Office of Tax Analysis
Technical Paper 12
May 2023

U.S. Treasury Individual Income Tax Model

Robert Gillette, Siva Anantham
Will Boning, Michael Cooper
Rachel Costello, Julie-Anne Cronin
Portia DeFilippes, John Eiler
Geoff Gee, Kye Lippold
Ithai Lurie, Ankur Patel

OTA Technical Papers is an occasional series of reports on the research, models, and data sets developed to inform and improve Treasury's tax policy analysis. The papers are works in progress and subject to revision. Views and opinions expressed are those of the authors and do not necessarily represent official Treasury positions or policy. *OTA Technical Papers* are distributed in order to document OTA analytic methods and data and invite discussion and suggestions for revision and improvement. Comments are welcome and should be directed to the authors. OTA Papers may be quoted without additional permission.

U.S. Treasury Individual Income Tax Model

Robert Gillette,[1] Siva Anantham
Will Boning, Michael Cooper
Rachel Costello, Julie-Anne Cronin
Portia DeFilippes, John Eiler
Geoff Gee, Kye Lippold
Ithai Lurie, Ankur Patel

May 2023

All authors, U.S. Department of the Treasury

This paper provides a detailed description of Treasury's Individual Income Tax model (ITM). The ITM is the central model used in the analysis of individual tax liabilities under current law and any proposed changes to current law. Building, maintaining and continually improving this model as new data and modeling techniques become available is the task of the Economic Modeling and Computer Applications Staff of the Office of Tax Analysis.

Any taxpayer data used in this research was kept in a secured Treasury or IRS data repository, and all results have been reviewed to ensure that no confidential information is disclosed.

[1] Robert Gillette: Office of Tax Analysis, U.S. Department of the Treasury, Robert.Gillette@treasury.gov

1. INTRODUCTION

The staff of the Treasury Department's Office of Tax Analysis (OTA) has the dual responsibility of forecasting Federal tax receipts and of analyzing the revenue, distribution, and other economic impacts of tax legislation. The analysis of a proposed change in the tax law must answer many questions, such as: how tax revenues will change in future years; how the change is likely to affect the distribution of tax burdens or the distribution of income after tax; how efficiently the change will operate in achieving its objective; and what effect the change is likely to have on private economic decisions. One of the most useful tools developed, maintained, and employed by OTA to estimate and evaluate individual income taxes is the Individual Income Tax Model (ITM).

The ITM is used for many types of analyses including the following:

Estimating Revenue Effects for Proposed Tax Changes. To examine the effects of a single tax provision or set of tax provisions, the ITM simulates two tax regimes (normally referred to as Plan X and Plan Y) that only differ by the provision or set of provisions being examined. Thus, the ITM estimates tax liability and various tax values (such as the number of taxpayers claiming certain itemized deductions and the level of those deductions) under Plan X and Plan Y and the change in tax liability and tax values. Plan X does not need to be current law, so that, for example, the ITM can be used to estimate the additional revenue effect of a particular provision in a multi-provision proposal.

Estimating Interactions Among Provisions. It may be important to estimate the interrelationships between various individual income tax provisions in a proposed tax change or under current law. Because of the interactions between provisions, a revenue estimate for multiple provisions, each estimated separately, will not necessarily add up to the revenue estimate for all provisions estimated simultaneously. For example, changing tax rates will affect the values of deductions, and changing the levels of deductions will affect the amount of

revenue generated by the rate structure. Using the ITM, an analyst can estimate the revenue change from the change in statutory rates alone and from the change in the level of deductions alone and then also estimate the combined effect of changing both statutory rates and the level of deductions.

Sensitivity Analysis. Often, an important consideration in policy analysis is how responsive the tax system is to certain tax parameters. Using the ITM, the analyst can determine the change in tax liabilities due to a change in a tax parameter. The analyst may also be interested in the sensitivity of the tax system to economic variables. For example, the analyst can get a sense of how interest rates affect certain tax variables by adjusting interest income and interest expense items.

Estimates of Average Marginal Tax Rates. One of the most useful attributes of the ITM is its ability to estimate average marginal tax rates. To obtain marginal tax rates, the Federal income tax liability under a given tax regime is calculated for each return on the ITM. Then, the value of an income source is increased by a small amount (usually one percent). The ITM recalculates the tax liability for each return and the change in tax associated with the change in income. The tax and income changes are summed over all returns, and their ratio computed to give the average marginal tax rate for that income source. This gives a rate that is weighted by the amount of the particular income source appearing on each return. Alternatively, the marginal tax rate for each return can be multiplied by that return's sample weight. This provides an average marginal tax rate that is weighted the number of returns (or the number of returns with that particular type of income).

Distributional Analysis. The ITM is also used extensively for distributional analysis.[1] For example, policymakers may be interested in how individual income tax liabilities, and changes in these liabilities are distributed among the population. Standard ITM output includes the distribution of tax liabilities across adjusted gross income (AGI) class for tax units (including

[1] See Cronin (2022) for a discussion of Treasury's Distributional Analysis Methodology and Distribution Model (DM). The DM is an extension of the ITM. The ITM can be used alone for certain types of distributional analysis, such as those based on AGI.

current non-filers). An additional feature of the ITM is that it can associate "parent" returns in the Individual and Sole Proprietorship (INSOLE) sample with the returns of their dependents who file. The dependent filers included directly in the INSOLE are then replaced with these "dependents-of-INSOLE" filers to avoid double-counting. This option allows constructing distributions of tax liabilities across families of associated tax units. Because the ITM uses microdata, the effects of tax reform can be seen on almost any subgroup of the population that is adequately represented in the sample (such as elderly tax units or tax units with children).

Estimating Winners and Losers. A common policy question that arises with any tax reform proposal is who would win and who would lose if the proposal were enacted. A winner is usually defined as a tax return or family that pays less tax, while a loser is a return or family that pays more tax. The model can also estimate the number of returns or families who are removed from the tax rolls (tax liability changes from positive to non-positive) and the number who were previously non-taxable but must pay some tax under the proposal.[2] Standard ITM output includes the number of winning returns, number of losing returns, and the number of unaffected returns. The output also includes the number of tax returns who are either removed or added to the income tax rolls.

Analysis of Other Taxes. The ITM is also used to analyze other taxes, such as the levels and distribution of Federal payroll taxes, and Federal estate and gift taxes.

Special Cross-Tabulations. The ITM's use of microdata also allows cross tabulations to be made on any set of variables before and after a tax change. Frequently, the analyst may want to know how a particular subgroup of the population fares under a tax proposal. The analyst often wants information on the levels of certain variables before and after the proposal (*e.g.*, the level of medical expenses deducted). The ITM is a highly flexible tool for addressing such questions.

The first version of the ITM was built in 1963. Since then, there have been over 20 versions of the ITM with a new version being released approximately every three years. During this period,

[2] The ITM includes a non-filer sample. The methodology for non-filers is discussed in Section 2.

the ITM has been used to analyze and estimate the effects of thousands of proposed changes to the tax code. The staff at OTA continue to improve the ITM as new data and modeling techniques become available. To capture a relatively settled version of the model, this document describes the version of the ITM based on returns for Tax Year 2016 that was used for the Fiscal Year 2021 to Fiscal Year 2023 budgets.[3]

In general, the ITM is comprised of five main components, each of which is discussed in detail in separate sections of this paper. Summaries of each section are given below.

Section 2: Population. To capture the full revenue and distributional effects of current and potential tax policies, the ITM population aims to include all those who could have individual income tax or payroll tax liability or who could receive individual income tax credits. The ITM represents the population with stratified random samples of both tax filers and families or individuals who do not file (non-filers). The filer sample is based on the Individual and Sole Proprietorship (INSOLE) data file prepared by the Statistics of Income Division of the Internal Revenue Service. The non-filer sample is chosen using a sampling design that mirrors the sampling design of the INSOLE. The non-filer sample is reweighted to match the mix of demographic characteristics of non-filers identified in matched Census CPS-tax return data. For the 2016-based ITM, the non-filer sample is then reweighted again to reflect a non-filer count derived from tax data and information provided by the Social Security Administration (SSA).[4]

Section 3: Data. This section discusses the tax and non-tax variables used to estimate tax liabilities. This data includes variables available on the INSOLE file, data directly matched from information returns such as wage splits for joint filers, data directly matched from other sources, such as age from social security records, and data that is imputed or extracted using

[3] More precisely, this document reflects changes made to the 2016-based ITM through calendar year 2022. In 2022, OTA began development of a version of the model based on data from Tax Year 2019, which was used in preparing the Fiscal Year 2024 budget estimates. Several areas of the 2019-based model are in active development as of early 2023, including the population represented, deduction imputations, and health insurance coverage information. This document describes the more stable 2016-based version of the ITM.

[4] The non-filer count targets for future years used in extrapolation are derived from a modified version of the Social Security Area Population (SSAP) forecast from SSA. The 2019-based ITM replaced the use of microdata to estimate non-filer totals for the base year with the base year modified SSAP.

various sources such as itemized deduction levels for non-itemizers.

Section 4: Extrapolation. To estimate the future effects of tax laws and tax law changes as well as to forecast individual income tax receipts over the ten-year budget window, the ITM extrapolates the level and distribution of income, deductions, and other tax variables to future years.

Section 5: Estimating Tax Liabilities. The ITM's tax calculator is a computer program that can simulate changes in tax liabilities arising from changes in any of the underlying variables or parameters. The overall program can be conceptually divided into three steps. In the first step, the program sequentially reads the data file of income tax returns. In the second step, the program uses a set of tax calculators to determine a series of tax values for each return. These values are aggregated and retained for the third step of the program, which prints the values in tabular and/or spreadsheet form. Included in this section are OTA's estimating assumptions for running the model, including assumptions concerning compliance, feedback effects, and behavior.

Section 6 concludes.

2. ITM POPULATION

2.1 The Ideal ITM Population

To support estimates of the revenue effects and distributional impacts of a broad range of individual tax policies, the ideal ITM population would represent everyone liable for Federal individual taxes or eligible for Federal individual credits, as well as their dependents. For analysis of reasonably likely alternative tax policies, the ITM population would ideally also include people who would be liable for Federal individual taxes or eligible for Federal individual credits under those policies.

Common uses of the ITM include people who do not currently file tax returns (non-filers), so

non-filers would be included in the ideal ITM population. Income and payroll taxes are withheld from non-filers' wages. Changes to refundable credits or the standard deduction could change who files, while distributional analysis aims to include all US families regardless of whether they file.

One exhaustive description of the ideal ITM population is:

1. All U.S. citizens, regardless of where they live.
2. All non-citizens, regardless of immigration status, who live on a long-term basis in the U.S.
3. Non-citizens who do not live in the U.S. on a long-term basis but who have U.S.-source income.

