Appendix A

IDDA Technical Documentation

Date: September 20, 2023

1. Data Sources

The Income Distributions and Dynamics in America (IDDA) statistics are produced from the universe of individual tax returns filed with the Internal Revenue Service (IRS) from 1998 to 2019, linked to detailed demographic information available from various sources through the U.S. Census Bureau. Statistics cover two primary samples: household income and earnings data aggregated from Form 1040 and individual-level earnings information reported on Form W-2 (available from 2005-2019). Records in both samples are linked to the Census Numident file using the "protected identification key" (PIK), a privatized person identifier as unique as a social security number. The Census Numident contains demographic information recorded by the Social Security Administration (SSA) and facilitates additional linkages to the decennial Census, American Community Survey, and other administrative summary files. The PIK also enables longitudinal linkage of the IRS data and the construction of IDDA statistics that track income mobility and changes for individuals and households over time.

The IDDA statistics contain summary income measures broken out by sex (men and women); race and ethnicity (Hispanic, non-Hispanic American Indian or Alaska Native, non-Hispanic Asian, non-Hispanic Black, non-Hispanic Native Hawaiian or Pacific Islander, non-Hispanic other or multiple races, and non-Hispanic White); age bracket; and place of birth (U.S.- or Foreign-born). They also include intersections of race and age, race and sex, and age and sex. See the IDDA codebook for complete definitions and source information for all demographic variables and values in the published dataset.

Sex, place of birth, and years of birth and death (which determine age) come from the SSA Numident file which records all transactions against all Social Security Numbers. The Numident file contains the sex reported on an individual's initial Social Security Number application unless

the individual applied and was able to change their sex recorded by the SSA. Place of birth indicates whether individuals were born inside or outside of the United States. Individuals born abroad to U.S. parents are considered U.S.-born.

Race and ethnicity is a summary variable coded from Census Bureau's 2020 Best Race File, the 2000 Decennial Census, the 2010 Decennial Census, and 2005-2019 American Community Survey (ACS). The Census Bureau's best race file pulls from a variety of sources including records from the Temporary Assistance for Needy Families program, the Department of Housing and Urban Development, and the Decennial Census and is our preferred method for identifying race and ethnicity. If a record does not appear in the Census Bureau Best Race File, the most recent race and ethnicity data from either the Decennial Census or ACS are used to code the summary race and ethnicity variable.

In the 1040 sample, household income is aggregated across all 1040 forms associated with a shared address, identified using the Census Bureau Master Address File ID (MAFID). The primary and secondary filers on each form 1040 are assigned the total income value pertaining to the common address. This allows for an expanded notion of household income that includes, for example: earners who are married filing separately, multiple individual filers living in the same household (such as housemates or partners), or multiple generations of a family living in a single household and filing multiple 1040s. It also allows us to provide statistics on the distribution of total household income across demographic groups including race, ethnicity, and age that vary among household members.

2. Sample Selection

In a given year, records from the universe of PIKs can be excluded from either or both the W-2 and 1040 samples. The most straightforward reason a record is excluded from either sample is that they did not receive a W-2 or did not file a 1040 form in the tax year. Filing requirements vary by the age of filer(s), filing status (single or joint), and over time. As an example, single individuals under 65 without dependents were required to file Form 1040 for the 2022 tax year if their taxable gross income was at least $12,950. Users can compare the probability that individuals or households move out of the W-2 or 1040 data over a 1- or 5-year period in the IDDA income transition matrix module.

Note, many low-income individuals filed a federal income tax return to receive economic stimulus payments under the Economic Stimulus Act of 2008 (affecting 1040 data in tax year 2007) and to receive stimulus payments during the Covid-19 pandemic (affecting 1040 data in tax year 2019).[1] Some measures of total household income tend to be lower in these years.

Records are excluded from both samples if they are missing key demographic or geographic information: sex or year of birth (from the Numident), state of residence (from 1040 or 1099 tax forms), or race/ethnicity (from any of the ranked sources described above). Individuals are also excluded if they are younger than 16, older than 100, or deceased, or if their state of residence is not listed as one of the 50 states or D.C. Due to sample size limitations, PIKs are excluded if their sex is not recorded by the SSA as male or female.

Additionally, individual records are excluded from the 1040 sample only if they:

- Do not appear in the Census Master Address File
- Have missing values for any of the 1040 income variables or if individuals in the same tax unit have different values for any of the 1040 income variables
- Have an unusually high count of individuals associated with the underlying MAFID
- Are not the primary or secondary filer on the 1040 form on which they are listed

Individual records are excluded from the W-2 sample only if they:

- Have missing values for any of the W-2 income variables
- Have a value of zero for either wage income or total compensation reported on form W-2

Once sample restrictions are made in each cross-sectional data file, records are linked longitudinally via the PIK over 1- and 5-year time horizons. In general, records appear in the panel data files if they are in the W-2 or 1040 data in either of the two years. If a PIK does not appear in the W-2 or MAFID-1040 sample in one year, then the relevant income values are set to missing and income growth measures are not calculated for that record.

The IDDA data also includes statistics for a secondary prime-age workers sample restricted to individuals aged 25-54 with earnings above a threshold, equivalent to working half-time for 13

[1] IRS Publication 1304, 2007 and 2019: https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-returns-complete-report-publication-1304#_IndReturns

weeks at the federal minimum wage, as measured by their wage compensation on form W-2. In the longitudinal data, individuals are allowed to join or leave the prime-age workers sample as they age or move above or below the minimum wage threshold. As before, individuals are considered part of the prime-age earners if they are aged 25-54 and earned more than the defined earnings threshold in either of the two years. Demographic information is taken from the base year, meaning that individuals aged 16-24 can "age in" a given income bracket in the prime-age earners sample and individuals aged 45-54 can "age out."

