, that all five alternative allocations would assign the minimum rate to the lowest positive income class and all five would assign a rate no higher than 0.12 percent to the next lowest positive income class. In addition, two of the designs would assign sampling rates of 0.12 percent or lower to the lowest negative income class whereas the current design samples this income class at 0.19 percent.

With lower sampling rates at the second and third highest positive and negative income strata, these alternative allocations would have even lower editing costs than those reported in Table IV.8. Thus the sample could be increased above the 252,588 assumed here while still achieving a substantial reduction in editing time. With a full redesign in which the stratum boundaries were also changed, however, lowering the stratum boundaries below their indexed 1991 values (which we believe would be desirable) would move more returns into higher income classes than assumed in Table IV.13, and this would partially offset the editing cost reductions reported earlier.

Table IV.14 reports CVs for each of the 34 variables under seven different scenarios that use the alternative sample design (new stratifier and new index) with different allocations of a fixed sample size. The first scenario uses the current sampling rates; these CVs are the same as those

[10] Both rates exceed 100 percent in the current version of this table, but these are being corrected.

reported in column four of Table IV.11. The second scenario uses the optimal allocation for each of the 34 variables. These CVs represent the highest precision that can be obtained with the new stratifier and new index and the implied sample size. The next five scenarios reflect optimal allocations based on each of five variables—that is, the allocations reported in Table IV.13.

Comparing the first two columns we see very small improvements for amounts that are reported on all returns (AGI, exemptions, taxable income, and income tax before credits) but larger improvements for most of the remaining items, including in particular Schedule D capital gain distributions and alternative minimum tax.

Allocations based on the five variables cannot match these minimum CVs across all 34 variables, but the two that do best are the allocations based on AGI (E00100) and alternative minimum tax (E09600). Table IV.15 shows the percentage increase in CV compared to the current design if the alternative design were allocated in turn to minimize sampling error on these two variables.

4. Specialized Strata

Of the two specialized strata, from which the returns are selected into the Individual sample with certainty, the high-income nontaxable returns included more than 32,000 returns in 2008 while the returns with high Schedule C receipts numbered fewer than 400. While the average editing time for the high-income nontaxable returns is not particularly high, compared to returns with incomes between $250,000 and $500,000, their large number accounted for 10 percent of the total editing time for the Individual sample in 2009. The number of such returns included in the Individual sample far exceeds their value for policy analysis, and their overall cost is disproportionate as well.

As we noted earlier, it remains unclear whether the SOI Division is constrained by law to capture all of the high-income nontaxable returns in order to prepare the mandated annual report. The SOI mathematical statisticians investigated whether a sample of high-income nontaxable returns selected at rates consistent with their distribution by Individual sample stratum would be adequate to

support the tabulations in the annual report. They concluded that such a sample (of 1,199 returns) would not be sufficient, but they did not address the question of how large a sample would be necessary to produce estimates of adequate precision (Testa no date). We note that there is a lot of room between 1,200 and 32,000 to design a sample that could very well meet the precision requirements for the annual report and still reduce editing costs by a non-trivial amount. If the SOI Division is not required to continue selecting high-income nontaxable returns with certainty, it would seem that this portion of the sample could be reduced significantly without harming the precision of the full sample estimates.

Returns with high Schedule C total receipts, by contrast, represent a small group. Their inclusion as a certainty stratum is consistent with sampling the open-ended income and loss strata with certainty, and they account for less than 1 percent of the total editing time for the Individual sample.

5. Stratification by Filing Mode

Cost data shared by the SOI Division show that among the simplest, low-income returns, the average editing time for electronic returns is 43 percent of the average editing time for paper returns. The relative efficiency achieve with electronic returns grows with income. In the highest income class the average editing time for electronic returns is only 11 percent of the average editing time for paper returns. The difference between paper and electronic returns at any income level arises primarily from the greater ease of capturing from electronic returns the additional fields that are needed for the SOI records but not captured in revenue processing. High-income returns contain proportionately more such fields than low-income returns—hence the greater cost differential for returns in the top income class versus the bottom income class.

Given the markedly lower editing cost of electronic versus paper returns, the average editing cost per return could be reduced by substratifying on filing mode and sampling electronic returns at a higher rate than paper returns. Filing mode could be added to the post-stratification scheme to

eliminate any bias that the differential sampling rates might introduce. While lowering editing costs, however, such a change might reduce the precision of sample estimates. This would occur if paper returns were more heterogeneous than electronic returns. In that case, oversampling electronic returns would be the exact opposite of what should be done to minimize sampling error. Even if paper and electronic returns were equally heterogeneous, oversampling electronic returns would be less efficient than sampling the two types of returns at the same rate (within the current strata) if unequal weights were necessary within strata. A large cost reduction might justify a small loss of precision, but the point is that there might be trade-offs that would require a careful assessment to determine what was optimal.

To provide some information on this issue, we compared paper and electronic returns by income class, using the alternative income definition and index. For each class we computed the mean and standard deviation of AGI and net capital gains by filing mode, and we did so for 2008 and 2009.[11] We found similar results for both years, so only those for 2008 are reported here. For AGI the means are very similar except for the highest positive and negative income classes and the two specialized strata (Table IV.16). The standard deviations are also very similar except for the top two positive income classes and the highest negative income class and the two specialized strata—that is, all but one of the classes sampled with certainty under the current design.[12] Among these classes, paper returns are much more heterogeneous than electronic returns. It is noteworthy that in these strata, the sampling rate of 100 percent does not allow for differential selection of paper and

[11] We selected these two variables to represent a broad measure of income and a narrow measure—but one that is particularly important to the principal customers.

[12] In 2008, high income nontaxable returns show far more variability than returns in the income class with the most similar mean AGI: $120,000 to under $250,000. In 2009, the standard deviations among high income nontaxable returns are much more similar to those of returns in the indicated income class. We have no explanation for the markedly higher variability in 2008. Also of note, returns with high total receipts have much higher means among both paper and electronic returns in 2008 than in 2009. This is a small stratum, however, making the means particularly vulnerable to outliers. That the means are lower in 2009 is consistent with the weaker economy, but outliers are a more likely explanation, given the magnitude of the difference between the two years.

electronic returns. For net capital gains the means for paper returns are generally higher than those for electronic returns but not substantially so (Table IV.17). The standard deviations also tend to be a little higher on paper than electronic returns but not consistently so.

We also find that paper returns are the majority in the certainty strata in both years but that the dividing line between majority paper and majority electronic shifts between the two years. In 2008, paper returns dominate the strata with at least $250,000 in positive income or with at least $120,000 in negative income. In 2009, paper returns do not become the majority until $1 million in positive income or $500,000 in negative income.

The largely uniform heterogeneity between paper and electronic returns in all but the certainty strata suggests that electronic returns may be able to substitute for paper returns, but this assessment is based on only two variables. If electronic returns were oversampled in order to reduce editing time, we would recommend that the weights be post-stratified to population totals by stratum and filing mode. This could produce some loss of precision as discussed above.

Another consideration arguing against any effort to sample differentially is the diminishing importance of paper returns. In 2008, paper returns were 48 percent of the returns selected into the Individual sample. By 2012 they were only 21 percent. Among returns outside the certainty strata, where differential selection is possible, paper returns were only 17 percent of the total. We project from this that by the time a differential sampling plan could be implemented, the minimal savings in editing time would not justify the effort.

6. Using the Secondary SSN in Sample Selection

Extending CWHS selection to the secondary SSN would improve the panel properties of the Individual sample and thereby improve the precision of estimates of year-to-year change for returns in the lower portion of the income distribution, where CWHS returns represent the entire sample. Expanding selection to include use of the secondary as well as primary SSN transform would extend such improvement to the upper income portion of the sample. However, these changes would

increase the selection probabilities for joint versus non-joint returns and require separate weighting of such returns.

If joint returns were more heterogeneous than non-joint returns, a higher selection rate for joint returns might be advantageous with respect to precision. Similar to our analysis of paper versus electronic returns we estimated means and standard deviations of AGI and net capital gains among joint versus non-joint returns by income class. For AGI the means and standard deviations are generally similar for joint and non-joint returns (Table IV.18). For net capital gains the results differ between the negative and positive income classes, with joint returns having somewhat higher means and very slightly higher standard deviations than non-joint returns in the negative income classes but lower means and standard deviations in the positive income classes (Table IV.19). If these findings are typical of other variables, sampling joint returns at nearly twice the rate of single returns would reduce the overall precision of sample estimates, assuming the overall sample size did not change. This could be offset by increasing the sampling rates for non-joint returns, but then the selection of returns based on the secondary SSN would not be parallel to the selection based on primary SSN.

Other variables might show the greater heterogeneity among joint returns that would make selection on the secondary SSN more viable. Two key variables do not, however, so there will need to be some accommodation to these variables and others that show similar patterns between joint and non-joint returns. In view of the complexities introduced, expanding selection to the secondary SSN does not appear to have enough merit, on balance, to be included as part of an overall Individual sample redesign.

7. Late Filing

Returns that were due to be filed during a particular calendar year but were filed too late to be included in the Individual sample are represented, in effect, by returns from prior filing periods that were filed during the same year. The vast majority of these prior year returns are from the tax year immediately preceding the nominal tax year of the SOI sample. Likewise, most of the returns that

were due to be filed but were filed late are filed in the next processing year. Staff at the JCT noted that a significant fraction of late returns are filed in time to be included in the first processing cycle for the next calendar year—that is, in week or cycle 4. Seemingly, extending the sample selection to week 4 of the next year would enable many of the late returns to be included in the appropriate sample, but this would also delay the completion of the file by at least four weeks. It would also complicate return selection and processing, as it would be necessary to have selection and processing for two different tax years overlap for a period of a few weeks.

We compared prior year returns across a range of items for three successive years in order to determine their degree of resemblance. Specifically, we compared tax year 2006 returns in the 2007 SOI file, tax year 2007 returns in the 2008 SOI file, and tax year 2008 returns in the 2009 SOI file. If prior year returns are indeed a good substitute for late returns, we ought to see that the prior year returns processed in one year are similar to the prior year returns processed in the next year.

We find similarities for some items but substantial differences for others. Comparing means (Table IV.20) and totals (Table IV.21) we find very pronounced differences among Schedule D items. For example, the mean taxable net gain increases from $9,035 to $18,903 between the 2007 and 2008 files but then falls to $4,484 in the 2009 file. The total shows a similar change. Means and totals of nearly all of the variables rise and fall over the three years, reflecting changes in the economy.

Our overall conclusion is that late returns may not be very well represented by prior year returns, but they are a small group, and there is not an obviously better option. While we do find that a disproportionate share of late returns are processed in cycle 4, this proportion varied across the three years, rising from 25 percent in 2007 to 35 percent in 2008 but then falling to 18 percent in 2009 (unweighted). The weighted fractions were much smaller, indicating that the late returns processed early the next year are disproportionately higher income returns. This over-representation of higher income returns in cycle 4 is good news for a strategy that would extend selection through

cycle 4, but it is not enough to offset the complexities—staffing and otherwise—that would be introduced by overlapping the processing years and having to conduct sample selection for two different studies at the same time, using the same data.[13] Furthermore, the improvement in the data (which has not yet been demonstrated) is not likely to be sufficient to justify a four-week or greater delay in the completion and delivery of the file. Users would be better served by accepting the Individual sample file for what it is—a representative sample of returns processed during a given calendar year—and not treating it as a representative sample of returns *due* to be filed in that year.

C. Other Findings

We also considered issues with regard to missing returns, advance estimates, and the use of individual tax data from the CDW.

1. Missing Returns

The problem presented by missing returns has diminished over time. Statistics provided by the SOI Division show 187 missing returns in tax year 2009 and 95 in 2010. In 2009, 21 percent of the missing returns and in 2010 35 percent of the missing returns were from the high positive or negative income strata sampled with certainty. These returns raise the most concern, but their small magnitudes and the downward trend do not suggest a need to address these differently than is being done currently. Unlike returns assigned excessive weights, missing returns do not threaten to distort the estimates.