2.2 Departures from the Ideal Population

The 2016-based ITM departs from the ideal population to balance providing accurate estimates against data and modeling challenges. While the model captures all people who file or are claimed as dependents on tax returns, the non-filing population used for this version of the ITM is limited to individuals in families who would have been required to file if they had sufficient income and who lived on a long-term basis in the 50 U.S. States or the District of Columbia (plus armed forces abroad). The ITM population thus excludes two groups of non-filers regardless of total income: residents of U.S. territories who do not file Federal income tax returns, and foreigners without Social Security income who have US-source income but do not file Federal income tax returns. These groups are excluded because modeling responses of these non-filers is technically challenging (given their lack of connection to the existing tax system), and their responses are not relevant to most tax reforms (which do not focus on changes to taxation in territories or abroad). Considering how best to model these groups is one of several areas in which OTA is working to improve the ITM.

2.3 Constructing Microdata to Represent the Population

The microdata that represent the population are constructed using the 2016 Individual and Sole

Proprietorship (INSOLE) file to represent filers and a supplemental sample to represent non-filers.

Filers. The 2016-based ITM uses the 2016 INSOLE file to represent filers. This file represents 148.6 million filed returns (including dependent returns) and 290.4 million people.

Non-Filers. As detailed in Appendix I, OTA uses information returns and records of births and deaths provided by the Social Security Administration to IRS to estimate that the total non-filing population for 2016 was approximately 39.5 million people.[5] Demographics for the non-filing population are estimated from U.S. Census linked Current Population Survey - tax data.

As Appendix II explains in greater depth, OTA constructs records on the ITM that represent non-filers from a sample of non-filers who appear on information returns. This sample is then reweighted to match the total count and estimated demographics of the full population of non-filers, including those who do not appear on information returns. OTA identified 95,000 potential non-filing Continuous Work History Sample (CWHS)[6] records. Removing those who would likely not have a filing obligation reduced this sample to 40,000 records. OTA then used 2016 Form 1095 data to construct the tax-unit (family) structure for this sample. However, only 75 percent of the sample had a 2016 Form 1095 and Form 1095 does not always contain information on marital status. As a result, certain records for singles were paired to create additional joint non-filers. In the end, 34,000 unweighted tax units were reweighted to represent the 39.5 million people (31.5 million tax units) that OTA estimates were non-filers in 2016.

2.4 The Target Population Total: Adjusted Social Security Area Population

To account for population growth and demographic change between the base year and each year in the budget window, the ITM uses a target population. The 2016-based ITM target

[5] The 2019-based version of the ITM instead uses an adjusted version of the Social Security Area Population to estimate the number of non-filers in the base year.
[6] The CWHS is a random sample of individuals included in the INSOLE file if they file. Selection is based on the last four digits of their social security number.

population is an adjusted version of the Social Security Area Population (SSAP) forecast provided by the Social Security Administration (SSA).[7] SSA produces population projections for the SSAP[8], which includes everyone living in the 50 U.S. States and the District of Columbia (including U.S. armed forces overseas, who are treated as though they live in the States), U.S. citizens living abroad, non-citizens living in U.S. territories, and non-citizens living abroad who are insured for Social Security benefits. Compared to the ideal population, the SSAP is too narrow because it does not include all non-residents with potential U.S. tax liability. The SSAP is also broader than the population used for the 2016-based ITM because it includes people living in U.S. territories who do not file Federal income tax returns, so OTA adjusts the SSAP to remove them. The population totals by age are used in the extrapolation routine described in Section 4 below.

3. DATA

ITM data includes individual income tax data sent by the taxpayer to the IRS (Form 1040), information returns which include SSNs sent by other entities to the IRS, data linked by SSN from the Social Security Administration (e.g., date of birth and death), and data imputed from other sources such as the Survey of Consumer Finances.

3.1 Base Year Tax Data

As discussed in Section 2, the tax data sample is the 2016 INSOLE, which is also used to prepare the SOI tabulation of tax return statistics, "Individual Tax Returns" or IRS Publication 1304. The ITM weights this sample to represent the filing population in each year.

Generally, the INSOLE is a stratified random sample. At the heart of the sample is a purely random component, known as the Continuous Work History sample (CWHS). The last four digits

[7] The 2014-based version of the ITM used estimates of the U.S. resident population from the Census Bureau, but the SSAP is closer to the ideal ITM population. The 2019-based model in use for the FY2024 budget continues to use the SSAP for extrapolation.

[8] See "The 2016 Annual Report of the Board of Trustees of the Federal Old-Age and Survivors Insurance and Federal Disability Insurance Trust Funds," table V.A2: https://www.ssa.gov/oact/tr/2016/index.html.

of an individual's SSN are randomly assigned by SSA. The CWHS are drawn from 10 unique 4-digit endings. Any individual with one of the ten 4-digit endings is included in the INSOLE sample every year. The sample is then stratified by income and other characteristics to ensure coverage of a broad range of tax situations. Strata that would otherwise have low representation using only the CWHS data are oversampled and the weights for these individuals are lower. In general, the CWHS returns have weights of 1,000 and the highest income individuals have weights of 1 (all are included in the sample). More information on the sample design and underlying data can be found on the SOI website.[9]

3.2 Data and Imputations from Information Returns

The ITM incorporates data from numerous information returns associated with the taxpayers in the INSOLE file. While there are differences in how information return types are incorporated into the ITM, there is a general procedure common to all of them:

1) Information returns are matched to the 1040 using a record identification number assigned to the tax return and related documents by SOI at the return level and SSNs (including ITINs) at the individual level.
2) To the extent aggregate amounts do not match, the 1040 amounts are assumed to represent the true values and the information returns are used to allocate amounts among individuals on the return (such as splits between primary and secondary filers) and also to add details (e.g., how many information returns per person).
3) Where needed, OTA makes ad hoc adjustments, such as dropping duplicates and adjusting or deleting implausible values. The exact techniques used vary by information return.

[9] SOI can be found at https://www.irs.gov/statistics. The 2016 sample documentation is at https://www.irs.gov/pub/irs-soi/16indescofsample.pdf.

Certain information returns are used to create variables that are needed for detailed analysis using the ITM. These include information returns related to earnings, Social Security benefits, retirement benefits and education benefits. Each is discussed below.

Wage Splits. The INSOLE data report the combined wages of both the primary and secondary on a return, whether or not both have earnings. Individual wages are needed to calculate payroll taxes, retirement contributions, and dependent care benefits. The following procedure describes how Form W-2 data is used to allocate total taxable wages reported on Form 1040 (WAS) between wages earned by the primary taxpayer (WASP) and wages earned by the secondary earner (WASS), for returns of married couples filing jointly:

For all returns without a spouse present (i.e., all non-joint returns) WASP=WAS and WASS=0. For notation purposes, the sum of taxable wages earned by the primary as reported on Forms W-2 is W2P; the corresponding value for the spouse is W2S.

W2P and W2S amounts are determined based on SSN matches from the Form 1040 sample to Forms W-2. Any duplicates are dropped, and W-2 fields are aggregated across all unique employee-employer pairings. The two aggregated employer W-2s (per taxpayer) with the highest-wage are retained on model, with all aggregated box details. The majority of 1040 wage dollars (97%) is captured by the top two highest-wage W-2's per taxpayer.

For married filing jointly tax returns, the W2P and W2S amounts are used to determine their wage split. This wage split is applied to the 1040 wage value (WAS).

Most of the 1040 wage dollars not captured by W2P + W2S, 1.7% out of the total of 3%, are from returns with no corresponding Form W-2s found. For these returns, the wage split between the primary and secondary is imputed. The first stage of the imputation is a probit to determine the likelihood of having a secondary worker; this is based on total taxable wages, number of dependents, and age of youngest taxpayers. If a secondary worker is likely, a logit regression determines their ratio of primary versus secondary earnings. For any return that is not married filing jointly or determined to be unlikely to have a secondary earner, all WAS is attributed to the primary taxpayer.

Social Security Benefits. The ITM also uses data available from the SSA-1099 and RRB-1099 information returns to allocate gross Social Security payments (SSINC, which includes both Social Security and Railroad Retirement benefits) into four subcategories for the primary taxpayer, as well as the secondary taxpayer if the tax unit is married and both spouses receive forms 1099. These categories are (1) Social Security retirement benefits, (2) Social Security disability benefits, (3) Railroad Retirement benefits, and (4) Railroad Retirement disability benefits. The following procedure explains how the gross benefits are allocated:

The Form SSA 1099s matched to the Form 1040 sample are slotted into the four categories outlined above, based on their Payer ID and trust fund code. This match accurately captures 94% of SSINC. For 6% of SSINC reported, 1099s exist but do not aggregate to the same total declared on the 1040. For this population, the categories are calibrated to match SSINC while maintaining their respective ratios found on existing 1099s. For the remaining <1%, an imputation is performed, using the 94% of sample of taxpayers successfully matched to 1099s, to determine the probability that a taxpayer has a given benefit, based on AGI and ages.

Retirement Contributions. The ITM uses data available on Form 5498 and Form W-2 (deferred compensation) to maintain information on both the primary and secondary taxpayers' contributions towards individual retirement accounts. This information includes Roth Individual Retirement Account (IRA), Traditional IRA, SEP, Simple, and defined benefit contributions, as well as their fair market values and any rollover amounts. The following outlines the process used to clean this data:

All Forms 5498 and W-2 are matched to the sample based on record identification numbers and SSNs. Duplicate, amended, and corrected returns are cleaned and aggregated by taxpayer and tax return. Form 5498 box details include excess contributions, rollovers, and conversions. There are clearly defined rules on contribution limits depending on the type of account and taxpayer characteristics, such as age. The tax law for the given year is applied to the aggregated data to more accurately capture contributions separate from the other amounts listed above. For example, if the maximum Roth contribution for an individual is $5,500 and the individual reports a contribution of $6,000, the contribution is reduced to $5,500 to account for the

excess contribution reported.

Higher Education. Using a combination of information from the Form 1098-T information return, as well a supplementary information from the National Postsecondary Student Aid Study (NPSAS) and the Integrated Postsecondary Education Data System (IPEDS), the ITM creates information pertaining to a student's estimated tuition and grants related to higher education, undergraduate versus graduate status, class year, and institution information. The process of creating this data is as follows:

The Form 1098-T is matched to the Form 1040 sample based on SSNs, and the top two forms for highest qualified tuition are retained. Qualified tuition being defined as the maximum of payments received and amount billed on a given 1098-T. Using an IPEDS dataset, information pertaining to the institution is merged onto each Form 1098-T. Using a combination of the half-time and graduate indicators on the Form 1098-T, as well as educational tax benefits claimed on the 1040, each student is given a corresponding half-time and graduate status.

For SSNs with evidence of student status on the Form 1040 but missing a Form 1098-T, half-time and graduate status is imputed based on probabilities constructed from NPSAS and tax return characteristics. In addition, institutional information is imputed from the IPEDS dataset. For all students, a class year imputation is performed, based on data provided by NPSAS as well as AGI, dependency status, educational tax benefits claimed, and half-time and graduate indicators.

Lastly, based on information reported on Form 1098-T and on Forms 8863 and 8917 attached to Form 1040, an estimate of qualified tuition and gross grants received is constructed. For those students without a Form 1098-T, an imputation is performed based on the NPSAS data, educational tax benefit claimed, dependency status, AGI, as well as half-time and graduate indicators.