3. Income Variables

3a. Income variable sources

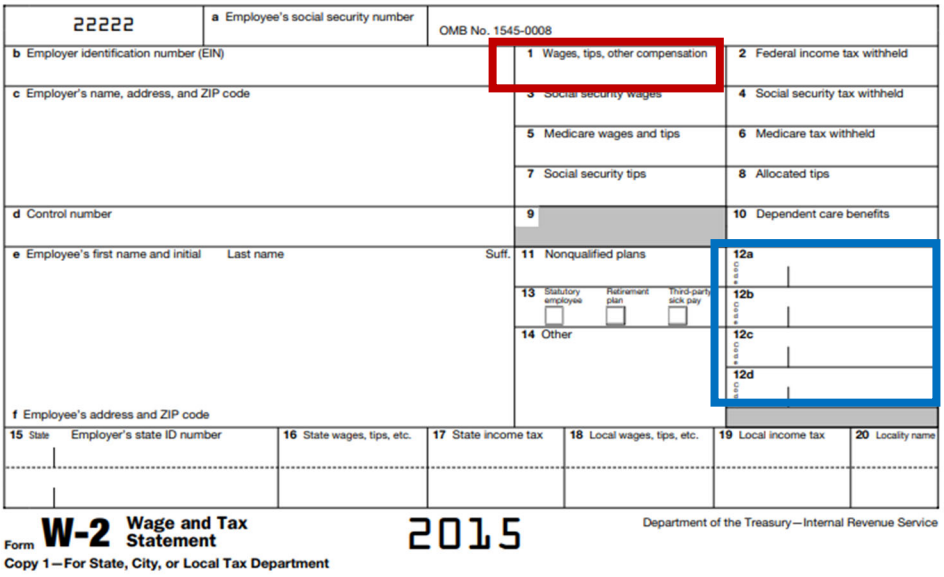
The W-2 sample includes three income variables: Wage compensation (WC), deferred compensation (DC), and total compensation (TC), shown on the 2015 Form W-2 in Figure 1. Wage compensation is the total income reported in Box 1 of all W-2s received by the individual in the tax year. Deferred compensation is the total income reported in Box 12a, 12b, 12c, and 12d of all W-2s received by the individual in the tax year. Total compensation is the sum of wage and deferred compensation and is the primary W-2 earnings measure available at the state level.

Figure 1: Individual Income Variables Derived from Form W-2

Wage compensation: For an individual earner, total of Box 1 compensation across all W-2s in the tax year

Deferred compensation: Total of Box 12a-d compensation across all W-2s. Includes payroll deductions and elective deferrals

Total compensation =



Note: 2015 version of Form W-2 shown for reference. Box numbers may change over time.

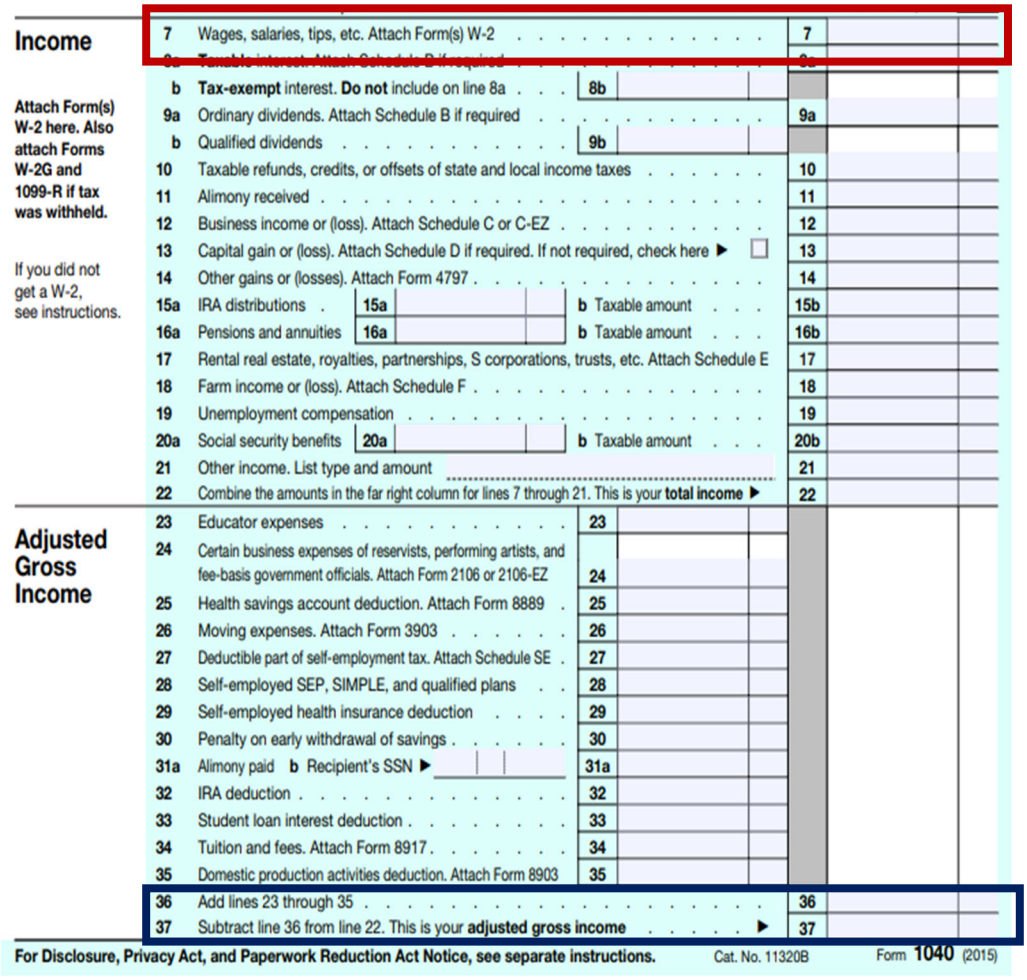
The 1040 sample includes household-level adjusted gross income (GI), wage and salary income (WS), and nonwage income (NW) aggregated across all 1040s filed at a shared address. Wage and salary income is the total wage compensation reported in Box 1 of all W-2's received by all filers in the tax unit, reported on line 7 of the 2015 Form 1040 (Figure 2).