2. Advance Estimates

We compared the published advance estimates of total returns and amounts with the final estimates presented in the Complete Report for the years 1996, 2000, 2005, and 2010. For each year

[13] The Corporate sample is selected for two consecutive tax years at the same time because corporations have varying fiscal years, and the tax years for two returns needed for different samples may overlap. The Corporation branch has adapted its sampling and editing procedures to this reality, but this would require major changes for the Individual branch.

we calculated the distribution of percentage errors for the advance estimates that could be matched with final estimates. We also examined the time trend in errors for every item for which we were able to calculate percentage errors in at least three of the four years. Our results document the declining accuracy of the advance estimates over this period although not all items show this pattern.

The number of items included in the advance estimates grew over this period. Counting those for which we could match a Complete Report estimate the number of individual advance estimates grew from 119 in 1996 to 194 in 2010 (Table IV.22). The upper half of the table shows a frequency distribution of errors, with categories ranging from less than 0.5 percent to 20 percent or more. The lower half of the table converts the frequency distribution to a percentage distribution. We focus on the latter.

Differences in the error distributions between 1996 and 2000 are relatively small and in the direction of the advance estimates for 2000 being slightly more accurate than those for 1996. Most notably, the 2000 estimates have fewer errors between 1 and 2 percent and more below 1 percent than the 1996 estimates. There is a pronounced change in 2005, the year that the automatic extension was increased to six months. Between 2000 and 2005 the fraction of errors below 0.5 percent drops from 33.3 percent to 20.8 percent while the fraction between 10 and 20 percent grows from 3.8 percent to 7.8 percent. Accuracy continues to decline between 2005 and 2010. The fraction of items with errors under 1 percent drops from 40.9 percent to 31.5 percent while the proportion of items with errors between 10 and 20 percent grows from 7.8 to 9.3 percent, and the fraction with errors of 20 percent or more nearly doubles, from 1.9 to 3.6 percent.

It is possible, of course, that the decline in accuracy is associated with the increase in the number of items for which advance estimates are prepared, as the newer items tend to be less common items and perhaps more likely to appear on late returns than the more common items. To address the possibility that the decline in accuracy was due to the addition of new estimates for progressively weaker items rather than a growth in error for a common set of items, we identified 66

items for which there were advance estimates matched to final estimates in at least three of the four years. For most of these 66 items we had estimates of both returns and amounts.

Percentage errors for the advance estimates by year are reported in Table IV.23. There are 57 items for which we have error estimates for all four years although not always for both returns and amounts. For these items we have indicated with a double asterisk (**) the ones where the error grew progressively larger from 1996 to 2010—that is, where each year's error was larger than the previous year's error. We observed this pattern for advance estimates of 23 amounts and 11 counts of returns. Among the more prominent items and the growth in error between 1996 and 2010 are:

- Taxable interest (amount): 3.60 to 15.57
- Tax-exempt interest (amount): 1.59 to 7.91
- Ordinary dividends (amount): 1.27 to 15.30
- Business or profession net income (amount): 2.48 to 4.74
- Net capital gain reported on Schedule D (returns): 0.15 to 2.08
- Net capital gain reported on Schedule D (amount): 3.49 to 18.67
- Sales of property other than capital assets, net gain (returns): 0.20 to 6.18
- Sales of property other than capital assets, net gain (amount): 5.44 to 32.27
- Total itemized deductions (amount): 1.69 to 3.99
- Total income tax (amount): 0.43 to 0.75

Most of these items show dramatic growth in error over the period, and this decline in accuracy is all the more striking because the items are not rare.

When we do not see steady growth in error over the four years we sometimes see higher error in both 2005 and 2010 than in 1996 and 2000—or, if we have estimates for only three years, higher error for the later estimates than the earlier estimates. We observed this pattern, which is indicated by a single asterisk (*) for 17 advance estimates of returns and 14 advance estimates of amounts. Among the items in this group are salaries and wages (amount), payments to an Individual Retirement Arrangement (returns and amount), self-employment tax deduction (returns), payments to a Keogh plan (returns), taxable income (amount), and total tax liability (amount).

If the error in the advance estimates is growing over time, due to an increasing fraction of returns being filed late, we would not necessarily expect to see such patterns among all items. Items that are not associated with late filing would not be expected to show an increase in error over time. Thus we note that AGI and total deductions do not show steady growth in error over time, although in both cases the error in 2010 is higher than the error in 1996.

Overall, these results are consistent with comments from customers that the advance estimates have become less accurate over time. If advance estimates continue to be important to these customers, then some investment in better understanding the sources of the growing error and in determining how the estimates might be improved may be warranted, although such activity might very well be ranked behind other priorities for the Division's limited resources..

If an effort to improve the advance estimates is considered, we note the following. The advance estimates continue to be produced by adjusting the weights of the advance sample so that they sum to projections of the final population counts by stratum. This approach corrects for the differential distribution of the advance and final samples by stratum, but it assumes that there are no differences between the advance and final returns within stratum (that is, their means on the items being estimated are identical). Czajka et al. (1992) proposed an approach to improving the quality of the SOI advance estimates using propensity score modeling within stratum to construct weights that adjusted the advance sample for differences between advance and final returns in the prior year. That is, observations in the advance sample that resembled late returns more closely were assigned higher weights than observations that resembled late returns less closely. Reconsideration of this approach may be indicated as a possible way to improve the quality of the advance estimates although simpler alternatives should not be ruled out either.

3. Individual Tax Data from the CDW

The individual tax data maintained in the CDW are the same electronic data used to build the returns selected into the Individual sample, and their availability lags the data used by the SOI

Division. Thus CDW data offer nothing for sample selection or record construction that the SOI Division is not already using. They have become useful because they are much more accessible to those with access privileges than the transaction file data used in developing the SOI Individual sample. Their principal usefulness to policy analysts is the large sample size—the full population—that they provide for rare items.

Because the CDW data are unedited, their quality is a significant concern. The SOI Division compared aggregate estimates of about 100 items from Individual Return Transaction File (IRTF) data—the source of the CDW data on individuals—with the corresponding estimates produced from edited data in the 2011 SOI Individual sample. The median percentage difference was just under 1 percent, but 22 items had discrepancies in excess of 20 percent, and 7 had discrepancies in excess of 100 percent. Some of these large differences may reflect sampling error, and for these items it would be useful to construct a multi-year comparison. The information compiled from these comparisons would be highly valuable to users and prospective users of the CDW data.

D. Review of Documentation

User-accessible documentation on the Individual sample design consists of a sample description chapter included in the annual Complete Report publication (see, for example, Statistics of Income Division 2013). This brief summary includes citations to papers presented at the Joint Statistical Meetings by Mathematica and SOI staff. These papers describe aspects of the sample design in considerable detail, but they were prepared before the design was implemented and do not reflect subsequent modifications. The user reading these papers will find a description of a sample designed to yield 95,000 returns—not the 300,000 plus returns that are being selected currently. Nor do the earlier papers address all aspects of the design. Most notably, they do not provide a complete description of how the degree of interest is defined. The biggest change from the original design, the five-fold increase in the minimum sampling rate, is not discussed in the Complete Report chapter; nor is the elimination of differential sampling by degree of interest.

A highly valuable component of the Complete Report sample description is a table showing population and sample counts by income class, degree of interest, and the collapsed form type classification used for weighting. The table includes the two specialized strata as well. Sampling rates would be a useful addition, although the user can calculate them from the information provided in the table. The annual sample tables can be used to investigate changes over time in the composition of the filing population by stratum, as we did in Chapter II.

The level of detail in the sample description provided in the Complete Report is sufficient for the average CR reader, but it should be supported by comprehensive documentation that is accessible on line. Ideally, this documentation would:

- Provide a more precise definition of the SOI universe than appears in the Complete Report
- List the components of gross positive and negative income
- Delineate the full form type classification used in sample selection
- Explain how the four levels of the degree of interest are determined
- Define expanded income (used to designate high income nontaxable returns)
- List the target sampling rates by stratum
- Explain how selection using the transform is coordinated with the selection of returns from the CWHS subsample
- Provide a history of the design since its 1991 implementation
- Describe how the processing and editing of paper and electronic returns differs
- Explain how items that are missing from the original returns are imputed
- Discuss the handling of missing returns
- Explain how returns may be misclassified (assigned to the wrong stratum prior to selection) and what corrections are applied

The document should be updated periodically. Between updates, the sample description in the Complete Report should detail anything that has changed since the last update.

Table IV.1. Percentage Change is Population Counts by Stratum with Alternative Changes to the Income Definition, 2008

Income Class	Current Definition	Adding Schedule C Other Income	Replacing Tested Total Pension Income with Taxable	Replacing Total Rent and Royalties Received with Net	Replacing Tested Total Social Security with Taxable	Removing Tax Exempt Interest
-$10,000,000 or less	3,000	-1.03	0.20	5.13	0.00	1.37
-$9,999,999 to -$5,000,000	5,245	-1.01	0.27	7.28	0.13	1.58
-$4,999,999 to -$2,000,000	19,555	-1.40	0.41	10.31	0.08	1.54
-$1,999,999 to -$1,000,000	39,787	-1.53	0.78	13.05	0.19	1.98
-$999,999 to -$500,000	93,232	-1.54	0.94	16.82	0.65	2.14
-$499,999 to -$250,000	201,584	-1.66	2.22	21.68	0.69	2.14
-$249,999 to -$120,000	407,893	-2.19	2.87	32.77	1.24	1.22
-$119,999 to -$60,000	541,041	-3.41	5.68	47.14	4.02	2.12
-$59,999 to -$1	1,364,504	-3.98	35.68	37.65	29.80	1.18
$0 to under $30,000	74,540,523	0.00	0.55	0.19	3.75	0.13
$30,000 to under $60,000	35,502,729	0.06	0.26	-0.42	-7.73	0.12
$60,000 to under $120,000	21,078,098	0.15	-2.05	-2.29	-1.91	-0.34
$120,000 to under $250,000	6,164,484	0.30	-6.69	-5.52	-1.22	-0.92
$250,000 to under $500,000	1,649,221	0.56	-8.71	-5.93	-0.53	-1.76
$500,000 to under $1,000,000	552,763	0.59	-6.57	-5.12	-0.38	-2.14
$1,000,000 to under $2,000,000	183,017	0.61	-4.67	-5.24	-0.13	-3.02
$2,000,000 to under $5,000,000	75,230	0.77	-3.02	-4.77	-0.12	-3.07
$5,000,000 to under $10,000,000	17,840	0.77	-1.28	-3.51	-0.06	-3.53
$10,000,000 or more	10,823	0.67	-0.60	-2.28	-0.02	-3.16

Note: All estimates include high income nontaxable returns and high gross receipt returns, which have been assigned to income classes based on their positive or negative income.

Table IV.2. Stratification of the 2008 Filing Population with All Changes to the Income Definition

Income Class	Current Definition	Combining All Income Changes	Net Change in Number of Returns	Percentage Change in Number of Returns
-$10,000,000 or less	3,000	3,161	161	5.37
-$9,999,999 to -$5,000,000	5,245	5,684	439	8.37
-$4,999,999 to -$2,000,000	19,555	21,707	2,152	11.00
-$1,999,999 to -$1,000,000	39,787	45,554	5,767	14.49
-$999,999 to -$500,000	93,232	110,829	17,597	18.87
-$499,999 to -$250,000	201,584	252,231	50,647	25.12
-$249,999 to -$120,000	407,893	556,460	148,567	36.42
-$119,999 to -$60,000	541,041	845,512	304,471	56.28
-$59,999 to -$1	1,364,504	2,940,520	1,576,016	115.50
$0 to under $30,000	74,540,523	77,766,244	3,225,721	4.33
$30,000 to under $60,000	35,502,729	32,736,477	-2,766,252	-7.79
$60,000 to under $120,000	21,078,098	19,724,058	-1,354,040	-6.42
$120,000 to under $250,000	6,164,484	5,322,341	-842,143	-13.66
$250,000 to under $500,000	1,649,221	1,385,336	-263,885	-16.00
$500,000 to under $1,000,000	552,763	479,077	-73,686	-13.33
$1,000,000 to under $2,000,000	183,017	160,896	-22,121	-12.09
$2,000,000 to under $5,000,000	75,230	67,729	-7,501	-9.97
$5,000,000 to under $10,000,000	17,840	16,497	-1,343	-7.53
$10,000,000 or more	10,823	10,256	-567	-5.24
	142,450,569	142,450,569	0	0.00

Source: Mathematica tabulations of the 2008 Complete Report file.