Other Information Returns. The ITM uses information based on the following additional information returns, with minimal cleaning:

- Form 5498-SA: Contributions to a Health Savings Account (HSA), Archer, Medicare & Choice Medical Savings Account (MSA)
- Form 1099-SA/MSA: Distributions from an HSA, Archer MSA, or Medicare Advantage MSA
- Form 1099-Q: Payments from Qualified Education
- Form 1098: Mortgage Interest Statement
- Form 1098-E: Student Loan Interest Statement
- Form 1099-R: Distributions from Pensions, Annuities, Retirement or Profit-Sharing Plans, IRA, Insurance Contracts

3.3 Imputations

Some tax proposals are concerned with demographic, income, or deduction items that are not part of the income tax structure that existed in 2016 or did not appear on particular tax returns for various reasons. Analysis of proposed tax changes concerning these items using the ITM requires imputing missing data to individual tax returns. OTA imputes age, itemized deductions for non-itemizers, health coverage for all units, and wealth for all units. Each imputation is discussed in detail.

Date of Birth, Age and SSN for Work. Many tax provisions and proposals include age requirements or require that an individual or dependent have an SSN that is valid for work. OTA imputes age and checks SSNs using a file that includes all valid SSNs and Taxpayer Identification Numbers, dates of birth, and dates of death that is maintained by the Social Security Administration and provided to IRS.

Imputation of Itemized Deductions to Non-Itemizers. Changes in tax law may lead individuals who do not itemize deductions in the base year to itemize deductions. To model such changes, the following itemized deductions are imputed for non-itemizers:

- Total medical expenses,
- State and local income taxes,
- Real estate taxes,

- Personal property taxes,
- Total home mortgage interest expenses,
- Investment interest expenses,
- Total charitable contributions,
- Fully deductible miscellaneous expenses,
- Miscellaneous expenses that are subject to the 2% AGI floor, and
- Total casualty and theft losses.

A key problem encountered when imputing itemized deductions is the lack of information about the behavior of the non-itemizing population. The methodology employed by OTA assumes that the non-itemizing population behaves no differently than the itemizing population. In addition, the imputation assumes that itemized deductions expressed as a share of non-negative AGI have a multivariate normal distribution truncated at zero. Given these assumptions, a version of the expectation-maximization (EM) algorithm can be used to estimate the levels for non-itemizers. The EM algorithm iterates between forming an expectation of the missing values and then calculating the maximum likelihood estimate of the imputation model given the previously formed expectations. This application of the EM algorithm was originally presented as one technique for imputing missing at-random data and developed by Little and Rubin (1987) in their work on missing data.

There are, however, several additional refinements used for the non-itemizer imputations. First, dependent filers (tax filers who are claimed as a dependent by another tax return) are assumed to have zero expenses and excluded from the imputation process. Second, since not everyone has all forms of itemized deductions, the presence of a specific itemized deduction for a non-itemizer is determined by an indicator vector whose values are determined by an "equivalent" itemizer. A hot deck procedure is used to determine equivalence within an AGI class. Returns from states without an individual income tax are assigned no state and local income tax.[10] Similarly, returns from states without personal property taxes are assigned no

[10] State is determined using the INSOLE state code.

personal property tax.

Next, imputed values of home mortgage interest are overwritten using Form 1098 Mortgage Interest Statements (nonitemizers receive these information returns). Also, the total number of returns with a home mortgage interest expense is reduced to match the number of owner-occupied houses with mortgages reported in the American Housing Survey (AHS).[11] Further, any non-itemizer with home mortgage interest expense is also imputed a real estate tax value. The number of returns imputed a real estate tax value is also restricted so that the total number of returns with real estate taxes, including itemizers, equals the number of owner-occupied houses in the United States as reported in the AHS. The dollar amounts are adjusted to equal the total amount of real estate taxes paid as reported in the AHS.[12] Finally, the age distribution of homeowners and mortgage payers is adjusted to match data reported in the AHS.

The imputed amounts of charitable contributions and state and local income taxes are also adjusted to match aggregate forecasts. In the case of charitable contributions, imputed amounts for non-itemizers are adjusted so that total charitable contributions (itemizer plus non-itemizer) match the amount reported by Giving U.S.A. The imputed amounts of state and local income taxes are adjusted to match values reported by the U.S. Census Bureau's Annual Survey of State and Local Government Finances.

Finally, since the itemized deductions for non-itemizers must (under the assumption that taxpayers optimize) be less than the standard deduction, the total imputed value of itemized deductions is adjusted to fall below the level of the standard deduction. A value between 1/2 of the standard deduction and the full standard deduction is randomly assigned to each non-itemizer. If the initial value of deductions is greater than the standard deduction, then each deduction is reduced, pro-rata, so that the new total equals the assigned value.

[11] Some taxpayers receive a Form 1098 for mortgage interest paid on rental property. Returns reporting mortgage interest expense on Schedule E (rental), Schedule C, or Schedule F were dropped first.
[12] The AHS does not directly report the total amount of real estate taxes paid on owner-occupied housing units. This can be estimated from the AHS table "Monthly Cost Paid for Real Estate Taxes".

Health Insurance Coverage and Premiums. Beginning with tax year 2015, all insurers (including self-insured employers, private insurers, and government entities) are required to report information about each enrollee's health insurance offers and coverage to the enrollee and the IRS using Form 1095 information returns. There are three forms in the 1095 series: Form 1095-A "Health Insurance Marketplace Statement," Form 1095-B "Health Coverage," and Form 1095-C "Employer-Provided Health Insurance Offer and Coverage." Individuals may receive more than one form in this series. OTA uses the information on these forms to assign a type of insurance coverage to each individual on the ITM. The coverage types are Marketplace, uninsured (no indication of coverage on the 1095 series), employer-sponsored coverage, non-group private coverage outside the Marketplace (off-exchange), and Medicare or other public coverage.

Marketplace premiums are imputed using information from Forms 1095-A and 8962 (Premium Tax Credit). Employer sponsored insurance premiums are imputed using months of coverage from Forms 1095-B/C and premium data from Form W-2 (where large employers are required to report the total premium for policies provided to employees each year). Appendix III provides more information on these imputations.

Wealth. OTA imputes wealth to each tax unit on the ITM. This part of the ITM program is referred to as the Wealth Module (WM). The primary uses of the WM are to forecast estate tax liability and to conduct policy analyses involving the estate tax and the taxation of wealth and unrealized capital gains more generally. The WM imputes SOI estate tax data and Survey of Consumer Finance data to the ITM. The SOI estate tax data covers all decedents who had a filing requirement, generally high wealth estates. The SCF data provides wealth information for tax units below the estate tax filing threshold. The WM imputes for each tax return on the ITM values of assets (namely, stocks, bonds, other financial assets, real estate assets (other than home values), business assets, retirement assets (other than IRAs), and other assets), liabilities (namely, mortgage debt and other debt), unrealized capital gains. Home values are imputed by grossing up real estate taxes using local property tax rates. Information returns provide the fair market value of IRAs. Appendix IV provides more information on the WM.

Pass-Through Income. Income for certain business entities, namely partnerships and S-corporations, is passed through the entity and taxed at the individual level. The pass-through entity must file a Form K-1, either a Form 1065 (partnership) or a Form 1120-S (S-corporation) with the IRS. These forms are matched by Employer Identification Number (EIN) and SSN to the tax units on the ITM. For individuals with more than one source of pass-through income, up to 10 partnership and up to 10 S-corporation returns are attached to the individual return. Appendix V provides more information on this imputation.

4. EXTRAPOLATION

To accurately forecast the effects of proposed changes in the tax code, the ITM is extrapolated to future years. Currently, OTA generates an extrapolated database for every year in the budget forecast. The extrapolated databases depend on assumptions about economic growth.

4.1 General Approach

As discussed in Gillette (1989), there are two approaches by which cross-sectional tax data can be extrapolated. The database can be reweighted so that the individual variables on each tax return are adjusted together in order to achieve some set of aggregate and/or distributional targets. Reweighting changes the number of tax returns represented by each return in the database. Alternatively, the individual variables on the database can be adjusted with varying growth factors to match a set of economic or demographic targets.

The reweighting approach has the advantage of preserving significant correlations between the various items on a return, such as the likelihood of owning a home and having retirement income and children. The multiple growth factor approach has the advantage of allowing the composition of income or itemized deductions to vary over time; it also allows the means of particular variables to vary over time.

OTA's current approach is a hybrid of the reweighting and multiple growth factor approaches. For example, when the sample is extrapolated over ten years, a 35-year-old married man with

two children and no sources of income other than wages and interest in the first year (call him record 1 with weight x) still represents a 35-year-old married man with two children and income solely from wages and interest in the tenth year (still record 1 but now with weight y). However, because we allow for differential growth factors, record 1 may have a different ratio of wages to interest in the tenth year relative to the first year.

The sections that follow detail the targets and formal procedure used in the ITM's hybrid approach.

4.2 Targets

OTA typically conducts two major budget exercises per year: the Administration's Annual (or winter) Budget and the Midsession Review. Each budget exercise's window is generally ten years beginning the year after the current fiscal year and the extrapolation is updated for each exercise to account for the latest economic forecast. The Office of Management and Budget (OMB) provides OTA with an economic forecast of income levels, employment, inflation levels, interest rates, and other important macroeconomic variables for each budget exercise.

The ITM also requires forecasts of tax-relevant macroeconomic aggregates not included in OMB's economic assumptions. For example, the ITM requires taxable interest, while the OMB provides total interest. Each target is forecast using the OMB assumptions, population projections from the SSA, and historical data on the target variable and variables in the OMB assumptions.

The tax-relevant aggregate forecasts are called targets, because the ITM's extrapolation procedure is constrained to hit these targets in each year of the forecast. There are 84 targeted variables including:

<u>35 Return (or Individual) Count Targets Including</u>: filing status (4 classes), age (11 classes), total exemptions, elderly exemptions, AGI over $1 million (4 classes), returns with interest income, returns with sole proprietor income, returns with AGI under certain poverty thresholds (3 classes), insurance statuses (6 classes), returns with health savings accounts, returns with

unemployment compensation, and returns with self-employed unemployment compensation.

49 Dollar Amounts Targets Including: total wages, total qualified dividends in AGI, total nonqualified dividends in AGI, total interest, total pensions in AGI, total unemployment insurance compensation, total Social Security income in AGI, total sole proprietorship income, total net positive Schedule E income (4 sources of income), total net negative Schedule E income (4 types of income), total positive capital gains by income class (9 income classes), total modified AGI by income class (15 income classes), total tax exempt interest, total discharge of mortgage debt, total ESI premiums, total non-filer income (5 sources of income) and total self-employed unemployment compensation.

4.3 Procedure

The extrapolation process consists of two stages. The first stage adjusts items on a tax return whereas the second stage adjusts a return's weight to hit aggregate yearly targets based on OMB and other forecasts. Stage I imposes growth factors that reflect per-capita real and inflationary growth on a set of dollar items in the database. Stage II reflects more complicated trends by adjusting the weights of each return to hit target amounts for a selected set of variables, while minimizing a nonlinear loss function where large changes in weights cost disproportionately more than small ones. The net effect of this process ensures that the ITM's extrapolated database is consistent with forecasted aggregate economic measures such as national income, inflation, and population.