Figure 2: Household Income Variables Derived from Form 1040

Wage and salary income (WS): Sum of Box 1 wage compensation across all W-2s in the tax unit (single or joint filers) (line 7). Aggregated across all tax units with a shared address.

Adjusted gross income (GI): WS + interest, dividends, capital gains/losses, self-employment income, retirement income, etc (lines 8-21). Minus adjustments (lines 23-36).

Nonwage Income (NW) = GI – WS



Income	7	Wages, salaries, tips, etc. Attach Form(s) W-2		7
	b	**Tax-exempt** interest. **Do not** include on line 8a	8b	
Attach Form(s) W-2 here. Also attach Forms W-2G and 1099-R if tax was withheld.	9a	Ordinary dividends. Attach Schedule B if required		9a
	b	Qualified dividends	9b	
	10	Taxable refunds, credits, or offsets of state and local income taxes		10
	11	Alimony received		11
	12	Business income or (loss). Attach Schedule C or C-EZ		12
	13	Capital gain or (loss). Attach Schedule D if required. If not required, check here ▶ ☐		13
If you did not get a W-2, see instructions.	14	Other gains or (losses). Attach Form 4797		14
	15a	IRA distributions 15a	b Taxable amount	15b
	16a	Pensions and annuities 16a	b Taxable amount	16b
	17	Rental real estate, royalties, partnerships, S corporations, trusts, etc. Attach Schedule E		17
	18	Farm income or (loss). Attach Schedule F		18
	19	Unemployment compensation		19
	20a	Social security benefits 20a	b Taxable amount	20b
	21	Other income. List type and amount		21
	22	Combine the amounts in the far right column for lines 7 through 21. This is your **total income** ▶		22
Adjusted Gross Income	23	Educator expenses	23	
	24	Certain business expenses of reservists, performing artists, and fee-basis government officials. Attach Form 2106 or 2106-EZ	24	
	25	Health savings account deduction. Attach Form 8889	25	
	26	Moving expenses. Attach Form 3903	26	
	27	Deductible part of self-employment tax. Attach Schedule SE	27	
	28	Self-employed SEP, SIMPLE, and qualified plans	28	
	29	Self-employed health insurance deduction	29	
	30	Penalty on early withdrawal of savings	30	
	31a	Alimony paid b Recipient's SSN ▶	31a	
	32	IRA deduction	32	
	33	Student loan interest deduction	33	
	34	Tuition and fees. Attach Form 8917	34	
	35	Domestic production activities deduction. Attach Form 8903	35	
	36	Add lines 23 through 35		36
	37	Subtract line 36 from line 22. This is your **adjusted gross income** ▶		37

For Disclosure, Privacy Act, and Paperwork Reduction Act Notice, see separate instructions. Cat. No. 11320B Form **1040** (2015)

Note: 2015 version of Form 1040 shown for reference. Box numbers may change over time.

Adjusted gross income includes wage and salary income plus self-employment earnings, interest and dividends, capital gains/losses, unemployment insurance, and the taxable components of social security income, supplemental security income, and other retirement income (e.g. from a pension

or IRA), minus deductions. The components of gross income in addition to wages, salaries, and tips are reported on IRS Schedules B, C, D, E and F, and Form 4797. Adjusted gross income does not include nontaxable transfer income—including benefits from the Supplemental Nutrition Assistance Program, Temporary Assistance for Needy Families, and most welfare payments generally—or nontaxable tax credits or refunds such as the Child Tax Credit or Earned Income Tax Credit. Deductions are listed in lines 23-35 of the 2015 Form 1040 (Figure 2) and include, for instance, the deductible part of self-employment tax, deductions from health savings accounts, and student loan interest deductions. After 2017, both additional income and adjustments to income are totaled on Schedule 1 before they are passed through to Form 1040. Gross income is the last line a filer reports on Form 1040 before claiming standard or itemized deductions.

Nonwage income is the difference between gross income and wage and salary income. Wage and salary income can be positive or zero, while nonwage income and gross income can be positive, negative, or zero.

3b. Aggregation of household income

Income variables as reported on Form 1040 represent total income earned by members of a tax unit—either one single or two joint filers, plus any dependents claimed. If multiple 1040 forms are filed by earners at a common address, the values of GI, WS, and NW are aggregated across all 1040 forms at the address. The summed, address-level income values are then assigned to each primary and secondary filer residing at the address, subject to the sample restrictions above. Thus, the household-level income statistics in IDDA should be interpreted as measures of total household resources available for individuals in a particular demographic group.

As a result, income dynamics at the household level capture both income changes for individuals who continue to reside at the same address as well as changes in household composition. For example, if an individual lives alone in 2005 but shares an address with two other people in 2010, that person's income growth over the 5-year horizon includes the income earned by those two additional tax filers, even if they are not represented on the same 1040. On the other hand, a couple that lives at the same address and files jointly would have the same household income in the aggregated dataset as if they filed separately.

At the U.S. level, the IDDA data also include a version of gross income (GI_ADJ) that divides household adjusted gross income by the square root of the household size, where household size is the total number of primary and secondary filers plus the number of dependents claimed on all 1040 forms associated with the address.