Note: All estimates include high income nontaxable returns and high gross receipt returns, which have been assigned to income classes based on their positive or negative income.

Table IV.3. Alternative Index Values for Stratum Boundaries

Year	SOI Index	GDP Price Index	Personal Income Index	Published Values: GDP Price	Published Values: Personal Income
1991	1.0000	1.0000	1.0000	75.557	5,126.1
1992	1.0000	1.0219	1.0712	77.212	5,490.9
1993	1.0000	1.0440	1.1138	78.883	5,709.2
1994	1.0000	1.0664	1.1757	80.572	6,026.6
1995	1.0000	1.0877	1.2299	82.180	6,304.7
1996	1.1030 [a]	1.1081	1.3137	83.721	6,734.3
1997	[b]	1.1260	1.3998	85.080	7,175.5
1998	1.1403 [c]	1.1377	1.4989	85.962	7,683.6
1999	1.1480	1.1561	1.5830	87.350	8,114.7
2000	1.1640	1.1837	1.6986	89.435	8,707.3
2001	1.1914	1.2074	1.7386	91.225	8,912.3
2002	1.1640 [d]	1.2295	1.7805	92.894	9,126.8
2003	1.2297	1.2551	1.8685	94.833	9,578.3
2004	1.2510	1.2954	1.9943	97.876	10,223.1
2005	1.2510 [e]	1.3408	2.0953	101.305	10,740.8
2006	1.3386	1.3792	2.2444	104.206	11,504.8
2007	1.3794	1.4156	2.3685	106.956	12,141.4
2008	1.4181	1.4459	2.4073	109.247	12,340.0
2009	1.4459	1.4535	2.3149	109.820	11,866.2
2010	1.4530	1.4802	2.4404	111.838	12,509.9

Source: Statistics of Income Complete Report and BEA, Table 1.1.4, Price Indexes for Gross Domestic Product, and Table 2.1, Personal Income and its Disposition, January 30, 2013.

[a] SOI did not start adjusting the stratum boundaries until 1996.

[b] The index value was not reported in the sample table.

[c] This is the first time the SOI index was described as a chain-type price index.

[d] The published value is identical to 2000 instead of falling between the 2001 and 2003 values. Since the GDP price index rose between 2001 and 2002, this is clearly an error.

[e] The published value is identical to the prior year, suggesting that the footnote was not updated.

Table IV.4. 2008 Dollar Value of SOI Individual Sample Income Class Boundaries with Alternative Price Indices

Income Class Boundary	Value in 1991 Dollars	Value in 2008 Dollars with SOI Index of 1.4181	Value in 2008 Dollars with an Alternative Index of 2.4073
1	30,000	42,543	72,219
2	60,000	85,086	144,438
3	120,000	170,172	288,876
4	250,000	354,525	601,825
5	500,000	709,050	1,203,650
6	1,000,000	1,418,100	2,407,300
7	2,000,000	2,836,200	4,814,600
8	5,000,000	7,090,500	12,036,500
9	10,000,000	14,181,000	24,073,000

Table IV.5. Income Stratification of the 2008 Filing Population with Alternative Indexing

Income Class	No Indexing	Indexed by Price Index for Gross Domestic Product 2008 = 1.4181	Indexed by Personal Income 2008 = 2.4073	1991 Fixed Shares
Total returns	142,417,596	142,417,596	142,417,596	142,417,596
-$10,000,000 or less	4,676	2,807	1,397	1,389
-$9,999,999 to -$5,000,000	7,956	5,045	2,131	1,826
-$4,999,999 to -$2,000,000	29,533	18,913	9,806	6,607
-$1,999,999 to -$1,000,000	59,576	38,841	19,746	12,780
-$999,999 to -$500,000	136,986	91,784	47,105	31,717
-$499,999 to -$250,000	284,573	200,185	110,756	76,042
-$249,999 to -$120,000	503,959	407,367	258,787	164,323
-$119,999 to -$60,000	546,872	540,982	434,730	238,893
-$59,999 to -$1	1,096,243	1,364,450	1,785,916	762,751
$0 to under $30,000	55,814,510	74,540,433	102,386,704	84,948,479
$30,000 to under $60,000	37,127,541	35,502,645	25,647,305	35,876,216
$60,000 to under $120,000	30,241,848	21,077,981	8,299,654	15,357,798
$120,000 to under $250,000	12,199,952	6,151,761	2,336,732	3,570,357
$250,000 to under $500,000	2,918,434	1,641,068	713,185	966,718
$500,000 to under $1,000,000	957,649	549,218	233,597	282,413
$1,000,000 to under $2,000,000	316,566	181,401	81,055	81,284
$2,000,000 to under $5,000,000	124,223	74,506	35,561	29,047
$5,000,000 to under $10,000,000	29,042	17,614	8,493	6,054
$10,000,000 or more	17,457	10,595	4,936	2,902

Note: All estimates exclude high income nontaxable returns and high gross receipt returns. The size of these specialized strata would not be affected by indexing or by fixing the shares.

Table IV.6. Distribution of the 2008 Filing Population by Income Class Under Alternative Sample Designs

Income Class	Current Design	New Stratifier with Current Index	Current Stratifier with New Index	New Stratifier with New Index
Total returns	142,450,569	142,450,569	142,450,569	142,450,569
High-income nontaxable	32,591	32,591	32,591	32,591
High Schedule C receipts	382	382	382	382
-$10,000,000 or less	2,811	2,949	1,397	1,458
-$9,999,999 to -$5,000,000	5,146	5,451	2,131	2,275
-$4,999,999 to -$2,000,000	18,926	20,951	9,806	10,635
-$1,999,999 to -$1,000,000	38,842	44,385	19,746	21,952
-$999,999 to -$500,000	91,778	108,937	47,105	54,593
-$499,999 to -$250,000	200,278	250,207	110,756	134,197
-$249,999 to -$120,000	407,323	554,929	258,787	332,559
-$119,999 to -$60,000	540,978	844,935	434,730	612,224
-$59,999 to -$1	1,363,441	2,940,158	1,785,916	3,603,011
$0 to under $30,000	74,535,462	77,765,757	102,386,704	103,177,656
$30,000 to under $60,000	35,507,627	32,735,654	25,647,305	24,267,323
$60,000 to under $120,000	21,078,572	19,722,202	8,299,654	7,292,748
$120,000 to under $250,000	6,151,777	5,311,569	2,336,732	1,977,489
$250,000 to under $500,000	1,641,057	1,379,625	713,185	607,767
$500,000 to under $1,000,000	549,382	476,576	233,597	204,080
$1,000,000 to under $2,000,000	181,441	159,710	81,055	72,393
$2,000,000 to under $5,000,000	74,502	67,191	35,561	32,541
$5,000,000 to under $10,000,000	17,621	16,322	8,493	7,985
$10,000,000 or more	10,632	10,088	4,936	4,712

Source: Mathematica tabulations of 2008 INSOLE file, Reject 0 records.

Table IV.7. Estimated Sample Counts by Income Class Under Alternative Sample Designs, 2008

Income Class	2008 Sampling Rates	Current Design	New Stratifier with Current Index	Current Stratifier with New Index	New Stratifier with New Index
Total returns		328,468	322,000	254,187	252,588
High-income nontaxable	100.00	32,591	32,591	32,591	32,591
High Schedule C receipts	100.00	382	382	382	382
-$10,000,000 or less	100.00	2,811	2,949	1,397	1,458
-$9,999,999 to -$5,000,000	100.00	5,046	5,451	2,131	2,275
-$4,999,999 to -$2,000,000	33.76	6,389	7,073	3,310	3,590
-$1,999,999 to -$1,000,000	15.84	6,153	7,031	3,128	3,477
-$999,999 to -$500,000	3.31	3,035	3,602	1,558	1,805
-$499,999 to -$250,000	0.99	1,986	2,481	1,098	1,331
-$249,999 to -$120,000	0.51	2,081	2,835	1,322	1,699
-$119,999 to -$60,000	0.31	1,702	2,658	1,368	1,926
-$59,999 to -$1	0.18	2,516	5,426	3,296	6,649
$0 to under $30,000	0.11	80,454	83,941	110,517	111,371
$30,000 to under $60,000	0.12	41,880	38,611	30,250	28,622
$60,000 to under $120,000	0.11	24,181	22,625	9,521	8,366
$120,000 to under $250,000	0.29	17,540	15,144	6,663	5,638
$250,000 to under $500,000	0.72	11,744	9,873	5,104	4,349
$500,000 to under $1,000,000	2.48	13,622	11,817	5,792	5,060
$1,000,000 to under $2,000,000	12.17	22,073	19,429	9,861	8,807
$2,000,000 to under $5,000,000	32.25	24,029	21,671	11,469	10,495
$5,000,000 to under $10,000,000	100.00	17,621	16,322	8,493	7,985
$10,000,000 or more	100.00	10,632	10,088	4,936	4,712

Source: Mathematica tabulations of 2008 INSOLE file, Reject 0 records.

Note: Estimates were calculated by applying the 2008 sampling rates by income class to the population totals in Table IV.6.

Table IV.8. Hours to Edit the SOI Sample by Income Class If the 2010 Editing Rates Were Applied to the Current Design and the Alternative Design with a New Stratifier and New Index, 2008 and 2009

Income Class	2010 Editing Minutes per Return	2008		2009	
		Current Design	Alternative Design	Current Design	Alternative Design
Total		57,171	31,845	48,380	29,165
High income nontaxable	8.34	4,530	4,530	4,886	4,886
High total receipts	64.39	410	410	331	331
-$10,000,000 or less	53.07	2,486	1,290	2,901	1,616
-$9,999,999 to -$5,000,000	25.02	2,104	949	2,406	1,183
-$4,999,999 to -$2,000,000	19.66	2,093	1,191	2,465	1,451
-$1,999,999 to -$1,000,000	15.01	1,539	869	1,777	1,081
-$999,999 to -$500,000	12.47	631	371	730	461
-$499,999 to -$250,000	9.87	327	216	377	272
-$249,999 to -$120,000	8.86	307	241	354	295
-$119,999 to -$60,000	8.45	240	259	276	296
-$59,999 to -$1	5.05	212	576	240	569
$0 to under $30,000	2.44	3,278	4,196	3,037	4,075
$30,000 to under $60,000	3.76	2,624	1,521	2,185	1,554
$60,000 to under $120,000	5.51	2,221	670	1,835	715
$120,000 to under $250,000	6.70	1,959	729	2,126	714
$250,000 to under $500,000	8.51	1,666	621	1,480	556
$500,000 to under $1,000,000	11.00	2,497	928	2,143	777
$1,000,000 to under $2,000,000	12.32	4,532	1,814	3,545	1,406
$2,000,000 to under $5,000,000	16.98	6,800	2,987	4,825	2,133
$5,000,000 to under $10,000,000	23.76	6,978	3,162	4,573	2,072
$10,000,000 or more	54.95	9,737	4,315	5,888	2,722

Note: Estimates of editing time are based on actual 2010 editing times by three-digit sample code and assume the 2010 composition (by return type, degree of interest, and paper versus electronic filing).