In Stage I, dollar variables on the database are adjusted for per-capita growth. The various income, deduction, and credit items on each return are multiplied by factors based on per-capita growth rates estimated from OMB's economic forecasts. We allow differential growth rates for income variables where growth has historically varied across the income distribution. For example, if we expect average wage income to grow by 10% and we expect high-wage returns to capture most of that growth, then we apply different factors across the wage distribution to meet those two beliefs about the future economy. For the 2016-based ITM, we adjust dollar values by one of roughly 40 growth factors.

Following Stage I, Stage II begins by directly adjusting the weights with growth factors reflecting expected changes in return counts by filing status. Stage II then recalculates the ITM's endogenous variables. These variables are transformations or combinations of other variables such that the tax return is internally consistent and consistent with the tax law. For example, after growing individual dollar fields the ITM recalculates adjusted gross income for the tax return. Alternatively, it might limit the contribution to a retirement account, such as a traditional IRA, to reflect that year's tax law.

Stage II then adjusts the sample weights so that all income and demographic targets are simultaneously achieved. The targets, as listed above, are variables likely to impact distribution and revenue estimates. Additional or alternative targets are possible and often discussed. However, OTA research has shown that adding additional targets does not necessarily improve the extrapolation (Gillette 1989). Also, adding more targets will eventually lead to non-convergence of the extrapolation routines. The extrapolation process is akin to inflating a balloon, with the targets forcing the balloon to fit a certain shape. Adding targets leads the balloon to bulge in an unconstrained dimension. These bulges may prove more problematic than the original justification for the additional target.

With several hundred thousand records and a limited number of targets, there is no unique solution to the problem of adjusting the weights to hit the extrapolation targets. The problem amounts to solving a number of linear equations equal to the number of targets, where the number of independent variables equals the number of records in the INSOLE file. Clearly, some criterion must be developed for choosing one of the many solutions. The ITM's criterion is stated in terms of an objective function, whose arguments include the new sample weights. Simply stated, the new weights are chosen to hit the targets while minimizing the objective function.

The convergence algorithm is summarized as follows. Let the targets be represented by t(j) (j=1, ..., N), where N is the number of targets. In addition, let w(i) be the original sample weight (i=1, …, M), where M is the number of returns, let s(i, j) be the amount of the j-th targeted item on

the i-th return and let (x(i)*w(i)) be the new weight. Thus, x is the ratio of the new weight to the old weight.

The problem is to minimize, by choosing the x(i)'s, the objective function,

$$\sum_{i=1}^{M} w(i) * \Phi(x(i)) ,$$

subject to the constraints

$$\sum_{i=1}^{M} \left(x(i) * w(i) * s(i,j) \right) = t(j)$$

for each target t(j) (j=1,...,N).

This is achieved by introducing Lagrange multipliers λ(j) (j=1,...,N) and seeking an extremum for the Lagrangian:

$$P = \sum_{i=1}^{M} w(i) * \Phi(x(i)) - \sum_{j=1}^{N} \lambda(j) * z ,$$

where,

$$z = \sum_{i=1}^{M} \left(x(i) * w(i) * s(i,j) \right) - t(j) .$$

The partial derivatives of P are:

$$\partial P / \partial \lambda(j) = \sum_{i=1}^{M} \left(x(i) * w(i) * s(i,j) \right) - t(j) \qquad (1)$$

(for each j), and

$$\partial P \big/ \partial x(i) = w(i) * \left(\Phi'(x(i)) - \sum_{j=1}^{N} \lambda(j) * s(i,j) \right) \qquad (2)$$

(for each i).

Setting these equal to zero gives a set of simultaneous nonlinear equations, one for each target and one for each return, which can be solved for the x(i)'s and λ (j)'s. The first step in the solution is to use the group of equations to express the x(i)'s in terms of the λ(j)'s.

For each i, setting the partial derivatives in (2) equal to zero implies:

$$\Phi'(x(i)) - \sum_{j=1}^{N} \lambda(j) * s(i,j) = 0 .$$

Solving for x(i) yields:

$$x(i) = (\Phi')^{-1} \left(\sum_{j=1}^{N} \lambda(j) * s(i,j) \right).$$

Setting the partials derivatives in (1) equal to zero and substituting the above values for the x(i)'s results in:

$$\sum_{i=1}^{M} \left((\Phi')^{-1} \left(\sum_{j=1}^{N} \lambda(j) * s(i,j) \right) * w(i) * s(i,j) \right) - t(j) = 0 \qquad (3)$$

(for each j).

This is a set of N equations that can be solved for λ(j). Then the x(i)'s from the equation can be determined given a specific distortion function Φ.

The following distortion function was chosen:

$$\Phi(x) = x^4 + x^{-4} - 2 .$$

There are many possible distortion functions that could be chosen, but functions of this class have several desirable features:

$$\Phi(1) = 0,$$

so there is no distortion when the weight is not changed. Further,

$$\underset{x}{Lim}_{o}\,\Phi(x) = \underset{x}{Lim}_{\infty}\,\Phi(x) = \infty,$$

so weights too far from 1 are not tolerated. Finally,

$$\Phi(x) = \Phi(1/x),$$

so the distortion function is symmetric with respect to both increases and decreases in the weights (i.e., it is as distorting to multiply a weight by 4 as it is to divide it by 4).

Unfortunately, the resulting system of equations is, in general, nonlinear and cannot be solved explicitly. For the particular choice of Φ the system cannot even be written explicitly, as $(\Phi')^{-1}$ cannot be expressed in terms of simple functions. Still, OTA has written a fairly general algorithm for calculating numeric solutions to the distortion equation specified. The user specifies the function Φ, whereupon the machine constructs tables of values for Φ, Φ', and $(\Phi')^{-1}$. Any further reference to Φ is interpreted in terms of the table of values.

The algorithm begins with a set of initial guesses for the λ(j)'s. The initial value of each λ(j) is 1.0. Then the left sides of the equations in (3) are computed, which requires one pass through the file of returns. At the same time, the Jacobean matrix is computed (which shows the effects of small changes in the λ(j)'s on the values of the left sides of the equations). With this information it is possible to select an improved guess for the λ(j)'s for use in the next pass. Normally, five to ten iterations will be sufficient to enable the algorithm to converge on the targets within the limit of accuracy of the computer.

The process of convergence is interesting to follow, since the left sides of the equations

represent the amounts by which the targets will be missed if the process is stopped with the current values of the λ(j)'s. Thus, while the λ(j)'s converge on their correct values, one can watch the entire file converge on its targets. In general, if convergence problems are encountered in applying the algorithm, it is generally the result of the target forecast being inconsistent and alternative targets are chosen

The final output of Stage II is a data file that contains a new return weight for each record on the file. A different file is created for each extrapolation year and for each set of growth factors and targets.

5. THE INDIVIDUAL INCOME TAX MODEL PROGRAM

The ITM consists of tens of thousands of lines of code in Python, Stata, SAS, and Fortran. The core tax calculator is written in Fortran and submitted via Python, while Python (Numpy/Pandas), Stata, and SAS are used to produce the input files for the ITM. Output tables are provided both as text (log files) and Excel spreadsheets. A standard ITM run spanning the budget window (10 years) will take 2 to 3 minutes to run while more complex runs such as those producing marginal tax rate tables can take 5 to 7 minutes.

The sections that follow present the major components of the ITM program. The first section briefly describes the overall design of the ITM program. We then discuss the operating assumptions used in ITM runs and the general limitations of the program.

5.1 ITM Program Design

A typical ITM run compares two tax regimes of tax parameters and code that we label Plan X and Plan Y. Plan X is a baseline chosen for comparison. Typically, this is the tax law currently in effect. Plan Y represents the proposal or alternative that OTA wishes to simulate. For most proposals, the number of differences between Plan X and Plan Y is small in relation to the total number of tax parameters in Plan X. The program takes advantage of this fact by setting the default to be the case where no differences exist between Plan X and Plan Y (i.e., Plan X = Plan

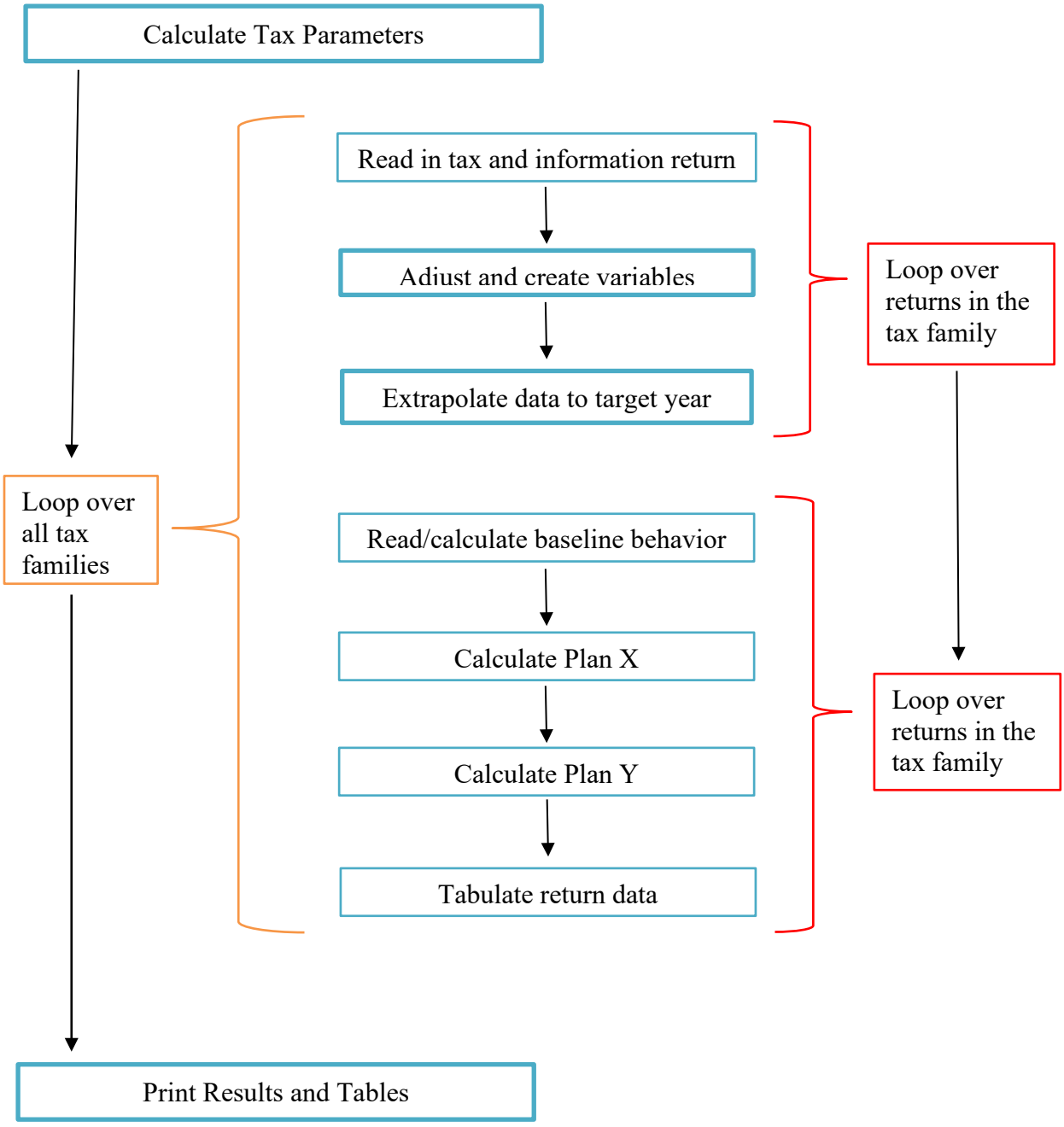
Y). This assumption normally simplifies the model preparation for the user, who has only to specify how Plan Y differs from Plan X.