4. Statistics Modules

The statistics in IDDA can be grouped into five modules: Percentiles of Income, Top Income Shares, and Top Income Population Shares (which measure income distributions in a given year); Income Change Distributions; and the Income Transition Matrix module (which measure movement and changes along the income distribution over time). Both the income change distributions and transition matrices leverage the longitudinal data linking individual tax filers and their associated households over 1- and 5-year time horizons.

All statistics are calculated within a geography (U.S.-level or an individual state), year (or base year and time horizon), income variable, and sample. They are then defined either within or across demographic groups. More formally, a statistic S on an income concept in IDDA exists at the *ktlg* level, where

- *k* is the household-1040, individual- W-2, or individual prime-age earners- W-2 sample, and a corresponding income concept (GI, NW, WS, WC, TC, or DC),
- *t* is the time horizon (1998 to 2019 + 1- and 5-year horizons),
- *l* is the geography (U.S. or an individual state), and
- *g* is a demographic group (overall sample, age, sex, race/ethnicity, birth location, or an intersection). In the downloadable IDDA datasets, the column *group_var* specifies the level of disaggregation (e.g., “by race”, or “age-by-sex”), and *group_var_val* identifies the particular demographic group within it (e.g., “non-Hispanic Asian” or “55-64 Female”).

The statistics included in each module are defined in sections 4a-4e. Appendix B summarizes the demographic disaggregation dimensions available by module, sample, and geographic level. For complete record layout information, please refer to the IDDA codebook.

To preserve confidentiality, all percentile values report the mean income (or change in income) for a small set of observations around the *p*th percentile rather than an individual person’s income (or

change in income). Sample size restrictions and confidentiality considerations may limit the availability of these defined statistics. See subsection 4.f titled *"Suppression in the IDDA data"* for a detailed discussion.

4a. Percentiles of Income Module

The percentiles of income module contains group-percentile income values at the 10^{th}, 25^{th}, 50^{th}, 75^{th}, 90^{th}, 95^{th}, and 98^{th} percentile (state and U.S.-level) and 99^{th}, 99.9^{th}, 99.99^{th}, and 99.999^{th} percentile (U.S.-level only). The statistics are calculated within demographic groups, meaning that the sorting, ranking, and assignment of percentile values is performed only for the subset of records belonging to the particular group of interest. For example, in 2019, the 90^{th} percentile of total W-2 compensation among Hispanic men in Illinois was $88,360.

Percentiles of household adjusted gross income and household nonwage income can be zero or negative. Percentiles of household wage/salary income and individual deferred compensation can be zero. Percentiles of individual wage and total compensation are positive.

4b. Top Income Shares Module

Top income shares are calculated both within and across groups. The "within" shares calculate the proportion of total income held by members of a demographic group (denominator) that is held by the top *p* percent of earners in that group's income distribution (numerator). Both the numerator and denominator are thus defined at the *ktlg* level. For example, in 2019, the top 10 percent of male earners in Massachusetts held 43 percent of the total W-2 compensation earned by men in Massachusetts. The dollar value cutpoint that determines whether a male earner in Massachusetts falls into the top 10 percent matches the value of *pctl90* when the geography is Massachusetts and demographic group is "Male" in the percentiles of income file in 2019.

The "across" shares are created using data for all sample members, not just the individuals belonging to the group of interest. The denominator is the sum of all positive income held by the top *p* percent of earners, regardless of demographic group. The numerator is the sum of income held by individuals in the top *p* percent who also belong to a particular demographic group. Therefore, the numerator is defined at the *ktlg* level but the denominator is defined at the *ktl* level. For example, nationally in 2019, men earned 73 percent of the total W-2 compensation held by the

top 10 percent of earners. In this case, the dollar value cutpoint that determines whether an earner falls into the top 10 percent matches the value of *pctl90* when the demographic group is "All sample members" in the percentiles of income file.

Top income shares are calculated for earners at or above the 90^{th}, 95^{th}, and 98^{th} percentiles (state- and U.S.-level), and 99^{th}, 99.9^{th}, 99.99^{th}, and 99.999^{th} percentiles (U.S-level.). The "across" shares are also provided for the full population (0^{th} percentile).

4c. Top Income Population Shares Module

The population shares module provides the demographic composition of a subset of the income distribution. Shares are calculated across demographic groups from the same underlying populations as the "across" income shares. The denominator is the total count of individuals at or above the *p*th percentile of income, regardless of demographic group. The numerator is the count of individuals at or above the *p*th percentile of income who also belong to a particular demographic group. The numerator is defined at the *ktlg* level but the denominator is defined at the *ktl* level. For example, nationally in 2019, men comprised 69 percent of the top 10 percent of earners based on total W-2 compensation. Based on the corresponding "across" income share, that means men held a slightly larger share of the top 10 percent of total W-2 compensation than their share of the population in that top 10 percent group. The dollar value cutpoint that determines whether an individual falls at or above the *p*th percentile matches the corresponding value in the percentiles of income file when the demographic group is "All sample members".

4d. Income Change Distributions Module

This module provides the distribution of 1-year income changes and 5-year income changes by base year income bin (typically an income quartile) and demographic group.. The base year income bins are defined within a sample, year, and geography (*ktl*) but across all sample members (across all *g*). In the income change distributions and transition matrix modules, the income bins are quartiles for state-level statistics. The top income quartile (above the 75^{th} percentile) is split into two smaller subsets at the national level: the 75-90^{th} percentile of initial income, and above the 90^{th} percentile.