Table IV.9. Sample Sizes by Item for Alternative Individual Sample Designs, 2008

Item	Description	Sample Size: Current Design	Sample Size: New Stratifier with Current Index	Sample Size: Current Stratifier with New Index	Sample Size: New Stratifier with New Index	Percent Reduction from Current Design
N1	Number of returns	328,468	304,880	241,557	239,226	27.2
n00200	Salaries and wages	251,828	233,856	186,333	184,406	26.8
n00300	Taxable interest	231,790	215,325	155,573	153,112	33.9
n00400	Tax exempt interest	100,624	94,519	64,069	61,771	38.6
n00600	Ordinary dividends	176,598	164,997	114,165	111,596	36.8
n00650	Qualified dividends	161,125	150,577	103,424	100,960	37.3
n00700	State income tax refunds	84,057	78,291	55,304	54,190	35.5
n00800	Alimony received	886	685	572	578	34.8
n00900p	Business or profession net income	56,914	50,384	35,181	34,745	39.0
n00900n	Business or profession net loss	25,634	23,762	16,972	16,849	34.3
n01100	Capital gain distributions	7,194	6,608	4,932	4,882	32.1
n01000p	D Taxable net gain	72,761	68,004	45,362	44,083	39.4
n01000n	D Taxable net loss	90,852	85,493	57,997	56,877	37.4
n22250p	D Net short-term capital gain	23,843	22,468	15,199	14,797	37.9
n22250n	D Net short-term capital loss	101,733	96,080	63,820	62,119	38.9
n21800	D Short-term loss carryover	21,977	21,164	13,185	13,006	40.8
n21600p	D Net short-term gain from SOCA	20,698	19,440	13,124	12,767	38.3
n21600n	D Net short-term loss from SOCA	86,264	81,249	54,307	52,738	38.9
n21620p	D Net short-term gain from other forms	21,556	20,775	14,532	14,152	34.3
n21620n	D Net short-term loss from other forms	11,105	10,686	7,282	7,083	36.2
n21775p	D Net short-term partnership/S-corps gain	19,320	18,504	12,299	11,952	38.1
n21775n	D Net short-term partnership/S-corps loss	36,705	35,122	23,966	23,259	36.6
n23250p	D Net long-term capital gain	86,921	81,344	54,407	52,783	39.3
n23250n	D Net long-term capital loss	69,221	65,206	44,067	43,305	37.4
n22300p	D Net long-term gain from SOCA	56,887	53,137	36,271	35,153	38.2
n22300n	D Net long-term loss from SOCA	65,281	61,339	41,547	40,534	37.9
n22390	D Net long-term loss carryover	30,709	29,158	18,805	18,568	39.5
n22320p	D Net long-term gain from other forms	55,137	52,376	34,644	33,693	38.9
n22320n	D Net long-term loss from other forms	7,627	7,326	4,922	4,788	37.2
n22365p	D Net long-term partnership/S-corps gain	42,076	40,226	26,809	26,023	38.2
n22365n	D Net long-term partnership/S-corps loss	26,634	25,459	17,281	16,782	37.0
n22370	D Capital gain distributions	87,957	82,407	55,420	53,669	39.0
n01200p	Sale of property other than capital assets, gain	20,547	19,681	12,505	12,199	40.6
n01200n	Sale of property other than capital assets, loss	26,469	25,446	16,408	16,060	39.3
n01400	Taxable IRA distributions	33,459	29,378	22,960	22,340	33.2
n01500	Pensions and annuities, total	75,497	64,963	51,733	50,012	33.8
n01700	Pensions and annuities, taxable	64,273	55,000	44,651	43,222	32.8
n25700p	Rent, net income	33,541	30,160	19,610	18,915	43.6
n25700n	Rent, net loss	36,127	33,518	22,626	22,835	36.8
n25800p	Royalty, net income	35,775	33,814	23,311	22,618	36.8
n25800n	Royalty, net loss	1,460	1,410	1,018	993	32.0
n27200p	Farm rental, net income	2,081	1,843	1,281	1,235	40.7
n27200n	Farm rental, net loss	635	571	396	388	38.9

Continued

Table IV.9 continued

Item	Description	Sample Size: Current Design	Sample Size: New Stratifier with Current Index	Sample Size: Current Stratifier with New Index	Sample Size: New Stratifier with New Index	Percent Reduction from Current Design
n27310p	Total rental and royalty, net income	59,291	54,622	36,603	35,425	40.3
n27310n	Total rental and royalty, net loss	26,784	24,784	17,017	17,475	34.8
n26270p	Partnership and S-corp, net income	67,308	63,026	38,030	36,771	45.4
n26270n	Partnership and S-corp, net loss	51,245	49,133	33,608	33,049	35.5
n26500p	Estate and trust, net income	9,405	8,719	5,987	5,726	39.1
n26500n	Estate and trust, net loss	3,193	3,077	2,225	2,170	32.0
n02100p	Farm net income	4,577	4,338	2,331	2,310	49.5
n02100n	Farm net loss	9,336	8,724	5,645	5,563	40.4
n02300	Unemployment compensation	12,576	11,225	10,371	10,344	17.7
n02400	Social Security benefits, total	69,194	60,323	48,620	47,694	31.1
n02500	Social Security benefits, taxable	53,712	46,676	37,145	35,856	33.2
n02700	Foreign earned income exclusion	10,291	10,101	9,174	9,157	11.0
n02600p	Other income, net income	48,137	44,961	31,296	30,515	36.6
n02600n	Other income, net loss	6,075	5,923	4,679	4,645	23.5
n02540	Net operating loss	13,727	13,539	8,487	8,649	37.0
n02800	Gambling earnings	9,083	8,387	6,375	6,318	30.4
n02610	Cancellation of debt	8,972	8,641	6,356	6,224	30.6
n02900	Statutory adjustments, total	130,593	119,914	83,867	82,560	36.8
n03220	Educator expenses deduction	5,654	5,151	4,359	4,322	23.6
n03700	Certain business expenses	189	169	147	146	22.8
n03290	Health savings account deduction	5,729	5,399	3,337	3,273	42.9
n03280	Moving expenses adjustment	2,040	1,895	1,580	1,563	23.4
n03260	Deduction for one-half of self-employment tax	82,760	74,926	49,475	48,562	41.3
n03300	Payments to a Keogh Plan	15,967	15,103	8,280	8,017	49.8
n03270	Self-employed health insurance deduction	37,481	35,191	20,834	20,385	45.6
n03400	Penalty on early withdrawal of savings	4,366	4,012	2,984	2,930	32.9
n03500	Alimony paid	5,213	4,966	3,237	3,182	39.0
n03150	IRA payments	8,597	7,795	5,505	5,422	36.9
n03210	Student loan interest deduction	10,870	10,097	9,550	9,585	11.8
n03230	Tuition and fees deduction	7,488	6,913	5,940	5,987	20.0
n03240	Domestic production activities deduction	19,917	19,158	11,737	11,367	42.9
n03900	Other adjustments	486	445	332	322	33.7
n04100	Basic standard deduction	114,741	104,972	100,719	100,797	12.2
n04250	Real estate deduction	17,756	19,228	17,546	13,947	21.5
n04200	Additional standard deduction	22,465	14,969	13,926	17,492	22.1
n04470	Total itemized deductions	193,947	180,082	128,253	125,516	35.3
n04600	Exemptions	317,875	294,783	231,589	229,259	27.9
n04805	Capital construction fund reduction	501	478	313	299	40.3
n04800	Taxable income	252,935	233,562	180,218	177,175	30.0
n09600	Alternative minimum tax	50,213	46,190	25,707	24,129	51.9
n05800	Income tax before credits	248,217	228,912	175,236	172,119	30.7

Source: Statistics of Income Division, special tabulation, and Mathematica.

Table IV.10. Estimated Coefficients of Variation of the Aggregate Number of Returns for Selected Items under Alternative Individual Sample Designs, 2008

Item	Description	Coefficients of Variation (Percent)				
		Current Design	New Stratifier with Current Index	Current Stratifier with New Index	New Stratifier with New Index	Current Design in 1996
N1	Number of returns	0.01	0.00 [a]	0.00 [a]	0.00 [a]	0.04
n00200	Salaries and wages	0.10	0.11	0.11	0.11	0.18
n00300	Taxable interest	0.24	0.26	0.27	0.27	0.39
n00600	Ordinary dividends	0.39	0.42	0.43	0.43	0.73
n00900p	Business or profession net income	0.35	0.39	0.41	0.41	0.60
n00900n	Business or profession net loss	0.95	1.05	1.10	1.09	1.74
n01100	Capital gain distributions	1.55	1.62	1.66	1.66	2.08
n01000p	D Taxable net gain	0.86	0.94	0.98	0.98	1.08
n01000n	D Taxable net loss	0.65	0.69	0.73	0.73	1.88
n21600p	D Net short-term gain from SOCA	1.64	1.75	1.84	1.84	n.a.
n21600n	D Net short-term loss from SOCA	0.85	0.91	0.97	0.96	n.a.
n22300p	D Net long-term gain from SOCA	1.05	1.15	1.21	1.21	n.a.
n22300n	D Net long-term loss from SOCA	0.85	0.91	0.96	0.95	n.a.
n22320p	D Net long-term gain from other forms	1.47	1.66	1.81	1.80	2.04
n22320n	D Net long-term loss from other forms	4.79	5.16	5.67	5.69	10.48
n22370	D Capital gain distributions	0.77	0.83	0.87	0.87	n.a.
n01400	Taxable IRA distributions	0.78	0.85	0.87	0.87	1.84
n01700	Pensions and annuities, taxable	0.48	0.54	0.55	0.55	0.93
n27310p	Total rental and royalty, net income	0.98	1.12	1.18	1.18	1.47
n27310n	Total rental and royalty, net loss	1.11	1.23	1.30	1.25	1.75
n26270p	Partnership and S-corp, net income	0.96	1.11	1.20	1.20	1.71
n26270n	Partnership and S-corp, net loss	1.36	1.46	1.58	1.56	2.56
n26500p	Estate and trust, net income	3.17	3.58	3.80	3.82	5.18
n26500n	Estate and trust, net loss	10.49	11.04	12.29	12.39	14.72
n02100p	Farm net income	2.80	3.17	3.37	3.37	3.79
n02100n	Farm net loss	1.43	1.58	1.67	1.68	2.31
n02300	Unemployment compensation	0.91	0.97	0.98	0.98	1.74
n02500	Social Security benefits, taxable	0.63	0.72	0.73	0.74	1.46
n02900	Statutory adjustments, total	0.36	0.37	0.38	0.38	0.74
n04470	Total itemized deductions	0.27	0.29	0.31	0.31	0.54
N2	Exemptions	0.15	0.16	0.17	0.17	0.28
n04800	Taxable income	0.13	0.13	0.14	0.14	0.24
n09600	Alternative minimum tax	0.65	0.78	1.08	1.08	n.a.
n05800	Income tax before credits	0.13	0.13	0.14	0.14	0.24

Source: Statistics of Income Division, special tabulation, and Mathematica.

[a] The CV for number of returns cannot be estimated from Reject 0 returns using SAS Proc Survey Means because the sample estimate and population estimate are identical.