An ITM flowchart is depicted in Figure 1. On the left is a high-level overview of the process broken into three steps: (1) calculating tax parameters such as rates and income breaks for the relevant years, (2) looping over tax families[13] to process return- and family-level data, and (3) summarizing/printing the simulation's results. Moving to the right, the second step is broken out into more detail. A tax family consists of the primary non-dependent tax return plus any returns filed by dependents on that return. The right side of Figure 1 shows two loops over the family's returns. The first loop is primarily concerned with loading the INSOLE and administrative data that references a particular base year (2016 as of this document) and adjusting it to target a year. The targeting during the first loop assumes a continuation of the tax law as of the base year. The second loop applies data adjustments that reflect behavior from tax law enacted after the base year, simulates a proposal's tax behavior, and tallies results at the return level. Processing a family in two loops allows tax behavior across returns in the family; e.g., choices in the nondependent return affect the dependent returns.

At the heart of the ITM are two versions of the tax calculator, one each for Plan X and Plan Y. Each calculator takes information from each potential tax-filing unit in the data file and calculates that unit's Federal individual income tax liability under the appropriate tax plan. While most fields are taken directly from tax and information returns with an adjustment to move the data from the base to target year, the tax calculator also computes the values of several variables that affect tax liability. Note that the calculator does not endogenously determine total levels of capital gains (or losses), IRA contributions, or Social Security income. The calculator only determines the portion of these items included in AGI.

[13] A tax family consists of all of the family members that file on one tax return.

Figure 1: ITM Flow Diagram



5.2 Modeling Assumptions

To make the modeling tractable, the ITM calculator makes several assumptions. These assumptions are discussed below. Of note, the ITM is only one tool that analysts use to estimate a response to tax changes. Analysts are aware of the ITM assumptions and limitations and can change the assumptions or make off-model adjustments to the ITM output.

Optimization. The ITM assumes all filers choose tax options that minimize their Federal individual income tax liabilities. The model does not include state calculators and does not optimize the sum of state and Federal income taxes. Note that to the extent that certain taxpayers appear eligible but do not claim certain credits under Plan X, the tax model flags these individuals as not taking the credit and would not, generally, switch them under Plan Y. However, analysts may decide to model a change in take-up if appropriate for a specific proposal.

Compliance. The ITM assumes taxpayers are compliant and have correctly reported their income and expenses. The information given by the taxpayer on their tax form is assumed to be accurate and is used by the tax calculator. While the IRS automatically checks for certain errors and the INSOLE file is edited to address inconsistencies within the tax form, the INSOLE is not generally a post-audit file.

Compliance Costs. The ITM assumes that compliance costs are zero. For example, it assumes that taxpayers who take the standard deduction would have total itemized deductions below the standard deduction amount. It does not allow a taxpayer to choose not to itemize because of the added effort required to itemize deductions. This simplifying assumption avoids the challenges involved in modeling the relationship between compliance costs and tax rules. When considering a proposal with significant compliance costs, the analyst may choose to make an "off-model" adjustment to the ITM output to reflect the compliance costs.

Feedback Effects. Estimates for years after the year of the sample data are inherently less reliable than estimates for the year of the data. Given the representativeness of the sample, the estimates will be valid under the assumption that the proposal under analysis has no

significant feedback effects on important variables exogenous to the ITM (*e.g.*, the level and distribution of income and itemized deductions). Clearly, the degree to which data can be assumed exogenous depends on the nature of the proposal, e.g., taxation of all realized capital gains as ordinary income is much more likely to have feedback effects than changes in the zero-bracket amount. Even when feedback is suspected, however, ignoring it is often a useful starting point.

Behavioral Responses. The ITM incorporates "micro" behavioral responses to changes in the tax code but does not include behavioral responses that would change the macroeconomic forecast. It automatically applies tax minimizing choices like the decision to itemize., On occasion, the analyst may explicitly define a behavioral response to a particular change in the tax code. For example, the ITM can use specified price and income elasticities of charitable giving to model how a particular change in the tax code affects charitable contributions. Similarly, estimates using the ITM can use price elasticities to adjust the level of capital gains realizations in response to changes in the tax rate on capital gains. The ITM also considers responses that do not expand or contract GDP, but merely change the level of taxable income. For example, if a proposal increases premiums for employer-sponsored health insurance (which is excluded from taxable wages), the model holds total compensation (wages plus health insurance premiums) fixed but captures the fact that taxable compensation and thus tax revenue will be lower under the proposal.

5.4 Program Limitations

The tax calculator is quite robust and can trace through most interactions between any income source and the various provisions of the tax code. However, it does have limitations. These include:

Certain Missing Calculations. The ITM calculator does not simulate the effects of changes in tax law on several small but important provisions in the tax code. Among these are state and local tax refunds, the foreign tax credit, certain general business credits, and depreciation expenses. The ITM takes the values of these variables as given,

though changes in these variables may be captured by other OTA models or off-model and then provided to the ITM as inputs.

Certain Missing Behavioral Responses. The calculator does not capture changes in individual behavior in response to changes in disposable income or prices. For example, the ITM does not by default capture changes in charitable giving due to changes in tax rates or incomes. Such changes are usually captured by making off-model adjustments, although the ITM is sometimes used to apply a behavioral adjustment on an ad hoc basis.

Imputation Errors. Simulations often require imputed data (data not contained on original tax returns but necessary in computing tax liability under one or more plans). Although every effort is made to correctly impute data items, errors are unavoidable. Within any simulation, variables containing imputed data will interact with other variables in the model. To the extent that imputation errors were made to one or more variables in some systematic way, the results of simulations will be biased.

Missing Dynamics. Many of the imputations for non-itemizers are based on values of variables on tax returns such as AGI. For example, the imputation for state and local income tax deductions is based, in part, on AGI. Therefore, these imputations should be adjusted when the base variables change. For simplicity, such adjustments are not made.

While these limitations are easy to identify, they are difficult to correct within the ITM. These limitations, however, may be corrected much more easily "off model." That is, if the analyst believes the ITM does not accurately portray the effects of a tax law change, they may adjust the output provided by the ITM.

6. CONCLUSION

As discussed in this paper, the ITM is the central model used by the Office of Tax Analysis to

analyze individual tax liabilities under current law and any proposed changes to current law. In addition to the ITM, the office also maintains corporate tax models, a foreign tax model, a depreciation model, tax receipts models, and a distribution model (Cronin 2022). Building, maintaining and continually improving these models as new data and modeling techniques become available is the task of the Economic Modeling and Computer Application (EMCA) Staff of the Office of Tax Analysis.

As tax provisions have expanded to cover more economic situations, the EMCA staff has added modules to the ITM. Modules for health, wealth, education, and retirement have become more sophisticated over time, allowing analysts to better understand the interrelationships among new and existing provisions and to present more detailed analysis.

OTA has also been expanding the ITM to include weighting schemes to represent Federal tax liabilities by state (Fisher and Lin 2015) and to represent tax liabilities by race and ethnicity (Fisher 2023, Cronin, DeFilippes and Fisher 2023). Understanding tax liabilities by state and by race and ethnicity allows OTA to answer a broader array of questions regarding the fairness of tax burdens.

APPENDIX I: ESTIMATING THE SIZE OF THE ITM POPULATION FOR 2016

This appendix describes how OTA estimated the total size and non-filer demographics of the 2016 population used for the 2016-based ITM.

The 2016 Individual and Sole Proprietorship (INSOLE) file represents 148,606,578 filed returns (including dependent returns) and 290,404,266 people (being sure to not double count dependent filers). Only 144,168,019 of these returns were filed for tax year (TY) 2016 or 2015 (representing 282,561,164 people). All the other 4,438,559 returns (and 7,848,102 people) were filed for prior tax years (1999 through 2013). The ITM assumes that these prior year returns represent future late filers for TY 2016. For example, TY 2015 returns filed during processing year (PY) 2017 (when the 2016 INSOLE was compiled) are a proxy for TY 2016 returns that were expected to be filed in PY 2018. Because these prior year returns proxy for future late filers, care is needed to avoid double-counting late filers.

The basis for the ITM's non-filer sample is a database containing the population of information returns.[14] OTA identified Continuous Work History Sample (CWHS) Social Security Numbers (SSNs) with information returns for TY 2016 that were not present on the 2016 INSOLE file. OTA also requires that each non-filer was not present on a 2016 tax return as of the construction of the 2016-based ITM in the spring of 2019 and that each non-filer does not have a recorded date of death prior to 2016. About 36,488,000 non-filing individuals were identified as non-filers using this data and methodology (assuming a weight of 1,000 for each person). A detailed outline of the steps used to construct the 2016 non-filer file is provided in Appendix II.

Combining the INSOLE population with the non-filer population yields a total tax-based population of 326,897,000. This population estimate double counts certain non-filers and omits others (those who do not appear on information returns). The rest of this section will focus on the double count problem that results from having a group of returns on the INSOLE file that proxy for people who file a return in PY 2016 or later.

[14] OTA uses several information returns to construct non-filers: Forms W-2, 1099-SSA, 1099-INT, 1099-DIV, 1099-MISC, etc.

Many of the prior year filers on the 2016 INSOLE also file a TY 2016 return in the 2016 INSOLE. People on prior year returns that do not also file a 2016 return are not double counted because these SSNs are all excluded from the SSN list used to generate the non-filer sample. Essentially, these prior year returns on the 2016 INSOLE represent themselves as future late filers for TY 2016. People on prior year returns that also file a 2016 return cannot represent themselves and must instead represent other non-filers who are expected to eventually file a late return for TY 2016. Because these SSNs were included in the SSN list to generate the non-filer sample, these people/returns represent the portion of the already sampled 36,488,000 non-filers that we expect to eventually file a return in future processing years. As a result, the 36,488,000 non-filers need to be reduced to account for these prior year returns that also file a 2016 tax return on the 2016 INSOLE.

Out of the 4,438,559 prior year returns (7,848,102 people) on the 2016 INSOLE, 3,281,167 also filed a 2016 return (5,752,199 people). This implies that the non-filer sample is too large by 5,752,199 people. This suggests a total ITM population of 321,145,067. To represent the correct population total, the weight applied to non-filer records in the CWHS sample is reduced from 1,000 to 842.