For individuals who are in the tax sample in both the base and subsequent year (y0 and y1), change in income is calculated as the nominal difference in individual or household level income from y0 to y1, divided by the time horizon (y1 minus y0). Individuals are ranked within an initial income bin and demographic group by this dollar value, and the mean and the 10th, 25th, 50th, 75th, and 90th percentiles of income changes are reported. Thus, this module shows what "strong" and "weak" income growth looks like at different points of the income distribution, and whether demographic groups experience different year-to-year patterns of income change even from similar initial income levels.

For example, among all individuals who started in the bottom earnings quartile (below the 25th percentile of total W-2 compensation in the national distribution) in 2018, the 10th percentile of 1-year income changes was a loss of $3,269. The median income change was $2000 and the 90th percentile was $16,410. Among Asian earners who started in the bottom earnings quartile, the 10th percentile of income changes from 2018 to 2019 was a loss of $2,931 and the median was a gain of $2,220, similar to the overall sample. However, the 90th percentile was $21,100, higher than in the overall sample.

Reporting the annualized difference rather than a percent change in income is useful as some of the income concepts (adjusted gross income and nonwage income) can take zero or negative values.

4e. Income Transition Matrix Module

The transition matrix statistics give the probability that an individual starting in a given income bin moves to another income bin after 1 or 5 years. As with the income change distributions module, the initial year and final year income bins are defined across all members in a sample, year, and geography (*ktl*), not just individuals belonging to the group of interest. The bins are quartiles of the state-level income distribution for state-level statistics. For national statistics, the top quartile is split into the 75-90th percentile of initial income and above the 90th percentile. The transition probabilities are calculated for a particular demographic group and in an initial income bin, meaning that they add up to 100% within each unique combination of these variables.

The transition matrices are computed using records that are in the tax sample in either of the two years. If an individual is not in the tax sample in one year, their income bin in that year is labeled

"missing." The transition probabilities when the initial year income bin is "missing" give the likelihood that members of a demographic group of interest without data in the initial year enter an income bin over the 1-year or 5-year period. Similarly, transitions from the W-2 or 1040 data into "missing" give a sense of how common movement into nonemployment or non-filing is for individuals at different points in the income distribution, though there are other reasons a record might be excluded from the sample in a year, as detailed above.

To provide an example, among Black workers in Minnesota who started in the lowest quartile of total W-2 compensation in 2014 (based on the statewide distribution), 36 percent remained in the lowest earnings quartile in 2019, 32 percent had earnings in the second quartile, and 9 percent had earnings in the third quartile of the statewide distribution. About 1.5 percent had earnings in the top quartile, and 21 percent were not in the W-2 sample in 2019. The transition matrices also track the probability that individuals move into or out of the W-2 prime-age workers subsample. Records that are not in the prime-age worker subsample in one year are assigned the earnings quartile "out-of-sample" in that year.

4f. Suppression in the IDDA data

The IDDA statistics are only published based on underlying samples and implicit samples that meet a certain minimum size. Where groups do not meet this minimum threshold, statistics are suppressed in order to protect the confidentiality of individuals in the tax data. This occurs most often in small states where a particular demographic group is not highly represented, in the upper tail of the income distribution, and for intersections of small race or ethnicity groups with age or sex. Suppressed values are excluded from the "long" format csv files available for download so will appear as missing values when the data are reshaped to "wide" format.

The population shares, top income shares, and transition matrix files contain sets of probabilities that add up to 100%. In these cases, if one statistic is suppressed, additional statistics are also suppressed so that no information can be "backed out" about a group smaller than the minimum size. The method for performing suppressions is described in Table 2. Note that the suppression rules try to preserve smaller race/ethnicity groups whenever possible instead of systematically favoring the larger groups.

Table 2: Suppression in the IDDA Data

Module	Rule
Percentiles of income	Sample sizes at and between each percentile value meet minimum threshold
Population shares	Shares for "across" statistics add up to 100% within a given subset of the income distribution. If one share is suppressed, then the rules for suppressing additional statistics depends on the group variable: *By age:* suppress the next smallest age category *By race/ethnicity:* suppress the non-Hispanic other or multiple races group. Then, suppress the largest race/ethnicity group. *By sex:* suppress the other sex included in the data (male/female) *By foreign-born status:* suppress the other group (foreign/U.S.-born) These prioritizations are preserved when suppressing statistics at the intersections of age, race, and sex.
Top income shares	Same as in the population shares module
Income change distributions	Sample sizes at and between each percentile value meet minimum threshold
Transition matrix	Probabilities add up to 100% within an initial earnings bin. If one probability is suppressed, then the transition probability representing the next smallest group is also suppressed until the total size of the excluded group meets the minimum threshold.

5. Coverage and Comparability

5a. Coverage of Statistics in IDDA

The levels of disaggregation available in IDDA vary by module, geography, and sample. Table 3 presents the total number of defined statistics in the household-1040, individual-W-2, and prime-age workers-W2 samples, along with availability rates at the state and U.S. level. Availability is simply the percent of all defined statistics that are available (not suppressed) within the module(s), geography, and sample. The "income levels" availability rate is the percent of total defined statistics available in the percentiles of income, top income shares, and top income population shares modules.