Table IV.11. Estimated Coefficients of Variation of Aggregate Dollar Amounts for Selected Items under Alternative Individual Sample Designs, 2008

Item	Description	Coefficients of Variation (Percent)				
		Current Design	New Stratifier with Current Index	Current Stratifier with New Index	New Stratifier with New Index	Current Design in 1996
E00100	Adjusted gross income less deficit	0.09	0.09	0.12	0.12	0.16
E00200	Salaries and wages	0.16	0.17	0.21	0.20	0.28
E00300	Taxable interest	0.65	0.69	0.82	0.84	1.18
E00600	Ordinary dividends	0.66	0.72	0.88	0.92	1.31
E00900p	Business or profession net income	0.76	0.79	1.01	1.01	1.13
E00900n	Business or profession net loss	1.30	1.42	1.71	1.70	2.26
E01100	Capital gain distributions	4.25	3.96	4.82	4.98	5.44
E01000p	D Taxable net gain	0.45	0.51	0.69	0.71	0.83
E01000n	D Taxable net loss	0.69	0.73	0.79	0.78	2.02
E21600p	D Net short-term gain from SOCA	3.09	3.27	3.52	3.54	n.a.
E21600n	D Net short-term loss from SOCA	0.76	0.83	1.14	1.15	n.a.
E22300p	D Net long-term gain from SOCA	0.74	0.82	1.09	1.12	n.a.
E22300n	D Net long-term loss from SOCA	0.94	1.01	1.31	1.32	n.a.
E22320p	D Net long-term gain from other forms	0.85	0.95	1.32	1.38	2.37
E22320n	D Net long-term loss from other forms	4.26	4.51	6.70	6.69	6.91
E22370	Schedule D Capital gain distributions	1.98	2.13	3.38	3.55	n.a.
E01400	Taxable IRA distributions	1.29	1.45	1.55	1.64	3.06
E01700	Pensions and annuities, taxable	0.73	0.79	0.83	0.85	1.36
E27310p	Total rental and royalty, net income	1.25	1.52	1.82	1.98	1.63
E27310n	Total rental and royalty, net loss	1.40	1.37	1.82	1.50	2.08
E26270p	Partnership and S-corp, net income	0.57	0.63	0.87	0.87	1.04
E26270n	Partnership and S-corp, net loss	0.75	0.82	1.06	1.07	1.84
E26500p	Estate and trust, net income	2.62	2.92	3.82	3.92	4.55
E26500n	Estate and trust, net loss	3.14	3.44	4.43	4.49	7.66
E02100p	Farm net income	3.10	3.20	4.14	4.12	4.54
E02100n	Farm net loss	2.03	2.21	2.65	2.66	3.03
E02300	Unemployment compensation	1.26	1.35	1.36	1.36	2.41
E02500	Social Security benefits, taxable	0.76	0.85	0.89	0.90	1.75
E02900	Statutory adjustments, total	0.65	0.69	0.83	0.84	1.21
E04470	Total itemized deductions	0.27	0.29	0.32	0.32	0.55
E04600	Exemptions	0.15	0.16	0.17	0.17	0.28
E04800	Taxable income	0.11	0.12	0.16	0.15	0.21
E09600	Alternative minimum tax	0.70	0.81	1.06	1.13	n.a.
E05800	Income tax before credits	0.14	0.15	0.19	0.19	0.23

Source: Statistics of Income Division, special tabulation, and Mathematica.

Table IV.12. Estimated Increase in Coefficients of Variation by Item for an Alternative Sample Design with a New Stratifier and New Index, 2008

Item	Description	Percentage Increase in CV for Number of Returns	Percentage Increase in CV for Aggregate Dollar Amount
E00100	Adjusted gross income less deficit	n.a.	33.3
E00200	Salaries and wages	10.0	25.0
E00300	Taxable interest	12.5	29.2
E00600	Ordinary dividends	10.3	39.4
E00900p	Business or profession net income	17.1	32.9
E00900n	Business or profession net loss	14.7	30.8
E01100	Capital gain distributions	7.1	17.2
E01000p	D Taxable net gain	14.0	57.8
E01000n	D Taxable net loss	12.3	13.0
E21600p	D Net short-term gain from SOCA	12.2	14.6
E21600n	D Net short-term loss from SOCA	12.9	51.3
E22300p	D Net long-term gain from SOCA	15.2	51.4
E22300n	D Net long-term loss from SOCA	11.8	40.4
E22320p	D Net long-term gain from other forms	22.4	62.4
E22320n	D Net long-term loss from other forms	18.8	57.0
E22370	Schedule D Capital gain distributions	13.0	79.3
E01400	Taxable IRA distributions	11.5	27.1
E01700	Pensions and annuities, taxable	14.6	16.4
E27310p	Total rental and royalty, net income	20.4	58.4
E27310n	Total rental and royalty, net loss	12.6	7.1
E26270p	Partnership and S-corp, net income	25.0	52.6
E26270n	Partnership and S-corp, net loss	14.7	42.7
E26500p	Estate and trust, net income	20.5	49.6
E26500n	Estate and trust, net loss	18.1	43.0
E02100p	Farm net income	20.4	32.9
E02100n	Farm net loss	17.5	31.0
E02300	Unemployment compensation	7.7	7.9
E02500	Social Security benefits, taxable	17.5	18.4
E02900	Statutory adjustments, total	5.6	29.2
E04470	Total itemized deductions	14.8	18.5
E04600	Exemptions	13.3	13.3
E04800	Taxable income	7.7	36.4
E09600	Alternative minimum tax	66.2	61.4
E05800	Income tax before credits	7.7	35.7

Source: Statistics of Income Division, special tabulation, and Mathematica.

Table IV.13. Current Sampling Rates and Alternative Sampling Rates by Income Class Based on Optimal Allocations for Different Items

Income Class	Current Sampling Rate (Percent)	Optimal Allocation for E00100	Optimal Allocation for E01000p	Optimal Allocation for E21600n	Optimal Allocation for E22370	Optimal Allocation for E09600
High-income nontaxable	100.00	100.00	100.00	100.00	100.00	100.00
High Schedule C receipts	100.00	100.00	100.00	100.00	100.00	100.00
-$10,000,000 or less	100.00	100.00	100.00	100.00	100.00	100.00
-$9,999,999 to -$5,000,000	100.00	35.76	29.61	100.00	9.71	20.69
-$4,999,999 to -$2,000,000	34.07	15.09	8.75	100.00	8.76	11.91
-$1,999,999 to -$1,000,000	16.08	6.64	3.55	13.74	3.30	5.86
-$999,999 to -$500,000	3.41	3.31	1.71	6.78	1.68	3.00
-$499,999 to -$250,000	0.99	1.62	0.72	3.40	1.24	1.66
-$249,999 to -$120,000	0.51	0.83	0.34	1.54	0.57	0.55
-$119,999 to -$60,000	0.31	0.42	0.14	0.74	0.40	0.25
-$59,999 to -$1	0.19	0.23	0.10	0.22	0.17	0.12
$0 to under $30,000	0.10	0.10	0.10	0.10	0.10	0.10
$30,000 to under $60,000	0.10	0.12	0.10	0.10	0.11	0.10
$60,000 to under $120,000	0.10	0.27	0.21	0.18	0.28	0.30
$120,000 to under $250,000	0.33	0.65	0.70	0.49	1.49	0.98
$250,000 to under $500,000	0.72	1.41	1.83	1.19	1.14	1.83
$500,000 to under $1,000,000	2.48	2.80	4.67	2.70	2.22	3.42
$1,000,000 to under $2,000,000	12.19	5.17	10.93	4.87	4.66	6.91
$2,000,000 to under $5,000,000	32.47	13.59	31.42	10.47	11.58	13.98
$5,000,000 to under $10,000,000	100.00	52.53	100.00	21.95	20.20	32.90
$10,000,000 or more	100.00	100.00	100.00	100.00	100.00	100.00

Source: Mathematica calculations based on using tabulations from the 2008 INSOLE file, Reject 0 records.

Note: Each set of sampling rates reflects a sample of size 252,588 with certainty selection in the specialized strata and the highest positive and negative income strata and a minimum sampling rate of 0.1 percent. The income classes are fixed across the allocations.

Table IV.14. Estimated Coefficients of Variation of Aggregate Dollar Amounts for Selected Items with a New Stratifier and New Index and Alternative Sample Allocations, 2008

		Basis of Sample Allocation						
Item	Description	Using Current Sampling Rates	Optimal Allocation by Item	Optimal Allocation for E00100	Optimal Allocation for E01000p	Optimal Allocation for E21600n	Optimal Allocation for E22370	Optimal Allocation for E09600
E00100	Adjusted gross income less deficit	0.12	0.11	0.11	0.12	0.12	0.12	0.12
E00200	Salaries and wages	0.20	0.17	0.18	0.18	0.19	0.18	0.18
E00300	Taxable interest	0.84	0.68	0.69	0.73	0.75	0.72	0.70
E00600	Ordinary dividends	0.92	0.77	0.79	0.80	0.88	0.84	0.79
E00900p	Business or profession net income	1.01	0.81	0.85	0.89	0.96	0.85	0.84
E00900n	Business or profession net loss	1.70	1.44	1.52	1.86	1.49	1.70	1.66
E01100	Capital gain distributions	4.98	4.03	4.62	4.77	5.01	4.89	4.82
E01000p	D Taxable net gain	0.71	0.56	0.67	0.56	0.81	0.79	0.67
E01000n	D Taxable net loss	0.78	0.65	0.72	0.79	0.76	0.74	0.77
E21600p	D Net short-term gain from SOCA	3.54	3.17	3.67	4.51	3.60	4.07	3.93
E21600n	D Net short-term loss from SOCA	1.15	0.89	1.05	1.37	0.89	1.26	1.15
E22300p	D Net long-term gain from SOCA	1.12	0.92	1.07	0.92	1.31	1.29	1.09
E22300n	D Net long-term loss from SOCA	1.32	1.04	1.17	1.51	1.08	1.46	1.33
E22320p	D Net long-term gain from other forms	1.38	1.05	1.21	1.06	1.39	1.37	1.18
E22320n	D Net long-term loss from other forms	6.69	4.21	5.57	7.87	4.55	6.90	6.36
E22370	Schedule D Capital gain distributions	3.55	1.97	2.17	2.28	2.43	1.97	2.05
E01400	Taxable IRA distributions	1.64	1.17	1.27	1.39	1.40	1.30	1.32
E01700	Pensions and annuities, taxable	0.85	0.65	0.74	0.79	0.80	0.76	0.77
E27310p	Total rental and royalty, net income	1.98	1.50	1.54	1.61	1.70	1.57	1.51
E27310n	Total rental and royalty, net loss	1.50	1.15	1.28	1.65	1.28	1.39	1.46
E26270p	Partnership and S-corp, net income	0.87	0.64	0.71	0.66	0.80	0.74	0.67
E26270n	Partnership and S-corp, net loss	1.07	0.83	0.99	1.31	0.86	1.24	1.10
E26500p	Estate and trust, net income	3.92	2.95	3.24	3.03	3.65	3.40	3.07
E26500n	Estate and trust, net loss	4.49	3.47	5.55	7.14	3.95	7.64	6.20
E02100p	Farm net income	4.12	2.98	3.30	3.31	3.72	3.25	3.07
E02100n	Farm net loss	2.66	2.34	2.41	2.83	2.53	2.55	2.56
E02300	Unemployment compensation	1.36	1.12	1.31	1.35	1.34	1.33	1.35
E02500	Social Security benefits, taxable	0.90	0.70	0.82	0.88	0.88	0.85	0.86
E02900	Statutory adjustments, total	0.84	0.64	0.67	0.72	0.75	0.67	0.67
E04470	Total itemized deductions	0.32	0.28	0.29	0.31	0.31	0.30	0.30
E04600	Exemptions	0.17	0.14	0.16	0.17	0.17	0.17	0.17
E04800	Taxable income	0.15	0.14	0.14	0.14	0.15	0.14	0.14
E09600	Alternative minimum tax	1.13	0.81	0.86	0.88	0.96	0.86	0.81
E05800	Income tax before credits	0.19	0.16	0.16	0.17	0.18	0.17	0.16

Source: Mathematica tabulations of 2008 INSOLE file, Reject 0 records.