OTA used population data following the method of Heim, Lurie, and Pearce (2017) to obtain upper and lower bounds that provide a check on the estimate of the total 2016 ITM population. This method identifies 318,283,000 people who appear on a 2016 tax return or information return in the population data. This count is a lower bound because it may not contain some dependents or spouses who did not appear on TY 2016 tax returns or information returns as of the construction of the 2016-based ITM in 2019 but who would appear on late-filed TY 2016 tax returns filed after 2019.[15] Upper bound estimates of the number of such dependents and spouses come from late-filed returns from pre-2016 tax years contained in the 2016 INSOLE. There are about 2,445,000 such dependents and 1,080,000 such spouses. These estimates are an upper bound because some such dependents and spouses appear on 2016 information

[15] As was the case with the INSOLE-based population counts outlined above, the CDW-based methodology also misses people who do not appear on tax returns or information returns.

returns and thus are already included in the lower bound count. Adding these additional dependents and spouses to the lower bound gives an upper bound for the total population appearing on tax returns or information returns of 321,808,000 people. The total 2016 ITM population estimate of 321,245,067 is between the upper and lower bound estimates obtained from this method.

APPENDIX II: IDENTIFYING AND CONSTRUCTING 2016 NON-FILERS

The 2016-based version of the ITM models a non-filer population that consists of individuals who did not file an individual income tax return but who would have had an individual income tax filing obligation if they had sufficient income.[16] Many, but not all, such individuals have income and payroll tax withheld from wages reported on Form W-2. The 2016-based model creates this non-filer population in four steps. The first step assembles a pool of potential non-filing individuals to represent all non-filers in 2016. The second step narrows that pool to exclude those who may be deceased, may have filed after the INSOLE sample of filers was constructed, and who should not appear in the non-filer population for other similar reasons. In the third step, the pool of representative individuals is formed into tax return units (adding spouses and dependents where appropriate). The final step weights the sample to represents the estimated demographics and total count of non-filers.

Step 1: Identifying Potential Non-Filers. The Continuous Work History (CWHS) sample forms the purely random component of the tax filer sample (INSOLE file). The non-filer sample follows the same random design.

CWHS individuals include anyone with one of 1 of 10 4-digit SSN endings. For example, for SSN xxx-xx-yyyy, yyyy is the 4-digit ending. There are 10 unique, unvarying values for yyyy that are followed over time and make up the CWHS sample. Anytime a CWHS individual files a return in any year, they are included in the INSOLE. Because the last four digits of an SSN are randomly assigned, the CWHS records form a purely random sample, each record represents 1,000 filers; has a weight of 1,000 (10 of 9,999 possible 4-digit endings, since SSA does not assign 0000 endings). [17]

[16] The 2016-based ITM also limits the non-filer population to those living on a long-term basis within the 50 U.S. States and District of Columbia or in the armed forces abroad. The scope of the population of non-filers to model is an area of active model development in the 2019-based ITM.

[17] The INSOLE file also includes an oversample of certain low probability returns, such as those with high income. This is done to improve the estimates of tax provisions which would otherwise have too small a sample from which to get a reliable estimate. Thus, the INSOLE is a stratified random sample. Certain low probability strata are over oversampled. The sample of taxpayers with 1 of the CWHS ending digits forms the base and random oversampling fills out each strata to achieve an efficient number of unweighted returns in each strata.

All potential members of the non-filer sample will likewise have 1 of the 10 4-digit SSN endings. Thus, the sample that represents non-filers uses the same sampling design as the sample that represents filers. If an individual has one of the CWHS ending digits and interacts with the tax system either by filing an individual income tax return, or by receiving or being present on an individual income tax return or an information return, then they will be part of the ITM sample.[18]

To find CWHS records that did not file but might have had a filing obligation if they had sufficient income, OTA assembles a pool of records for all ***potential*** non-filers gleaned from two data sources:

1) SOI-provided information returns for those with CWHS ending digits and,

2) A TY 2016 file constructed by OTA using the population of Forms 1095.

The first source contains 307,000 records for approximately 92,000 unique SSNs that were not associated with an individual income tax return filed for 2016. Over 50 types of information returns were included in the information returns population data from SOI in 2016 and all were used to search for potential non-filers. Most records meeting the criteria were from the following types of information returns: W-2 Wage and Tax Statements, 1095-B Health Coverage, 1099-B Proceeds from Broker and Barter Exchanges, SSA-1099 Social Security Benefit Statement, 1099-INT Interest Income, and 1099-R Distributions from Pensions, Annuities, IRA.

The second data source, the population of Forms 1095, identifies CWHS individuals who were not the recipient of a Form 1095 (health insurance coverage form) but were listed as covered on a Form 1095 received by someone who was not in the CWHS.[19] This source identifies an additional 3,000 potential non-filers.

[18] The ITM represents people who do not appear on information returns by increasing the weights applied to non-filers who appear on information returns to represent the full estimated total population of non-filers.
[19] The population data also contain some Forms 1095 provided to CWHS members that do not appear in the SOI-provided information returns file, perhaps because the 1095s arrived after the SOI file was constructed.

In total, the initial pool contains around 95,000 potential non-filing individuals.

Step 2: Refining the Sample. OTA narrows the pool of potential non-filing individuals to those likely to meet the population definition used for the 2016-based ITM. The reasons individuals are removed from the pool and approximate counts are:

1. Drop those who appear as a primary or secondary filer on a Tax Year 2016 Form 1040 in the population data as of the construction of the 2016-based ITM in calendar year 2019. These are late filers captured by the population data but not the INSOLE (4,000 records).

2. Drop those who appear as dependents on a Tax Year 2016 Form 1040 in the population data as of calendar year 2019 (35,000 records).

3. Drop those whose SSNs are not verified in information on SSNs, dates of birth, and dates of death provided to the IRS by the SSA (2,000 records).

4. Drop those with a date of death before 2016 (5,000 records).

5. Drop those whose date of birth is after 2016, who were older than 105 years, or who are missing a date of birth (less than 500 records)

6. Drop those whose date of birth indicates they are younger than 19 (4,000 eliminations)

7. Drop those who appear to be foreign residents (appear on Form 1042-S or 8288-A) (less than 500 records).

8. Drop those who appear to be residents of Puerto Rico, the U.S. Virgin Islands or another U.S. Territory (2,000 records)

9. Other (2,000 records)

After these refinements, the sample contains 40,000 records.

Step 3: Forming Tax Units. OTA uses Forms 1095 (A, B, and C) to form tax units (add secondary filers and children). Forms 1095 provide SSNs and names of covered family members. This

matches approximately 75% of the non-filer records to a Form 1095, and thereby approximates filing status (single, joint, or head of household) and the number of dependents. OTA also attaches information returns, such as Form 1098-T information forms, using the population data and ages from the file provided by the SSA to the attached secondaries and dependents. This information is used to exclude people who appear on the same Form 1095 but would not be dependents for tax purposes (such as children age 19-25 who are not students).

Unfortunately, some Forms 1095 do not show spouses (such as those filed for Medicare recipients), and 10,000 individuals in the non-filer sample do not have Form 1095s. OTA imputes joint filers to a fraction of these returns by "marrying" individuals, assigning a share of single returns to instead be spouses on other returns.

After adjusting the totals for these "marriages", there are about 34,000 total non-filer tax units.

Step 4: Targets and Weights. As described in Appendix I, OTA's estimate of the total 2016 ITM population contains 39.5 million people and 31.5 million families not represented by the 2016 INSOLE. The demographics of these non-filers were estimated from U.S. Census Current Population Survey data linked to tax returns. From these estimates, the non-filer population contains 1.8 children (ages 0 to 18) and 2.0 million dependents of any age. Even after "marrying" some individuals, the non-filer sample contained too few joint and head of household non-filers and too many single non-filers relative to the estimated demographics and total ITM population. OTA corrects this imbalance by multiplying the weight on joint and head of household tax units by 1.5 and the weight on single tax units by 0.9.

Table A1 shows counts of unweighted non-filer records and weighted non-filer tax units, non-filer dependents, and non-filer primary and secondary filers over the age of 65 by filing status and dependents, by filing status and age, and by family size. Most non-filers are single. Few have dependents. This pattern is consistent with the greater incentives to file (even when not required to do so) that much larger refundable credits provide to families with children than to low-income households without children. Around a third of non-filer tax units have a primary filer over age 65, consistent with lower income among retirees and the exclusion of some Social

Security income from filing requirements.

Table A1 Non-filers in 2016 ITM	Unweighted Non-Filer Records	Non-filer units (weighted) (millions)	Non-filer Dependents (weighted) (millions)	Non-filer Primaries Age 65+ (weighted) (millions)	Non-filer Secondaries Age 65+ (weighted) (millions)
All	34,000	31.5	2.0	11.0	1.7
By Filing Status and Dependents					
Single, no dep.	27,000	24.5	0	9.3	0
Joint, no dep.	6,000	5.7	0	1.7	1.7
Joint, with dep.	*	0.4	0.7	**	**
HoH, with dep.	1,000	0.9	1.2	**	0
By Filing Status and Age					
a Single, <65	17,000	15.2	0	0	0
b Single, 65+	10,000	9.3	0	9.3	0
c Joint, both <65	4,000	4.1	0.7	0	0
d Joint, one 65+	1,000	0.6	**	.3	.3
e Joint, both 65+	1,000	1.4	**	1.4	1.4
f HoH, <65	1,000	0.8	1.2	0	0
g HoH, 65+	*	**	**	**	0
By Family Size					
Family size 1	27,000	24.5	0	9.3	0
Family size 2	6,000	6.2	0.6	1.7	1.7
Family size 3+	*	0.4	0.6	**	**

**Less than 500 unweighted*
***Less than 50,000 weighted*
"HoH" is "Head of Household" filing status

APPENDIX III: HEALTH MODULE

The health module of the ITM estimates health insurance coverage to model a tax unit's changes in tax liability arising from changes to health-related tax provisions, such as the Premium Tax Credit (PTC) for Marketplace coverage and the exclusion of employer-sponsored insurance (ESI) from taxable compensation.

Coverage Assignments

Beginning with tax year 2015, all insurers (including self-insured employers, private insurers, and government entities) are required to report months of health insurance coverage to enrollees and the IRS using Form 1095. Although the 2016 ITM mostly includes people who filed for tax year 2016, the base data for the model uses all taxpayers who filed in *processing* year 2017, some of whom filed for earlier years. OTA used Form 1095 coverage information from 2016 for taxpayers who filed tax year 2016 returns, and year 2015 coverage information for all other taxpayers. The coverage of the tax units for the model is merged from OTA's population-level health insurance coverage data, as described by Lurie and Pearce (2021).

OTA's measure of coverage is based on a hierarchical assignment of coverage on a monthly basis. We used the following hierarchy: Marketplace, employer-sponsored coverage, non-group private coverage outside the Marketplace (off-exchange), public coverage (Medicare, Medicaid, VA healthcare, and other public sources), and uninsured (no indication of 1095 coverage). As noted by Lurie and Pearce (2021), the number of uninsured people using the Form 1095 measure alone is much higher than in survey data. However, there is an indicator on the Form 1040 for 2016 where some people without Form 1095 coverage report having full year coverage (to claim an exemption from the individual mandate penalty that was in effect at that time). Instead of imputing coverage to those individuals to try and match survey information, we assign those people into a separate coverage group called "unverified" coverage. Not having third-party reported information about coverage for those people suggests that even if they do truly have coverage, it is somewhat different from people who have coverage reported on a 1095. The unverified coverage group represents about 26 million person-years of coverage in

2016. Other coverage totals include 156 million in ESI, 7.6 million in off-exchange coverage, 101 million in public, 25.8 million uninsured and 9.9 million people in marketplace coverage.