Table 3: Coverage in IDDA by geography and sample

	U.S.-level				State-level			
	1998-2004		2005-2019		1998-2004		2005-2019	
Household-1040	Defined	Available	Defined	Available	Defined	Available	Defined	Available
Income Levels	49,056	86%	105,120	88%	187,068	97%	400,860	98%
Transition Matrix	22,050	100%	37,800	100%	514,080	97%	881,280	97%
Income Changes	18,900	100%	32,400	100%	514,080	97%	881,280	98%
Individual-W2								
Income Levels	-		121,680	87%	-		1,054,170	85%
Transition Matrix	-		129,360	97%	-		499,392	95%
Income Changes	-		110,880	>99%	-		499,392	99%
Prime-age workers-W2								
Income Levels	-		80,370	89%	-		-	
Transition Matrix	-		156,672	76%	-		-	
Income Changes	-		72,000	100%	-		-	

For the set of defined statistics, both the income change distributions and transition matrix modules feature near-complete coverage at the U.S. level, and coverage above 95 percent at the state level. In general, availability is low at the intersections of age and race, which has 36 categories (6 race/ethnicity groups and 6 age categories). The intersection of age and race is reported in the U.S. and state W-2 income levels, U.S. W-2 transition matrices and income change distributions, and U.S. 1040 income levels. The effect of including age-by-race statistics is especially large in the income levels, which provide statistics far into the tail of the income distribution. For instance, coverage in the Household-1040 sample at the U.S. level, where age-by-race is included, is around 87 percent, while the state-level availability rate for the same sample is around 98 percent. Transition matrix coverage is low in the prime-age workers sample because individuals can "age in" or "age out" of the sample over time, leading to a number of statistics that are defined for small populations.

Table 4 provides a more detailed analysis of availability for individual race/ethnicity subgroups by sample and geography. Availability is given as a percent of all defined statistics pertaining to a specific subgroup, including intersections of age and race and race and sex. For example, for a fixed income concept and *ktl*, a statistic defined for the groups Asian Male, Asian Female, and Asian 25-34 each counts as 1 in the denominator. As in Table 3, coverage is lower at the intersections of age and race and in the tail of the income distribution.

Table 4: Detailed coverage by race/ethnicity

	Quartiles +	Intersections	Hispanic	AIAN*	Asian	Black	NHOPI*	White
U.S. Household-1040								
Income Levels	10, 25, 50, 75, 90, 95, 98 + tail	age	92%	76%	89%	88%	69%	95%
Transition Matrix	lt25, 25-50, 50-75, 75-90, gt90	-	100%	100%	100%	100%	100%	100%
Income Changes	lt25, 25-50, 50-75, 75-90, gt90	-	100%	100%	100%	100%	100%	100%
U.S. Individual-W2								
Income Levels	10, 25, 50, 75, 90, 95, 98 + tail	age, sex	90%	74%	88%	91%	66%	92%
Transition Matrix	lt25, 25-50, 50-75, 75-90, gt90	age, sex	100%	95%	100%	100%	83%	100%
Income Changes	lt25, 25-50, 50-75, 75-90, gt90	age, sex	100%	100%	100%	100%	98%	100%
U.S. Prime-age workers-W2								
Income Levels	10, 25, 50, 75, 90, 95, 98 + tail	age, sex	94%	77%	91%	93%	71%	94%
Transition Matrix	out-of-sample, lt25, 25-50, 50-75, 75-90, gt90	age, sex	79%	76%	79%	78%	69%	79%
Income Changes	lt25, 25-50, 50-75, 75-90, gt90	age, sex	100%	100%	100%	100%	100%	100%
State -1040								
Income Levels	10, 25, 50, 75, 90, 95, 98	-	100%	97%	100%	98%	72%	100%
Transition Matrix	lt25, 25-50, 50-75, gt75	-	100%	95%	99%	98%	62%	100%
Income Changes	lt25, 25-50, 50-75, gt75	-	100%	97%	100%	99%	74%	100%
State-W2								
Income Levels	10, 25, 50, 75, 90, 95, 98	age, sex	91%	76%	89%	88%	44%	95%
Transition Matrix	lt25, 25-50, 50-75, gt75	-	97%	84%	94%	94%	47%	100%
Income Changes	lt25, 25-50, 50-75, gt75	-	100%	98%	100%	100%	78%	100%

*Notes: The column "Quartiles +" shows which base year percentile values or income quantiles a set of statistics is calculated for, to contextualize the corresponding availability. For instance, income levels are provided further into the tail of the income distribution than the transition matrix and income change measures, resulting in additional suppression. The column "intersections" indicates whether statistics in the given geography/sample are provided for the intersection of age and race, race and sex, or neither. Rates are averages across all years. *NHOPI abbreviates Native Hawaiian or Pacific Islander; AIAN abbreviates American Indian or Alaska Native.*

5b. Comparability

A supplementary dataset compares the income distribution in IDDA to the Current Population Survey Annual Social and Economic Supplement (CPS ASEC), a standard source for income measurement in the U.S. The IDDA CPS supplement dataset and accompanying documentation will be available for download from the project website. It contains percentiles of income in the CPS disaggregated by race/ethnicity, sex, and place of birth (foreign-born or U.S.-born) for a subset of tax years, as well as analysis of a linked CPS-IRS sample. This section shares high-level findings from the IDDA CPS supplement that contextualize two key features of the IDDA statistics: the household aggregation method and the gross income measure.

Households and addresses in IDDA: Table 5 shows how the address identifier ("MAFID") used to construct household income in IDDA maps onto households in the Current Population Survey, using the tax year 2012 as an example. The first column compares the rate of overlap between MAFIDs and households in the CPS ASEC: A CPS household is considered a "match" to its IRS counterpart if all the household members in the CPS have the same value of MAFID, and all the members of the MAFID that show up in the CPS data are part of the same household. This rate is high, at least 90 percent for all demographic groups. The second and third columns show that household size is slightly larger in the IDDA than the CPS on average, possibly because some tax filers claim dependents who do not physically reside with them (for example, students or elderly family members). However, the relative differences in mean household size across demographic groups is preserved between both sources. Finally, column 4 suggests that the CPS and IRS household concepts are less tightly aligned for individuals who are nonwhite or foreign-born: even though the difference in mean household size between the CPS and IRS data is not particularly large for these groups, they are less likely to reside in a CPS household that has the exact number of members as are associated with their value of MAFID.[2]

[2] Prior research has also shown that the probabilistic matching technique used to assign the PIK identifier yields more false matches for some groups, especially non-U.S. citizens and recent movers, which may affect linkage to the CPS. For a discussion of bias in PIK assignment, see https://www.census.gov/content/dam/Census/library/working-papers/2014/adrm/carra-wp-2014-08.pdf.