Table IV.15. Estimated Increase in Coefficients of Variation of Selected Amounts for an Alternative Sample Design with a New Stratifier and New Index and Optimal Allocations by Item or Based on Adjusted Gross Income or Alternative Minimum Tax, 2008

Item	Description	Percentage Increase in CV if Allocated Optimally by Item	Percentage Increase in CV if Allocated Optimally by E00100	Percentage Increase in CV if Allocated Optimally by E09600
E00100	Adjusted gross income less deficit	24.6	24.6	28.2
E00200	Salaries and wages	6.4	10.1	11.6
E00300	Taxable interest	4.9	6.2	7.3
E00600	Ordinary dividends	16.8	20.0	19.7
E00900p	Business or profession net income	6.6	12.4	10.1
E00900n	Business or profession net loss	10.5	16.8	27.4
E01100	Capital gain distributions	-5.1	8.8	13.4
E01000p	D Taxable net gain	24.8	48.1	48.8
E01000n	D Taxable net loss	-6.4	4.2	11.4
E21600p	D Net short-term gain from SOCA	2.5	18.6	27.1
E21600n	D Net short-term loss from SOCA	17.2	37.9	51.3
E22300p	D Net long-term gain from SOCA	23.9	44.3	47.5
E22300n	D Net long-term loss from SOCA	10.3	24.7	41.6
E22320p	D Net long-term gain from other forms	24.1	42.1	38.3
E22320n	D Net long-term loss from other forms	-1.1	30.8	49.3
E22370	Schedule D Capital gain distributions	-0.7	9.6	3.8
E01400	Taxable IRA distributions	-9.6	-1.5	2.4
E01700	Pensions and annuities, taxable	-10.8	0.8	5.0
E27310p	Total rental and royalty, net income	19.6	22.9	20.7
E27310n	Total rental and royalty, net loss	-17.6	-8.8	3.9
E26270p	Partnership and S-corp, net income	12.4	25.0	17.0
E26270n	Partnership and S-corp, net loss	10.6	32.1	46.9
E26500p	Estate and trust, net income	12.5	23.5	17.3
E26500n	Estate and trust, net loss	10.4	76.7	97.4
E02100p	Farm net income	-3.9	6.5	-1.1
E02100n	Farm net loss	15.2	18.7	25.9
E02300	Unemployment compensation	-11.1	4.1	6.9
E02500	Social Security benefits, taxable	-8.5	7.3	13.5
E02900	Statutory adjustments, total	-1.7	3.4	3.0
E04470	Total itemized deductions	5.4	8.4	11.7
E04600	Exemptions	-6.0	9.3	11.7
E04800	Taxable income	24.1	25.9	27.9
E09600	Alternative minimum tax	15.9	22.5	15.9
E05800	Income tax before credits	13.5	16.8	15.5

Source: Statistics of Income Division, special tabulation, and Mathematica.

Table IV.16. Means and Standard Deviations of AGI: Paper and Electronically Filed Returns by Income Class for the Alternative Sample Design, 2008

Income Class (1991 dollars)	Percent Paper	Mean ($1,000s)		Standard Deviation ($1,000s)	
		Paper	Electronic	Paper	Electronic
-$10,000,000 or less	80.7	-14,499	-12,694	59,766	45,169
-$9,999,999 to -$5,000,000	72.2	-3,931	-4,609	7,354	8,022
-$4,999,999 to -$2,000,000	67.0	-1,429	-1,557	4,127	3,802
-$1,999,999 to -$1,000,000	61.3	-467	-424	2,743	2,513
-$999,999 to -$500,000	58.4	-142	-115	2,121	1,971
-$499,999 to -$250,000	55.7	-16	-24	2,196	2,102
-$249,999 to -$120,000	50.6	15	18	2,007	1,914
-$119,999 to -$60,000	47.2	20	28	1,385	1,352
-$59,999 to -$1	44.7	15	19	1,135	1,108
$0 to under $30,000	32.5	25	28	593	576
$30,000 to under $60,000	33.6	91	93	791	726
$60,000 to under $120,000	40.1	164	168	1,214	1,133
$120,000 to under $250,000	47.7	306	315	1,764	1,751
$250,000 to under $500,000	52.7	613	614	2,041	2,064
$500,000 to under $1,000,000	57.9	1,273	1,230	1,855	1,967
$1,000,000 to under $2,000,000	63.0	2,680	2,620	2,052	2,216
$2,000,000 to under $5,000,000	68.8	5,948	5,888	3,650	3,523
$5,000,000 to under $10,000,000	74.0	14,167	14,027	12,288	6,605
$10,000,000 or more	81.8	57,849	43,458	123,636	66,228
High income nontaxable	69.5	386	326	9,960	2,264
High total receipts	66.2	6,132	4,210	43,678	18,565

Source: Mathematica tabulations of 2008 INSOLE file, Reject 0 records.

Table IV.17. Means and Standard Deviations of Net Capital Gains: Paper and Electronically Filed Returns by Income Class for the Alternative Sample Design, 2008

Income Class (1991 dollars)	Percent Paper	Mean ($1,000s)		Standard Deviation ($1,000s)	
		Paper	Electronic	Paper	Electronic
-$10,000,000 or less	80.7	7,883.6	2,925.4	49,855.4	30,875.3
-$9,999,999 to -$5,000,000	72.2	703.0	777.9	2,092.0	4,200.1
-$4,999,999 to -$2,000,000	67.0	261.1	204.4	873.5	765.1
-$1,999,999 to -$1,000,000	61.3	92.2	91.2	338.1	344.6
-$999,999 to -$500,000	58.4	43.7	34.1	179.1	141.1
-$499,999 to -$250,000	55.7	13.0	10.2	73.3	64.4
-$249,999 to -$120,000	50.6	5.5	3.7	36.6	29.6
-$119,999 to -$60,000	47.2	1.1	0.7	14.3	12.8
-$59,999 to -$1	44.7	-0.6	-0.7	10.7	7.0
$0 to under $30,000	32.5	0.1	0.0	2.0	1.3
$30,000 to under $60,000	33.6	1.0	0.4	7.9	5.0
$60,000 to under $120,000	40.1	5.0	3.1	23.2	18.8
$120,000 to under $250,000	47.7	23.1	15.5	73.7	60.2
$250,000 to under $500,000	52.7	70.6	55.3	184.1	166.2
$500,000 to under $1,000,000	57.9	221.3	194.6	454.0	439.1
$1,000,000 to under $2,000,000	63.0	605.0	559.2	1,047.5	1,048.6
$2,000,000 to under $5,000,000	68.8	1,867.7	1,726.0	3,075.5	2,868.4
$5,000,000 to under $10,000,000	74.0	5,723.8	5,356.9	13,117.9	7,931.9
$10,000,000 or more	81.8	29,112.7	20,102.6	96,939.1	35,771.1
High income nontaxable	69.5	134.5	92.1	2,841.6	1,391.6
High total receipts	66.2	2,411.1	142.7	20,767.9	789.3

Source: Mathematica tabulations of 2008 INSOLE file, Reject 0 records.

Table IV.18. Means and Standard Deviations of AGI: Joint and Non-joint Returns by Income Class for the Alternative Sample Design, 2008

Income Class (1991 dollars)	Percent Joint	Mean ($1,000s) Joint	Mean ($1,000s) Non-joint	Standard Deviation ($1,000s) Joint	Standard Deviation ($1,000s) Non-joint
-$10,000,000 or less	70.6	-13,326	-16,133	60,557	48,276
-$9,999,999 to -$5,000,000	72.1	-4,298	-3,657	7,232	8,303
-$4,999,999 to -$2,000,000	72.6	-1,447	-1,542	3,195	3,160
-$1,999,999 to -$1,000,000	74.2	-444	-468	1,420	1,348
-$999,999 to -$500,000	74.4	-103	-211	703	686
-$499,999 to -$250,000	72.0	-1	-63	345	334
-$249,999 to -$120,000	68.1	30	-11	176	166
-$119,999 to -$60,000	61.8	34	9	92	80
-$59,999 to -$1	37.5	32	9	65	32
$0 to under $30,000	25.5	39	23	21	17
$30,000 to under $60,000	75.8	94	86	25	27
$60,000 to under $120,000	83.5	169	153	55	66
$120,000 to under $250,000	84.3	316	284	134	156
$250,000 to under $500,000	83.9	618	590	292	320
$500,000 to under $1,000,000	82.8	1,261	1,221	582	627
$1,000,000 to under $2,000,000	82.1	2,662	2,637	1,088	1,110
$2,000,000 to under $5,000,000	80.7	5,924	5,949	2,660	3,613
$5,000,000 to under $10,000,000	80.4	14,105	14,235	10,351	13,727
$10,000,000 or more	79.0	53,643	61,205	115,348	115,508
High income nontaxable	68.8	346	416	10,040	1,911
High total receipts	72.3	4,225	8,759	32,789	46,595

Source: Mathematica tabulations of 2008 INSOLE file, Reject 0 records.

Table IV.19. Means and Standard Deviations of Net Capital Gains: Joint and Non-joint Returns by Income Class for the Alternative Sample Design, 2008

Income Class (1991 dollars)	Percent Joint	Mean ($1,000s) Joint	Mean ($1,000s) Non-joint	Standard Deviation ($1,000s) Joint	Standard Deviation ($1,000s) Non-joint
-$10,000,000 or less	70.6	8,341.7	3,514.3	53,679.3	22,797.3
-$9,999,999 to -$5,000,000	72.1	727.8	713.4	2,118.2	4,159.0
-$4,999,999 to -$2,000,000	72.6	257.4	200.4	865.8	760.3
-$1,999,999 to -$1,000,000	74.2	100.0	68.5	355.4	293.3
-$999,999 to -$500,000	74.4	40.7	36.5	161.9	170.8
-$499,999 to -$250,000	72.0	12.1	10.9	70.4	67.2
-$249,999 to -$120,000	68.1	5.5	2.6	35.9	27.1
-$119,999 to -$60,000	61.8	1.0	0.8	13.1	14.2
-$59,999 to -$1	37.5	-0.1	-1.0	12.1	6.3
$0 to under $30,000	25.5	0.1	0.0	1.8	1.5
$30,000 to under $60,000	75.8	0.4	1.1	5.2	8.3
$60,000 to under $120,000	83.5	3.0	8.0	17.8	31.5
$120,000 to under $250,000	84.3	16.3	34.3	59.2	97.5
$250,000 to under $500,000	83.9	55.8	103.5	162.5	231.5
$500,000 to under $1,000,000	82.8	194.3	287.0	426.5	535.0
$1,000,000 to under $2,000,000	82.1	563.5	700.3	1,021.8	1,155.3
$2,000,000 to under $5,000,000	80.7	1,787.6	1,969.1	2,800.9	3,768.0
$5,000,000 to under $10,000,000	80.4	5,566.7	5,882.9	11,289.2	14,520.3
$10,000,000 or more	79.0	27,574.4	27,096.7	94,580.6	64,254.1
High income nontaxable	68.8	128.0	107.4	2,887.5	1,222.0
High total receipts	72.3	881.4	3,633.5	6,379.6	30,462.4

Source: Mathematica tabulations of 2008 INSOLE file, Reject 0 records.

Table IV.20. Estimated Means for Selected Items among Prior Year Returns

Item	Description	Tax Year 2006 Returns in 2007 File	Tax Year 2007 Returns in 2008 File	Tax Year 2008 Returns in 2009 File
E00100	Adjusted gross income less deficit	53,610	77,556	48,806
E00200	Salaries and wages	34,166	43,170	36,575
E00300	Taxable interest	1,915	3,539	1,479
E00600	Ordinary dividends	1,599	3,390	1,333
E00900p	Business or profession net income	4,589	5,973	4,525
E00900n	Business or profession net loss	-994	-1,460	-1,230
E01100	Capital gain distributions	37	64	16
E01000p	D Taxable net gain	9,035	18,903	4,484
E01000n	D Taxable net loss	-124	-137	-185
E21600p	D Net short-term gain from SOCA	393	645	189
E21600n	D Net short-term loss from SOCA	-390	-794	-1,579
E22300p	D Net long-term gain from SOCA	3,059	7,123	2,104
E22300n	D Net long-term loss from SOCA	-503	-693	-1,252
E22320p	D Net long-term gain from other forms	2,943	5,265	1,647
E22320n	D Net long-term loss from other forms	-17	-33	-67
E22370	Schedule D Capital gain distributions	277	662	101
E01400	Taxable IRA distributions	612	913	972
E01700	Pensions and annuities, taxable	2,259	2,161	2,605
E27310p	Total rental and royalty, net income	686	1,048	882
E27310n	Total rental and royalty, net loss	-882	-1,027	-890
E26270p	Partnership and S-corp, net income	5,664	10,423	4,568
E26270n	Partnership and S-corp, net loss	-2,643	-5,069	-2,971
E26500p	Estate and trust, net income	333	294	406
E26500n	Estate and trust, net loss	-65	-73	-66
E02100p	Farm net income	43	42	47
E02100n	Farm net loss	-259	-288	-265
E02300	Unemployment compensation	210	190	334
E02500	Social Security benefits, taxable	513	562	640
E02900	Statutory adjustments, total	903	1,325	939
E04470	Total itemized deductions	9,164	12,171	8,925
E04600	Exemptions	6,481	6,648	7,175
E04800	Taxable income	39,484	62,128	35,763
E09600	Alternative minimum tax	187	326	175
E05800	Income tax before credits	8,409	14,234	7,646

Source: Mathematica tabulations of Individual sample file, 2007 to 2009.