It is important to note that Form 1095s are sent for what we like to refer to as "major medical" coverage. For example, 1095 information is not sent for retiree coverage that is secondary to Medicare (also known as a "wraparound" policy), or short-term limited duration (STLD) plans. OTA does not impute these types of policies unless necessary for specific proposals.

Premium Imputations

Marketplace Premiums. To compute the PTC, OTA uses Form 1095A and Form 8962 to calculate the monthly Marketplace premiums and monthly Second Lowest-Cost Silver Plan (SLCSP) associated with each return. When Form 8962 data was available, OTA used that information first. The PTC uses an income concept of Modified Adjusted Gross Income (MAGI) relative to the Federal Poverty Level (FPL), which we computed using Form 1040 income and household size information.

The Advance Premium Tax Credit (APTC) depends on the "Advance MAGI" taxpayers reported to the Marketplace as their projected income for the year. This information is not available on tax returns, but one can calculate the implied income reported by using the APTC and SLCSP taxpayers received (from the 1095A) and the number of people on the return (assumed to be the same at Marketplace signup and on the 1040). However, this imputation is not exact--a taxpayer's family size may change during the year, and taxpayers with different values of Advance MAGI may qualify for the same APTC in certain cases. Consistent with this, OTA identified some taxpayers with imputed Advance MAGI below 100% of FPL in states that did not expand Medicaid and below 138% of FPL in expansion states, most of whom would not have received the APTC if the imputed Advance MAGI were correct. OTA edited the imputed Advance MAGI variable to ensure the MAGI to FPL ratio is in the right range.

ESI Premiums. As noted earlier, large employers are required to report the total premium for ESI policies provided to employees each year on Form W-2. However, because some employees begin or leave employment mid-year, the Form W-2 reports premiums that in many cases do

not represent a full-year premium. Also, reported premiums could be for the worker alone or for a policy with family members. Using the coverage information from Form 1095 for ESI policy holders, OTA can observe the number of people covered by the policy and the number of months the policyholder held the policy. Using that information combined with W-2 premiums enables OTA to calculate an annualized measure of premiums for W-2s. As shown by Lurie and Miller (2022), the shape of the premium distribution by family size from the tax data closely matches the premium distribution from the MEPS-IC, which helps validate the tax-based premiums. OTA then used tax-based premiums to impute ESI premiums to policies without W-2 premium information (using a random draw from the premium distribution by family size).

One issue OTA has encountered is that the total of implied ESI premiums from tax data is different from total premiums in the MEPS-IC. This is not surprising because the two data sources include different types of ESI premiums. Accounting for MEPS-IC not having federal policies (and Tricare) and OTA not having "wraparound" policies suggest that OTA and MEPS-IC totals are fairly close.

Extrapolation

The steps above result in coverage and premium assignments for 2016 data. The ITM extrapolates the 2016 data through the budget window by reweighting returns and adjusting dollar values on each return to hit targets based on OMB's economic forecasts. Targeted counts include the number of tax-filing and non-filing units by filing status and by income relative to the poverty level; we assume that the ratio of advance to final MAGI for each return remains constant throughout the budget window. In addition, targets for the number of subsidized and unsubsidized persons in the Marketplace were obtained from the Department of Health and Human Services' Office of the Actuary (OACT). Similarly, targets for the number of people with ESI coverage, non-group coverage outside of the Marketplace, and uninsured status were constructed from National Health Expenditure Accounts (NHE) growth rates for these categories. For premium growth, OTA used OACT projected growth rates for Marketplace premiums and NHE projections of total ESI premiums to grow reported 2016 premiums to future levels. Premium tax credit estimates for the Budget period are obtained using the

projected distribution of returns by coverage, premiums, and incomes.

Estimating Alternative Health Policies

When modeling changes to ESI or Marketplace subsidies, it is important to consider how taxpayers will change their choice of coverage in response to the policy. For example, the expansion of PTC subsidies in the American Rescue Plan was projected to cause some uninsured individuals to start buying coverage on the Marketplace. For the 2016 base year ITM, OTA has used its Simulation of the Health Insurance Market (SHIM) model to project how the distribution of coverage by income will change in response to policies, and then impute corresponding changes in coverage status for people in the ITM. The model assigns new Marketplace enrollees a premium (equal to the SLCSP value for their location, age and family size), computes their MAGI to FPL ratio (assuming advance MAGI will be equal to final MAGI), and computes total PTC spending for the new enrollees. Similarly, if individuals leave ESI coverage, the model captures the long-run shift of the value of their ESI exclusion from nontaxable to taxable compensation. These capabilities help the ITM provide accurate budgetary estimates of changes to health provisions in the tax code.

APPENDIX IV: WEALTH MODULE

OTA's Wealth Module (WM) imputes wealth to individuals on the ITM. The primary uses of imputed wealth values are to forecast estate tax liability and to conduct policy analyses involving the estate tax. Because, in general, estate tax returns are filed only by estates above the filing threshold, they are of limited use in estimating proposals that involve changing the filing threshold. The WM provides estimates of wealth for taxpayers below the estate tax filing threshold, which directly addresses this shortcoming.

Briefly, the WM imputes for each tax return on the ITM values of assets (namely, stocks, bonds, other financial assets, real estate assets, business assets, retirement assets, and other assets), liabilities (namely, mortgage debt and other debt), unrealized capital gains (namely for home values, stocks, and business assets) and the holding periods associated with these assets and liabilities.

The remainder of this document describes the primary data sources and imputation methodologies used in the construction of the WM and describes how the various data files were merged and extrapolated. The WM for the 2016-based model described in this documentation was based on return information for tax years 2013 and 2014.

Primary Data Sources

The WM's estimates combine data from the following core data sets:

- INSOLE2013 – The WM is based on estate tax returns for individuals who died in 2014. The tax year 2013 Income and Sole Proprietorship (INSOLE) file from the Statistics of Income Division (SOI) of the Internal Revenue Service (IRS) provided the core sample of taxpayers on which the WM is based because it is the last tax year for which individuals who died in 2014 report a full year of income.

- Real estate assets were derived directly from INSOLE data by capitalizing property taxes paid (found on Schedule A of Form 1040 for individuals who itemize their deductions) using ZIP code-level property tax rates. These property tax rates were obtained from the 2008 American Community Survey, produced by the U.S. Census Bureau. Property taxes for non-itemizers were imputed using an expectation-maximization (EM) algorithm. Roughly, the shares of income that itemizers spent on itemized expenses such as property taxes and charitable contributions were calculated, and a covariance matrix of these shares was constructed. This covariance structure was then applied to non-itemizers to impute their values of property taxes paid. (A match randomly chosen from an itemizing near neighbor return was used to determine the presence of a particular deduction.) The algorithm iterated until aggregate targets for property taxes paid from the U.S. Census Bureau's 2013 American Housing Survey were reached.

- Information on defined contribution plan assets was obtained directly from the ITM. IRA assets were obtained by matching tax records on the ITM to the Form 5498 "IRA Contribution Information". Assets for other types of defined contribution plans were imputed using information on contributions from Form W-2 and other sources.

- Information on defined benefit plan benefits was obtained directly from the ITM, which in turn was imputed using accrued retirement benefits reported in the 2014 Survey of Consumer Finances (SCF); Form 5500, Annual Return/Report of Employee Benefit Plan, data; the Department of Labor's Private Pension Plan Bulletin Abstract (2014); and various publications related to pension benefits for government employees. The flow of benefits obtained from these sources was converted into defined benefit plan assets by using the U.S. Treasury's Second Segment Rate in a present-discounted-value calculation.

- ESTATE2014_SOI – The 2014 estate tax sample from SOI was used to provide data on total assets and liabilities, as well as asset components (other than for real estate and retirement assets) for returns that could be matched to the INSOLE2013.

- ESTATE2014_POPULATION – This dataset, containing the entire population of estate tax filers who died in 2014, was used to provide data on total assets and liabilities for additional returns to be matched to the INSOLE2013, the last complete year tax return. (Detailed asset data is not available in the population file.)

- DECEDENTS (2014) – This dataset, the population of 2014 decedents from the Social Security Administration, was matched with INSOLE and estate tax returns to develop a model that predicted, conditional on income, age, and marital status, the probability of each living income taxpayer's having wealth above or below $5 million.

- SCF2013 – The 2013 SCF was used to impute asset and liability components (other than for real estate and retirement assets) for those taxpayers with imputed wealth below $5 million. It was also used to impute basis for selected assets from both the SCF and on estate tax returns.

- MORTALITY RATE – In order to calculate unrealized gains held at death and expected estate and gift tax revenues, mortality rates were required. Mortality rates based on age, sex, and income were estimated by merging the SSA Decedent data with the population of tax returns (including non-filers) to create the Mortality Rate dataset. The average mortality rate was estimated for 72 cells: 3 income groups by 2 sexes by 12 age groups.

Imputation Method

There are a number of methods for imputing values from one dataset to another, each with certain advantages and disadvantages. Except where noted, the WM uses hot decking or a closely related method called predictive mean matching (PMM) to impute values.

The basic hot decking method begins by using available categorical variables to group all observations in both the donor and recipient datasets into discrete cells. Any continuous variables used to create cells must first be converted into discrete categories. Next, each recipient observation is randomly matched to a donor observation in the same cell, after which all values to be imputed for that recipient observation are imputed from the same donor.

PMM is a more general case of the hot decking imputation. In this case, some or all of the variables used to create the cells are instead used in a regression. The predicted values from the regression are then used within each cell to match recipients and donors. Generally, the match is not the closest one, but rather is randomly chosen from a fixed number of the closest matches (generally known as k-nearest neighbors). While regression was used as part of the PMM process, the predicted values were only used for matching purposes, and never to directly impute values.

Choosing between a hot deck or PMM method at different stages of the WM creation generally depends on practical issues. Using the PMM version of Stata's "mi" (multiple imputation) command is fairly automated and easy to program, but it also has some tradeoffs. If a large number of cells are used, one might end up with cells containing too few donors. But if a smaller number of cells is used, the speed of computation is unfeasibly long. Conversely, pure hot decking must be manually programmed in Stata, but it is generally fast and robust. Hence, PMM was generally used if it worked well and quickly; otherwise, pure hot decking was used.

In contrast to hot decking or PMM that uses regression analysis to match recipients and donors, pure regression-based methods use regression results from the donor dataset to directly impute values in the recipient dataset. The primary advantage of pure regression-based methods is that they are straightforward to apply.

Nevertheless, hot decking (and PMM) methods were the main methods used to construct the WM because they have two key advantages. The first is that they only use actual values from the donor data set. For example, most of the imputed variables represent dollar values that are non-negative and not excessively large. In other words, with hot decking, one is constrained to imputing reasonable values as only those are contained in the donor set. With pure regression-based techniques, it is common to get negative or unreasonably large imputed values.