Table 5: Households in the Current Population Survey and IRS Data, 2012

Demographic Group	MAFID overlap rate with CPS households	Mean CPS household size	Mean IDDA household size	Exact MAFID household size match
All Sample Members	94%	2.87	3.12	63%
Hispanic	90%	3.5	3.9	47%
American Indian or Alaska Native	91%	3.1	3.4	52%
Asian	94%	3.3	3.6	58%
Black	90%	3.0	3.3	49%
Native Hawaiian or Pacific Islander	91%	3.7	4.0	52%
White	95%	2.7	3.0	68%
Foreign-born	93%	3.3	3.7	54%
U.S.-born	94%	2.8	3.1	64%

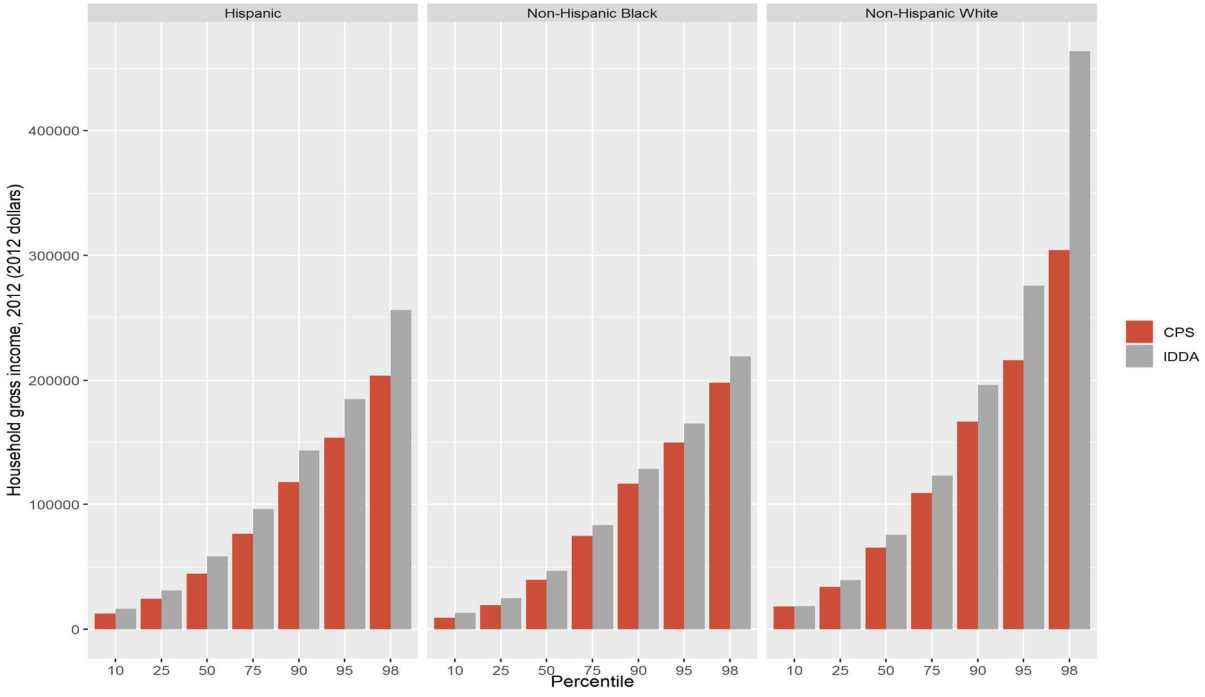
Source: Federal Reserve Bank of Minneapolis, Income Distributions and Dynamics in America.

Gross Income: Figure 3 compares the distribution of gross income (GI) in IDDA to the distribution of "total household income" (HTOTVAL) reported in the CPS ASEC in 2012 for three large race/ethnicity groups—Hispanic, non-Hispanic Black, and non-Hispanic White individuals. By definition, total household income in the ASEC contains most of the same income types as are included in gross income: wage and salary earnings, self-employment earnings, interest and dividends, social security and retirement income, and other types of rental or farm income. For individuals in lower-income households, CPS total household income may be larger than IDDA gross income because it includes some transfer income that is not reported on Form 1040, such as supplemental security income or SNAP benefits.

Differences in income between IDDA and existing sources may not be purely data-driven. Incomes reported in the CPS show a well-documented pattern of "mean reversion"—low-income respondents tend to overreport their income and high-income respondents tend to underreport it in a survey setting. High-income respondents often exclude nonwage sources that may comprise a meaningful portion of their annual income. Both income reporting and nonresponse in the ASEC is influenced by social dynamics between interviewer, interview participant, and other household

members that differ across race and gender.[3] All these reasons may help explain why gross income tends to be higher than total household income reported to the CPS, particularly at the highest-income percentiles and for White respondents. It's also possible that the CPS sample skews higher income for certain demographic groups relative to IDDA, and users will be able to explore this and other trends using the supplementary dataset.

Figure 3: Gross income and CPS total household income by race and ethnicity, 2012



Note: The red and grey bars show the 10th through 98th percentile of each group's income distribution in the CPS Annual Social and Economic Supplement and IDDA dataset, respectively. Income values from tax year 2012 reported in nominal dollars.
Source: Federal Reserve Bank of Minneapolis, Income Distributions and Dynamics in America.