Table IV.21. Estimated Aggregate Amounts for Selected Items among Prior Year Returns

Item	Description	Tax Year 2006 Returns in 2007 File	Tax Year 2007 Returns in 2008 File	Tax Year 2008 Returns in 2009 File
E00100	Adjusted gross income less deficit	174,878	190,777	137,911
E00200	Salaries and wages	111,451	106,192	103,350
E00300	Taxable interest	6,246	8,706	4,179
E00600	Ordinary dividends	5,217	8,339	3,766
E00900p	Business or profession net income	14,969	14,693	12,787
E00900n	Business or profession net loss	-3,243	-3,590	-3,475
E01100	Capital gain distributions	120	158	46
E01000p	D Taxable net gain	29,472	46,498	12,671
E01000n	D Taxable net loss	-406	-337	-522
E21600p	D Net short-term gain from SOCA	1,281	1,587	534
E21600n	D Net short-term loss from SOCA	-1,272	-1,954	-4,461
E22300p	D Net long-term gain from SOCA	9,978	17,520	5,944
E22300n	D Net long-term loss from SOCA	-1,641	-1,705	-3,538
E22320p	D Net long-term gain from other forms	9,598	12,951	4,654
E22320n	D Net long-term loss from other forms	-55	-81	-190
E22370	Schedule D Capital gain distributions	905	1,628	285
E01400	Taxable IRA distributions	1,997	2,246	2,746
E01700	Pensions and annuities, taxable	7,369	5,316	7,362
E27310p	Total rental and royalty, net income	2,237	2,577	2,492
E27310n	Total rental and royalty, net loss	-2,876	-2,527	-2,516
E26270p	Partnership and S-corp, net income	18,476	25,639	12,908
E26270n	Partnership and S-corp, net loss	-8,621	-12,468	-8,396
E26500p	Estate and trust, net income	1,086	724	1,146
E26500n	Estate and trust, net loss	-212	-180	-186
E02100p	Farm net income	141	104	133
E02100n	Farm net loss	-846	-708	-750
E02300	Unemployment compensation	686	468	943
E02500	Social Security benefits, taxable	1,674	1,383	1,809
E02900	Statutory adjustments, total	2,945	3,258	2,652
E04470	Total itemized deductions	29,892	29,940	25,220
E04600	Exemptions	21,142	16,353	20,273
E04800	Taxable income	128,799	152,825	101,056
E09600	Alternative minimum tax	611	803	495
E05800	Income tax before credits	27,431	35,014	21,605

Source: Mathematica tabulations of Individual sample file, 2007 to 2009.

Table IV.22. Distribution of Percentage Error in Advance Estimates, 1996 to 2010

Percentage Error	1996	2000	2005	2010
Total estimates	119	132	154	194
Under 0.5 percent	36	44	32	37
0.5 to < 1 percent	16	24	31	24
1 to < 2 percent	26	19	35	43
2 to < 5 percent	21	23	20	38
5 to < 10 percent	13	15	21	27
10 to < 20 percent	5	5	12	18
20 percent or more	2	2	3	7
Total percent	100.0	100	100	100
Under 0.5 percent	30.3	33.3	20.8	19.1
0.5 to < 1 percent	13.4	18.2	20.1	12.4
1 to < 2 percent	21.8	14.4	22.7	22.2
2 to < 5 percent	17.6	17.4	13.0	19.6
5 to < 10 percent	10.9	11.4	13.6	13.9
10 to < 20 percent	4.2	3.8	7.8	9.3
20 percent or more	1.7	1.5	1.9	3.6

Source: Statistics of Income Division Bulletin and Individual Complete Report, various issues..

Table IV.23. Percentage Error in Advance Estimates of Selected Items, 1996 to 2010

Item	1996	2000	2005	2010
Number of returns, total	**0.24**	**-0.08**	**0.07**	**-0.03**
Adjusted gross income (less deficit)	**0.41**	**-0.55**	**-0.79**	**-0.55**
Salaries and wages:				
Number of returns	0.40	0.17	0.46	0.36
Amount	0.61	0.86	1.57	1.42 *
Taxable interest:				
Number of returns	0.51	0.23	0.13	0.40
Amount	-3.60	-7.02	-13.80	-15.57 **
Tax-exempt interest				
Number of returns	1.09	0.30	-0.62	-0.20
Amount	1.59	0.41	-5.69	-7.91 **
Ordinary dividends:				
Number of returns	0.18	-0.12	-0.27	-0.20 *
Amount	-1.27	-3.25	-7.66	-15.30 **
State income tax refunds:				
Number of returns	0.21	0.40	0.77	1.06 **
Amount	-4.80	-6.68	-5.60	-8.05
Alimony received:				
Number of returns	1.71	-0.61	0.06	0.52
Amount	0.51	-3.66	-0.45	-2.36
Business or profession net income:				
Number of returns	0.08	-0.24	-0.29	-0.18
Amount	-2.48	-3.13	-4.49	-4.74 **
Business or profession net loss:				
Number of returns	0.20	0.14	0.52	-0.84 *
Amount	-5.11	-3.37	-4.29	-5.86
Net capital gain reported on Schedule D				
Number of returns	-0.15	-0.20	-1.25	-2.08 **
Amount	-3.49	-9.03	-12.52	-18.67 **
Capital gain distributions reported on Form 1040:				
Number of returns	-0.50	0.49	0.63	0.12
Amount	-0.58	0.78	0.21	-6.82
Net capital loss:				
Number of returns	0.62	-1.11	-0.28	-0.38
Amount	1.66	-1.22	-0.28	-0.29
Sales of property other than capital assets, net gain:				
Number of returns	0.20	-1.51	-4.13	-6.18 **
Amount	-5.44	-6.79	-11.07	-32.27 **

Continued

Table IV.23 continued

Item	1996	2000	2005	2010
Sales of property other than capital assets, net loss:				
Number of returns	-3.47	-4.36	-7.58	-9.61 **
Amount	-12.05	-12.20	-15.61	-15.32 *
Taxable Individual Retirement Arrangement distributions:				
Number of returns	1.63	1.37	0.80	0.56
Amount	1.34	1.01	-0.51	-0.90
Taxable pensions and annuities:				
Number of returns	1.18	1.39	1.30	1.41
Amount	1.68	1.73	1.88	1.65
Rent and royalty net income:				
Number of returns	-6.63	-0.98	5.00	5.97
Amount	-9.99	-3.28	3.05	2.52
Rent and royalty net loss:				
Number of returns	-3.25	-1.76	12.87	12.96 *
Amount	-5.46	-6.37	9.92	13.31 **
Partnership and S corporation net income:				
Number of returns	-5.04	-7.41	-8.98	-10.43 **
Amount	-7.13	-12.42	-13.65	-14.25 **
Partnership and S corporation net loss:				
Number of returns	-6.27	-7.96	-11.12	-10.30 *
Amount	-26.78	-28.96	-31.25	-33.61 **
Estate and trust net income:				
Number of returns	-3.21	-4.65	-6.67	-7.79 **
Amount	-11.49	-14.32	-16.42	-10.24
Estate and trust net loss:				
Number of returns	-6.25	-4.01	-14.49	-19.82 *
Amount	-45.19	-52.36	-54.29	-42.54
Farm net income:				
Number of returns	1.56	1.71	3.35	3.28 *
Amount	3.31	3.86	8.88	10.80 **
Farm net loss:				
Number of returns	0.08	-0.48	-1.02	-1.75 **
Amount	-1.22	-3.15	-3.92	-2.84
Unemployment compensation: [5]				
Number of returns	0.27	0.67	0.71	1.05 **
Amount	-0.31	0.41	0.50	0.71 **
Taxable Social Security benefits:				
Number of returns	1.55	1.35	1.17	1.12
Amount	1.79	1.83	1.27	1.02

Continued

Table IV.23 continued

Item	1996	2000	2005	2010
Total statutory adjustments:				
Number of returns	**-0.49**	**-0.65**	**-0.50**	**-0.08**
Amount	**-2.14**	**-3.47**	**-4.79**	**-3.41**
Payments to an Individual Retirement Arrangement:				
Number of returns		0.59	1.15	1.92 *
Amount	1.18	0.95	1.70	2.36 *
Student loan interest deduction:				
Number of returns		0.30	0.56	1.67 *
Amount		-0.26	0.46	2.26 *
Health savings account deduction:				
Number of returns			-1.86	-3.61
Amount			-4.29	-4.82
Medical savings account deduction:				
Number of returns		-1.06	-2.22	*
Amount		-0.03	-2.28	*
Moving expenses adjustment:				
Number of returns	-2.33	-0.04	-0.58	0.35
Amount	-5.57	-0.02	-0.69	-1.40
Self-employment tax deduction:				
Number of returns	-0.44	-0.93	-1.12	-1.09 *
Amount	-2.87	-4.10	-6.01	-6.86 **
Self-employed health insurance deduction:				
Number of returns	-2.48	-3.77	-5.91	-6.61 **
Amount	-3.06	-5.49	-8.47	-9.24 **
Payments to a self-employed retirement (Keogh) plan:				
Number of returns	0.20	-2.22	-4.70	-4.41 *
Amount	-1.75	-4.66	-8.75	-9.02 **
Penalty on early withdrawal of savings:				
Number of returns		0.00	-0.83	9.41 *
Amount		-1.15	0.52	501.07
Alimony paid:				
Number of returns	-0.57	-0.82	-0.88	-0.57
Amount	-3.95	-4.73	-4.80	-3.48

Continued

Table IV.23 continued

Item	1996	2000	2005	2010	
Total deductions:					
Number of returns	**0.26**	**0.00**	**0.12**	**0.10**	
Amount	**-0.72**	**-1.80**	**-2.17**	**-1.15**	
Basic standard deduction:					
Number of returns	0.45	0.06	0.32	0.30	
Amount	0.56	0.25	0.57	0.63	*
Additional standard deduction:					
Number of returns	1.23	1.23	1.23	3.71	
Amount	1.27	1.27	1.32	3.40	*
Total itemized deductions (after limitation):					
Number of returns	**-0.20**	**-0.11**	**-0.22**	**-0.30**	
Amount	**-1.69**	**-2.99**	**-3.60**	**-3.99**	**
Itemized deductions in excess of limitation:					
Number of returns	0.07	-0.02	-0.01	3.10	
Amount	-0.95	-4.34	-6.12	103.34	**
Medical and dental expenses deduction:					
Number of returns	-0.82	-0.67	-0.46	-0.76	
Amount	-0.80	-0.97	-1.88	-2.85	**
Taxes paid deduction:					
Number of returns	-0.13	-0.07	-0.18	-0.27	
Amount	-0.62	-1.74	-2.36	-2.81	**
Interest paid deduction:					
Number of returns	-0.16	-0.19	0.80	-0.35	*
Amount	-2.37	-3.62	1.73	-3.05	
Charitable contributions deduction:					
Number of returns	-0.04	0.14	0.00	0.06	
Amount	-2.20	-5.03	-6.20	-7.08	**
Taxable income:					
Number of returns	**0.23**	**0.11**	**0.17**	**0.11**	
Amount	**0.36**	**-0.55**	**-0.87**	**-0.78**	*
Alternative minimum tax:					
Number of returns	-5.04	8.56	1.02	0.03	
Amount	-14.22	-1.19	-8.88	-11.47	
Income tax before credits:					
Number of returns	**0.23**	**0.11**	**0.17**	**0.13**	
Amount	**0.41**	**-0.53**	**-0.77**	**-0.84**	**