Second, most of the imputations involved simultaneous imputations of multiple values from a donor observation to a recipient observation. Hot decking preserves the covariance structure of the imputed data. Conversely, with regression methods it is necessary to perform sequential

imputation of the imputed variables, making it much more difficult to preserve the covariance structure.

The main disadvantage of hot decking (and PMM) methods in our imputation was that the donor data set was smaller than the recipient data set. As a consequence, donor cells had to be reused.[20]

Imputation Procedure

The first step in the WM is to create a single data set containing wealth values for individual taxpayers across the wealth distribution. Estate tax returns are used for taxpayers with wealth in excess of the filing thresholds and the Survey of Consumer Finances is used for taxpayers with wealth below the filing thresholds.

When available, the estate tax data are generally preferred to SCF data. The former represents reported tax values, and taxpayers are legally liable for any false claims. In contrast, the latter contains data that are voluntarily provided, and there are no penalties for misrepresentation. While estate taxpayers do have an incentive to minimize asset values, the primary purpose of the WM is to calculate estate and gift tax revenues, which will generally reflect those reported values.

Estate tax returns are required to be filed for all decedents with gross estates above $5.34 million in 2014. Estate tax data are also available for many decedents with gross estates between $3 million and $5.34 million, as those estates sometimes file returns due to portability of the spousal exemption[21] and Generation-Skipping Tax (GST) elections. Additionally, returns are required to be filed for some decedents with a gross estate below $5.34 million because they have made gifts prior to death. Nevertheless, because coverage below the $5.34 million

[20] See https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3130338/ for more details about hot decking and predictive mean matching.

[21] Since 2010, the portability rule allows any unused lifetime estate and gift tax exemption of a deceased spouse to be transferred to the surviving spouse, ensuring it isn't lost. While predicted values based on the SCF were not compared with the actual values from the estate tax returns because of sampling issues, the resulting distribution is smooth at the $5.34 million net worth threshold.

threshold is difficult to determine, we used data from the SCF to impute asset and liability values when the imputed net worth is below the threshold.

The SCF-based dataset was prepared by aggregating wealth into seven asset categories (real estate, bonds, stocks, business, retirement, other financial and other non-financial) and two liability categories (mortgage debt and other debt). These categories were chosen so that variables in both the SCF-based and estate tax-based datasets could be aggregated into comparable categories. Preparation of the SCF-based dataset also involved calculating basis in percentage terms for real estate assets, stocks, and business assets by dividing the reported cost price of aggregated assets by the reported market value of aggregated assets.

Preparation of the estate tax dataset similarly required aggregation into the above asset and liability categories. It also required combining the population data (ESTATE2014_POPULATION) that lacked detail on assets and liabilities with the sample data (ESTATE2014_SOI) that contain those details. To impute this detail to the population, a simple hot deck from the SOI sample to the population data was performed, with cells consisting of income, age, and marital status.

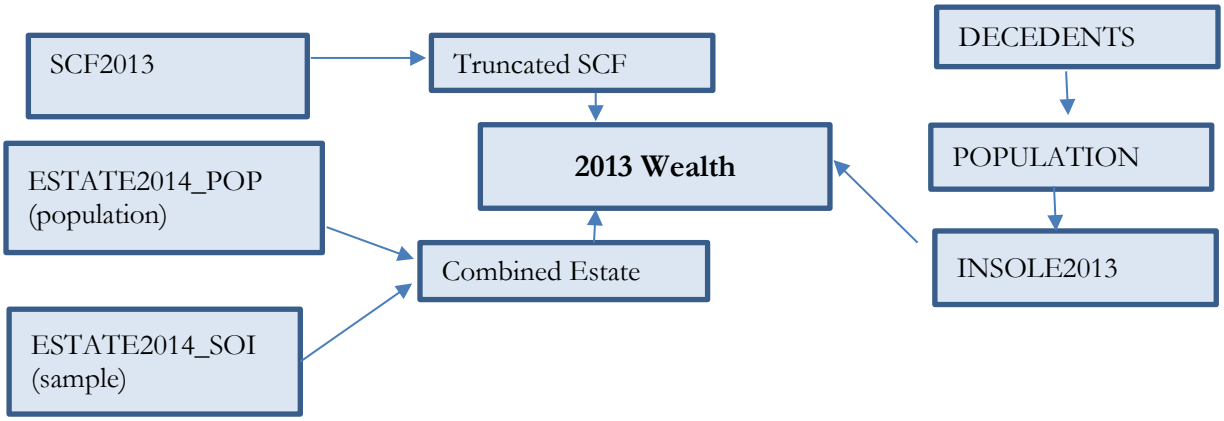
Before merging the INSOLE2013 with the SCF-based and estate tax-based datasets, it was necessary first to impute a binary variable indicating if a taxpayer in the INSOLE file had a net worth above or below $5.34 million. This was done by merging the 2013 INSOLE to the population of decedents from 2014 and the population of estate tax returns from 2014. A logit model was estimated on INSOLE decedents with capital income (interest, dividends, etc.), retirement balances (IRA, SEP, etc.), and marital status as covariates. This model was then used to predict if the net worth of non-decedents was above or below $5.34 million.

Conditional on this predicted net worth indicator, asset and debt components were imputed via hot decking to INSOLE2013 from either the SCF-based dataset if predicted net worth was below $5.34 million or the estate-based dataset if predicted net worth was above $5.34 million. Basis for each type of asset, measured as a percentage of asset values, were imputed via hot decking to the entire INSOLE13 using the SCF-based dataset since the estate-based dataset does not

contain information on basis.[22]

Figure A1 shows a general diagram of the matching that leads to the estimation of 2013 wealth in 2013.

Figure A1: Files Used to Merge INSOLE2013 Data to Non-INSOLE Data



Calibration and Extrapolation

In order to create the 2016 WM, we hot decked the asset and liability data created from the INSOLE2013 to INSOLE2016, using total income, age, marital status, interest income, dividend income, and property taxes to create the cells for the hot decking. We limited our imputation to these 6 control variables since increasing the number of control variables causes the number of

[22] For the purposes of imputing asset and debt components, the 2013 SCF file was truncated to exclude individuals with reported net worth exceeding $5.34m since the combined estate tax file was used to impute these components for taxpayers with predicted net worth above this level. For the purposes of imputing basis for each type of asset, the entire 2013 SCF file was used to impure these components for all taxpayer records in the INSOLE2013.

cells to grow exponentially, which in turn leads to empty donor cells. Hence, while our methodology assigns tax-preferred retirement wealth to only those with tax-preferred retirement savings, housing wealth to only homeowners, stocks holdings to those with dividend income and bond holdings to those with interest income, it does not assign business assets necessarily to those with business income.

After asset and liability fields were created, they were calibrated to produce an estate tax estimate comparable to the reported tax revenues for tax years 2015 and 2016 both in aggregate and by gross estate size and marital status class, which consists of ten cells (5 gross estate size groups and 2 marital status groups). The latter necessitated the use of both 2015 and 2016 due to the small size of the estate tax sample. (The basis of assets is determined by taking the calibrated asset values and then applying imputed basis percentages from the SCF. The basis is also adjusted to be consistent with existing revenue estimates from taxing capital gains at death.

At each budget cycle, all monetary amounts in the WM are grown at the GDP growth rate on the ITM plus an additional 0.1 percent a year, which is in line with the growth in historical receipts for the estate and gift tax. Non-monetary fields such as expected mortality (conditional on age, sex, and income), as well as basis (as a percentage of asset value), are held constant across all years.

APPENDIX V: PASS-THROUGH INCOME

The K-1 information return has the partner's/shareholder's tax id (SSN) as well as the entity's tax id (EIN). So, both SSN and EIN already exist on the K-1. Edited K-1s are available for INSOLE taxpayers and their dependents. We add EIN-level data, using the EIN from the K-1, by matching the EIN to CDW files; in this way, we are able to attach entity-level information (such as total assets and interest deduction) to a particular K-1.

We keep information for as many as 10 partnership K-1s plus as many as 10 S-Corp K-1s for each INSOLE return. The INSOLE only has aggregated Schedule E data. So, we are matching as many as 20 K-1s in order to identify the source of the Schedule E income. Among INSOLE returns with non-zero schedule E income or loss, there are a number of returns with more than 10 K-1s. For these returns, we rank the K-1s by the absolute value of income or loss (active income + passive income + abs(active loss) + abs(passive loss)) and add the EIN level data for the top 10.

The entity-level information is also used to define a "business" flag which may be used at the discretion of the analyst.[23]

In total, 27 fields are extracted from the population files for Forms 1065 and 1120S. These include the NAICS industry code, total assets, gross receipts, net receipts, various deductions and types of compensation. In addition, 12 more entity-level fields are created by aggregating K-1 information by EIN (so, all K-1s found in the population file for given tax-year & EIN, rolled-up to the EIN-level). These aggregated fields include ordinary income, dividends, interest, net long-term gains or losses, net short-term gains or losses, real estate, other rental, and a count of the number of K-1s aggregated.

[23] One example of a non-business is where individuals or entities form partnerships to re-distribute earnings that are passed through from other partnerships. These entities are conduits that merely redistribute funds. For further information, see Knittel et. al. (2016).

REFERENCES

Cronin, Julie-Anne. 2022. *U.S. Treasury Distributional Analysis Methodology*. OTA Technical Paper 8. Washington, DC: U.S. Department of the Treasury. https://home.treasury.gov/system/files/131/TP-8.pdf.

Fisher, Robin and Emily Lin. 2015. *Re-weighting to Produce State-Level Tax Microsimulation Estimates.* OTA Technical Paper 6. Washington, DC: U.S. Department of the Treasury.

Gillette, Robert. 1989. *Measures of Goodness of Fit for Extrapolations: Initial Results Using the Individual Tax Model Database.* OTA Paper 62. Washington, DC: U.S. Department of the Treasury. https://home.treasury.gov/system/files/131/TP-6.pdf.

Heim, Bradley T., Ithai Z. Lurie, and James Pearce. 2017. "What Drove the Decline in Taxpaying? The Roles of Policy and Population." *National Tax Journal* 70 (3): 585–620. https://doi.org/10.17310/ntj.2017.3.03.

Knittel, Matthew, Susan Nelson, Jason DeBacker, John Kitchen, James Pearce. Prisinzano, Richard. *Methodology to Identify Small Businesses.* OTA Technical Paper 4. Washington, DC: U.S. Department of the Treasury. https://home.treasury.gov/system/files/131/TP-4.pdf.

Little, Roderick and Donald Rubin. 1987. Statistical Analysis with Missing Data. John Wiley and Sons Inc., New York.

Lurie, Ithai Z., and James Pearce. 2021. "Health Insurance Coverage in Tax and Survey Data." *American Journal of Health Economics* 7 (2): 164–84. https://doi.org/10.1086/712213.

Lurie, Ithai Z. and Corbin L. Miller. 2022. "Employer-Sponsored Health Insurance Premiums and Income in US Tax Data." *Journal of Public Economics* (forthcoming). https://www.corbinmiller.website/publication/premiums/premiums.pdf.