[3] For an overview of how these and other factors influence earnings reporting and the measurement of inequality in the Current Population Survey, see Gideon, Heggeness, Murray-Close, and Myers (2021): https://doi.org/10.25071/1874-6322.40385

6. Citing the IDDA Data

The IDDA dataset recommended standard citation term is:

Kondo, Illenin, Brandon Hawkins, Kevin Rinz, John Voorheis, Andrew Goodman-Bacon, Natalie Gubbay, and Abigail Wozniak. (2023). Income Distributions and Dynamics in America: Version 1.0 [Dataset]. Minneapolis, MN: Federal Reserve Bank of Minneapolis.

Appendix B: Demographic Disaggregation Dimensions in IDDA

Percentiles of Income					
Geography (l)	Sample and income concept (k)		Demographic disaggregation (g)	y0 description	y0 quartiles+
U.S.	Individual W-2	TC, DC, WC	xall, xaged, xred, xsex, xfb, xagedXrea, xagedXsex, xredXsex	percentiles of income	p = 10, 25, 50, 75, 90, 95, 98, 99, 99.9, 99.99, 99.999
	Household-1040	GI, NW, WS, GI_ADJ	xall, xaged, xrea, xfb, xagedXrea		p = 10, 25, 50, 75, 90, 95, 98, 99, 99.9, 99.99, 99.999
	Prime-age workers W-2	TC, DC, WC	xall, xaged, xred, xsex, xfb, xagedXrea, xagedXsex, xredXsex		p = 10, 25, 50, 75, 90, 95, 98, 99, 99.9, 99.99, 99.999
State	Individual W-2	TC	xall, xaged, xrea, xsex, xfb, xagedXrea, xagedXsex, xreaXsex		p = 10, 25, 50, 75, 90, 95, 98
	Household-1040	GI, NW	xall, xaged, xrea, xfb		p = 10, 25, 50, 75, 90, 95, 98

Top income shares and population shares					
Geography (l)	Sample and income concept (k)		Demographic disaggregation (g)	y0 description	y0 quartiles+
U.S.	Individual W-2	TC, DC, WC	xall, xaged, xred, xsex, xfb, xagedXrea, xagedXsex, xredXsex	population shares across demographic groups top income shares within and across demographic groups above a percentile of income	p = 0 (across only), 90, 95, 98, 99, 99.9, 99.99, 99.999
	Household-1040	GI, NW, WS, GI_ADJ	xall, xaged, xrea, xfb, xagedXrea		p = 0 (across only), 90, 95, 98, 99, 99.9, 99.99, 99.999
	Prime-age workers W-2	TC, DC, WC	xall, xaged, xred, xsex, xfb, xagedXrea, xagedXsex, xredXsex		p = 0 (across only), 90, 95, 98, 99, 99.9, 99.99, 99.999
State	Individual W-2	TC	xall, xaged, xrea, xsex, xfb, xagedXrea, xagedXsex, xreaXsex		p = 0 (across only), 90, 95, 98
	Household-1040	GI, NW	xall, xaged, xrea, xfb		p = 0 (across only), 90, 95, 98

Dynamic Measures:

Income Change Distributions							
Geography (l)	Sample and income concept (k)		Demographic disaggregations (g)	y0 description	y0 quartiles+	y1 description	y1 quartiles+
U.S.	Individual W-2	TC, DC	xall, xaged, xrea, xsex, xfb, xagedXrea, xagedXsex, xredXsex	Income quartile within ktl, across g	lt25, 25t50, 50t75, 75t90, gt90	Percentiles of nominal income changes	10, 25, 50, 75, 90, mean
	Household-1040	GI, NW, WS	xall, xaged, xrea, xfb		lt25, 25t50, 50t75, 75t90, gt90		10, 25, 50, 75, 90, mean
	Prime-age workers W-2	TC, DC	xall, xaged, xrea, xsex, xfb, xagedXrea, xagedXsex, xredXsex		lt25, 25t50, 50t75, 75t90, gt90		10, 25, 50, 75, 90, mean
State	Individual W-2	TC	xall, xaged, xrea, xsex, xfb		lt25, 25t50, 50t75, gt75		10, 25, 50, 75, 90, mean
	Household-1040	GI, NW	xall, xaged, xrea, xfb		lt25, 25t50, 50t75, gt75		10, 25, 50, 75, 90, mean

Transition Matrix							
Geography (l)	Sample and income concept (k)		Demographic disaggregations (g)	y0 description	y0 quartiles+	y1 description	y1 quartiles+
U.S.	Individual W-2	TC, DC	xall, xaged, xrea, xsex, xfb, xagedXrea, xagedXsex, xredXsex	Income quartile within ktl, across g	miss, lt25, 25t50, 50t75, 75t90, gt90	Income quartile within ktl, across g	miss, lt25, 25t50, 50t75, 75t90, gt90
	Household-1040	GI, NW, WS	xall, xaged, xrea, xfb		miss, lt25, 25t50, 50t75, 75t90, gt90		miss, lt25, 25t50, 50t75, 75t90, gt90
	Prime-age workers W-2	TC, DC	xall, xaged, xrea, xsex, xfb, xagedXrea, xagedXsex, xredXsex		miss, out, lt25, 25t50, 50t75, 75t90, gt90		miss, out, lt25, 25t50, 50t75, 75t90, gt90
State	Individual W-2	TC	xall, xaged, xrea, xsex, xfb		miss, lt25, 25t50, 50t75, gt75		miss, lt25, 25t50, 50t75, gt75
	Household-1040	GI, NW	xall, xaged, xrea, xfb		miss, lt25, 25t50, 50t75, gt75		miss, lt25, 25t50, 50t75, gt75