Continued

Table IV.23 continued

Item	1996	2000	2005	2010
Total tax credits:				
Number of returns	**0.54**	**0.12**	**0.43**	**0.39**
Amount	**-4.39**	**-2.84**	**-1.84**	**-1.59**
Child care credit:				
Number of returns	0.08	0.01	0.76	1.52 *
Amount	-0.03	-0.09	0.74	1.61 **
Credit for the elderly or disabled:				
Number of returns	1.00	-0.79	-0.39	-6.23
Amount	0.70	-0.46	-2.41	-10.09 *
Child tax credit:				
Number of returns		0.12	0.39	0.63 *
Amount		0.06	0.43	0.66 *
Education tax credits:				
Number of returns		0.65	1.09	1.28 *
Amount		0.71	1.23	1.52 *
Adoption credit:				
Number of returns		-1.80	-1.31	2.45
Amount		-0.60	-5.06	2.07 *
Foreign tax credit:				
Number of returns	0.07	-0.59	-0.45	0.10
Amount	-11.05	-15.83	-10.91	-13.96
General business credit:				
Number of returns	-5.96	-6.48	-10.40	-12.03 **
Amount	-12.69	-16.24	-24.93	-23.30 *
Prior year minimum tax credit:				
Number of returns	-5.92	-6.29	-6.36	-7.32 **
Amount	-4.96	-2.58	-0.08	-3.68
Self-employment tax:				
Number of returns	-0.43	-0.93	-1.12	-1.09 *
Amount	-2.87	-4.10	-6.01	-6.86 **
Total earned income credit (EIC):				
Number of returns	1.31	0.44	1.11	1.49
Amount	**1.91**	**0.72**	**1.61**	**2.30**
EIC used to offset income tax before credits:				
Number of returns	1.25	0.41	0.53	1.42
Amount	1.41	0.36	1.02	2.22
EIC used to offset other taxes:				
Number of returns	1.83	0.42	1.65	1.57
Amount	2.13	0.19	1.90	1.89

Continued

Table IV.23 continued

Item	1996	2000	2005	2010	
Excess EIC, refundable portion:					
Number of returns	1.48	0.56	1.35	1.73	
Amount	1.96	0.80	1.59	2.34	
Additional child tax credit:					
Number of returns		-6.39	0.75		
Amount		-7.27	0.61	1.29	
Total income tax:					
Number of returns	**0.13**	**0.09**	**0.07**	**-0.01**	
Amount	**0.43**	**-0.44**	**-0.70**	**-0.75**	**
Total tax liability:					
Number of returns	**0.12**	**5.27**	**-0.07**	**-0.21**	
Amount	**0.28**	**-0.10**	**-0.95**	**-1.08**	*

Source: Statistics of Income Division, SOI Bulletin and Complete Report.

* Error in 2005 and later exceeds the error prior to 2005, but the error does not grown progressively across all four years.

** Error grows progressively across the four years.

V. RECOMMENDATIONS

The SOI Individual sample continues to serve the needs of its principal customers exceedingly well, more than 25 years after the last redesign. With a relatively recent five-fold increase in the minimum sampling rate, which now covers 92 percent of the return population, and an increased fraction of the population being sampled with certainty, due to significant income growth at the upper end of the distribution, the sample currently provides substantially more precise estimates than it was designed to provide, and supplies users with a very large sample base for simulating a wide range of tax policy options. The sample is larger than it needs to be, however, and while unit editing costs for Individual returns have declined markedly, reallocating some of the SOI Division's resources away from the Individual sample toward other Division needs merits serious consideration at this time.

A. Overview of Recommendations

With regard to the Individual sample itself we recommend the following:

- Continued use of gross positive and negative income in the income stratifier but replacement of some of their current components and addition of one or more other components
- Assessment of whether gross positive income should be replaced by adjusted gross income when the latter is larger
- Revision of the income stratum boundaries to reflect both inflation and real income growth
- Replacement of the current index, which is based on GDP, with one that is based on personal income
- Retention of form type as the second stratifier, with cross-sectional sampling rates by income class undifferentiated across form types except in foreign study years
- Elimination of sub-stratification by degree of interest
- Retention of certainty selection for high-income nontaxable returns if legally required; otherwise, sampling by stratum at enhanced rates sufficient to meet the annual reporting requirements
- Retention of the current minimum sampling rate of 1 in 1,000, which is very popular with the principal customers
- Continued selection of electronic and paper returns at the same rate

- No additional assessment of the merits of selecting sample returns based on both the primary and secondary SSN
- Retention of prior year returns as an integral part of the processing year sample rather than presenting them as representative of late returns
- Reallocation of the sample to maximize efficiency across a wide range of items in light of the increased minimum sampling rate and the substantial growth of income in the upper tail of the distribution
- Retention of certainty selection for returns with high business gross receipts
- Continuation of current procedures for handling misclassification error, which is likely to be reduced by recommended changes in the income stratifier
- Continuation of current procedures for handling missing returns, which are rare and becoming more so

With these recommendations the basic structure of the current design would be retained, but most of its elements would be modified to some degree.

With regard to other aspects of the Individual sample we recommend that the SOI Division:

- Maintain the current release schedule for the final file; there is no particular reason to accelerate delivery, but neither should it be delayed
- Develop as an annual product a person-level database of non-filers, using information returns with CWHS SSNs
- Follow up on customer comments about the declining usefulness of advance estimates, and if this decline is related to an actual deterioration in quality, investigate ways to improve the quality of these estimates
- Follow up on customer comments about the declining value of the SOCA study, due to the decreasing proportion of capital asset sales reported on Schedule D
- Explore whether SOI data show evidence of declining quality in SSNs
- Make available to CDW users the recent comparison of SOI and CDW aggregates
- Consider ways to assess the quality of CDW items that are too rare to estimate precisely with the Individual sample; the SOI Division can make an important contribution here
- Determine how any sample design changes might be reflected in the public use file and communicate this information to the major user of these data
- Ascertain what post-audit data might be available and whether it might be used to provide some sense of what the Individual sample data might look like if it were post-audit
- Develop comprehensive documentation of the sample design to supplement the description provided in the Complete Report

Of these the development of a database of non-filers, is the most significant undertaking but the one that will most enhance the value of SOI Individual data to its principal customers. Such an undertaking should build on the work these customers have already produced. A joint effort would further reduce the demand on SOI resources.

B. Implications of Selected Recommendations

By changing the income definition and the index for income growth, the size of the SOI sample could be reduced by 23 percent and the editing costs reduced by as much as 40 percent. Such large changes, however, would result in a significant loss of precision if the current sampling rates by stratum were retained. A critical element in developing a new design is to determine the optimal allocation of the sample across the newly defined strata. This and some of the other recommendations are discussed further below.

1. Optimal Allocation

With the increase in the minimum sampling rate through expansion of the number of CWHS endings from 2 to 10 and the shift in the income distribution, the optimal allocation under the current design is certain to have changed. If the stratifier and the index are changed as part of a redesign, the optimal allocation is likely to shift even more. Research will be needed to determine the optimal allocation once these other elements are determined. With the knowledge that the homogeneity of strata is susceptible to change over time, the sample allocation exercise should include multiple years of data

Our exploration of a more optimal allocation of the sample suggests that the precision can be improved with changes to the rates. Optimal sample allocation would represent a major component of the development of a new sample design.

2. Income Class Boundaries

Another part of the development of a new design is to specify the income class boundaries in dollars that are current with the implementation. The use of an index will then limit subsequent

growth at the upper end more effectively than the current index. We did not explore the specification of income class boundaries except to note that the boundaries obtained by applying the personal income index to the 1991 boundaries yields strata that in a number of cases may be much too broad. For example, the first positive income stratum would have run from zero to $72,000 in 2008 and to $80,000 in 2012. By the time that a new design is implemented, the ceiling on the first income class using this index will be well over $80,000, and the class would include three-quarters of all returns. In 1991 the first income class included 58 percent of all returns.

Where the boundaries are set will help to determine the size of the new sample, given the fixed minimum and maximum sampling rates. The higher the boundary between sampled and certainty returns, the smaller the certainty sample. Likewise, the higher the upper boundary for selection based solely on CWHS endings, the smaller the fraction of the filing population sampled at higher rates. Unless the intermediate sampling rates are changed appreciably, the placement of the income class boundaries will largely determine the size of the balance of the sample.

The establishment of new income boundaries and the development of new sampling rates should be assessed jointly, as both have implications for the efficiency of the sample.

3. Refining the Income Stratifier

Our analysis explored five specific changes to the income stratifier, but we also found evidence suggesting that other changes might be needed as well. One of the reasons that some returns are reassigned to higher income strata after selection is that their AGI contains large amounts of income from sources that are not included in the current income stratifier. Rather than adding several additional fields, however, it may be more effective to compare gross positive income to AGI and set the stratifier equal to AGI if AGI is larger than gross income.

4. Fixed Shares as an Alternative to Indexing

An alternative to indexing as a way to limit future sample growth is to define the strata so that they represent fixed shares of the population. With this approach the relative sizes of the strata

would not be affected by changes in the composition of the population. The ultimate question, of course, is whether this approach would produce more homogenous or less homogenous strata than an approach that allows for composition change while limiting growth through an index. Answering this question will require empirical analysis, which would be part of a redesign effort. Another issue is whether the fluctuation in the relative sizes of the strata during the lead-up to and aftermath of the Great Recession is a desirable feature that would be lost with a fixed shares approach.

5. Improved Documentation

A brief description of the sample design in the Complete Report publication satisfies the immediate needs of users, but comprehensive documentation of the sample design and sample selection procedures should be available on line. If the sample is redesigned, such documentation should be prepared before the release of the first sample data. After the design has been implemented the documentation should be updated periodically or as necessary. The description in the Complete Report should explain any changes that are not yet reported in the on-line documentation, although SOI staff may find it just as easy to keep the comprehensive documentation up to date once it has been prepared, as changes will be infrequent and relatively minor, presumably.

Items that should be included in the comprehensive documentation are:

- How the sampling strata are constructed, detailing the component variables and how they are used
- Sampling rates
- The sample selection methodology
- What adjustments, if any, are applied to correct for stratum classification errors
- How the processing and editing of electronic and paper returns differ
- A history of the design since its implementation

If there are any new elements introduced with the redesign, they should be covered as well.

If the sample is not redesigned, or if a redesign is not implemented for several years, we recommend the development of a comprehensive description of the current design. The description in Chapter II of this report covers what we view as important. The SOI Division is welcome to draw on this material.

REFERENCES

Bryan, Justin. "High-income Tax Returns for 2009." *SOI Bulletin*, Spring 2012, pp. 6-61.

Czajka, John L., Sharon M. Hirabayashi, Roderick J.A. Little, and Donald B. Rubin. "Projecting from Advance Data Using Propensity Modeling: An Application to Income and Tax Statistics." *Journal of Business and Economic Statistics,* vol. 10, April 1992.

Schirm, Allen L. and John L. Czajka. "Alternative Designs for a Cross-sectional Sample of Individual Tax Returns: The Old and the New." *Proceedings of the Section on Survey Research Methods.* Alexandria, VA: American Statistical Association, 1991, pp. 163-168.

Statistics of Income Division. *Individual Returns 2011.* Washington, DC: Internal Revenue Service, 2013.

Testa, Valerie. "What Would a Sample of HINTS Look Like?" Washington, DC: Statistics of Income Division, no date.

MATHEMATICA
Policy Research

www.mathematica-mpr.com

Improving public well-being by conducting high-quality, objective research and surveys

Princeton, NJ ■ Ann Arbor, MI ■ Cambridge, MA ■ Chicago, IL ■ Oakland, CA ■ Washington, DC