IDDA Technical Documentation[1]

July 8, 2024

0 Release Notes

This document describes the data sources, sample selection, and construction of the over 6 million statistics in the Income Distributions and Dynamics in America (IDDA) dataset. IDDA is a collaborative research project of the U.S. Census Bureau and the Minneapolis Fed. IDDA statistics summarize extensive data from restricted Internal Revenue Service and Census Bureau records to advance our understanding of income distributions and mobility across demographic and geographic groups in the United States.

This version of data and documentation was published July 8, 2024. The previous version was published November 7, 2023. Updates to the available data include:

- Suppressed values are represented as blank cells for machine readability. Note, only the percentiles of income module includes cells representing suppressed statistics.

Updates to the documentation include:

- Additional tabulations of the coverage of IDDA statistics and correction to the availability of statistics in the transition matrix module, prime-age working sample (see section 5).
- Additional tabulations of IDDA sample size and sample selection (see table 1).
- Correction to the description of total compensation (see section 3).

Please reach out to MplsInstitute@mpls.frb.org with any further questions on the IDDA statistics or documentation.

[1] The opinions and conclusions expressed here are those of the authors and should not be interpreted as reflecting the views of the Federal Reserve Board of Governors, the Federal Reserve Bank of Minneapolis, or any other person associated with the Federal Reserve System. Any opinions and conclusions expressed herein are those of the authors and do not represent the views of the U.S. Census Bureau. The Census Bureau has ensured appropriate access and use of confidential data and has reviewed these results for disclosure avoidance protection (Project 7511151; Disclosure Authorization Numbers CBDRB-FY23-0277, CBDRB-FY23-0373, CBDRB-FY23-CES014-019, CBDRB-FY23-CES014-016, and CBDRB-FY24-0131.)

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 "protected identification keys" (PIK), which are privatized person identifiers assigned by the Person Validation System (PVS).[2] 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. PVS assigned keys also enable 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 other 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 the Census Bureau's 2020 Best Race and Ethnicity Administrative Records Composite File, the 2000 decennial census, the 2010 decennial

[2] See Wagner and Layne (2014) for more details.

census, and 2005-2019 American Community Survey (ACS). The Best Race and Ethnicity Administrative Records Composite 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 Best Race and Ethnicity Administrative Records Composite 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 and living at the same address but 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.[3] However, W-2 earnings records are still observed for non-filing individuals. 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. Statistics

[3] Note that 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). Some measures of total household income tend to be lower in these two years. For more details, see IRS Publication 1304 on Tax Returns for years 2007 and 2019 at this webpage.

in IDDA are not reweighted to be representative of the overall (filing and non-filing) population.

From the universe of 1040 and W-2 records, a limited number are excluded for the following data validation reasons:

- Records are excluded from the 1040 sample if they 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
- Records are excluded from the 1040 sample if they are not the primary or secondary filer on the 1040 form on which they are listed
- Records are excluded from the W-2 sample if they have missing values for any of the W-2 income variables
- Records are excluded from the W-2 sample if they have a value of zero for either wage income or total compensation reported on form W-2

The validated W-2 and 1040 tax records are linked to the Census Numident via the PIK. The Census Bureau's Person Identification Validation System (PVS) uses personally identifiable information (PII) (e.g. name, date of birth, etc.) to uniquely map social security numbers (SSNs) and individual taxpayer identification numbers (ITINs) into PIKs. As a result, datasets that contain SSNs or ITINs, such as tax records, have very high PIK assignment rates (97 percent or higher). PIK assignment rates for files that do not contain this information (such as data from the decennial census and ACS, which do not collect it) are somewhat lower (90 to 93 percent) but still high in absolute terms. False assignment rates, however, are extremely low (Layne et al., 2014) datasets naturally inherit these high PIK match rates, as shown by the large sample sizes in the first row of Table 1 for the year 2010.

The next rows of Table 1 show how the size of the W-2 and 1040 samples change after successive sample restrictions are implemented, again using tax year 2010 as an example. 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 information returns), 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 if they do not have an address that appears in the Census Master Address File (do not have a MAFID) or if there is an unusually high count of individual tax filers associated with the underlying MAFID (above the 99th percentile).[4]

Table 1 also shows the demographic composition in 2010 in IDDA data sources and in the Current Population Survey samples obtained through IPUMS CPS (see Flood et al., 2023). Overall, demographic shares are similar. The IDDA 1040 sample has fewer 16 to 24-year-old individuals than the W-2 sample or the CPS data. IDDA also has fewer Hispanic, foreign-born, or non-Hispanic NHOPI shares compared to shares in CPS samples. Non-Hispanic White and U.S.-born shares are higher in IDDA samples compared to these groups' shares in the CPS.

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 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 an additional sample of prime-age workers restricted to individuals aged 25-54 with earnings above a threshold, equivalent to working half-time for 13 weeks at the federal minimum wage, as measured by their wage compensation on form W-2. Table 2 shows the demographic composition of the IDDA prime-age working W-2 sample, compared with a Current Population Survey sample using the same earnings criteria. As in the full W-2 sample, demographic shares in IDDA are similar to those in the CPS. In the longitudinal modules, individuals are allowed to join or leave the prime-age working sample as they age or move above or below the minimum wage threshold. Demographic information is taken from the base year, meaning that individuals aged 16-24 can "age in" to a given income bracket in the prime-age earners sample and individuals aged 45-54 can "age out."

[4] For reference, in the 2010 American Community Survey, the median number of individuals aged 16 or older living in a given household was two. The 90th, 98th, and 99th percentiles of the number of individuals aged 16 or older were 4, 7, and 8 individuals per household, respectively. Above the 99th percentile, the distribution of adults per household quickly spreads out. This restriction helps to exclude individuals living in large group settings who are unlikely to be sharing resources as a household.

Table 1: IDDA Sample Sizes and Composition (2010)

	Household-1040	Individual-W2	CPS Household	CPS Individual
In Numident	182,200,000	150,400,000	–	–
Has age, gender, and state	181,000,000	146,700,000		
Has race/ethnicity	178,000,000	144,300,000		
Has valid MAFID	169,300,000	–		
Final Sample N	169,300,000	144,300,000	153,586	95,094
Demographic Composition				
Female	52.1%	49.6%	51.7%	48.0%
Male	47.9%	50.5%	48.3%	52.0%
Hispanic	11.3%	12.9%	14.7%	14.8%
Non-Hispanic AIAN	0.7%	0.9%	0.7%	0.6%
Non-Hispanic Asian	4.6%	4.6%	5.1%	5.1%
Non-Hispanic Black	10.1%	12.0%	11.5%	10.8%
Non-Hispanic NHOPI	0.1%	0.2%	0.3%	0.3%
Non-Hispanic Other	1.3%	1.5%	1.3%	1.2%
Non-Hispanic White	71.8%	68.0%	66.5%	67.1%
Foreign born	13.9%	13.2%	15.5%	15.8%
Not Foreign born	86.1%	86.8%	84.5%	84.2%
16-24	8.8%	16.4%	16.1%	13.6%
25-34	18.5%	21.2%	16.9%	21.9%
35-44	18.7%	20.2%	16.6%	21.3%
45-54	20.9%	21.9%	18.3%	22.5%
55-64	17.3%	15.3%	15.6%	15.9%
65+	15.8%	5.0%	16.5%	4.8%

Note: IDDA is built from the universe of W-2 and 1040 income tax returns merged to demographic and geographic information via the PIK. Table 1 shows the total number of PIK-level records in this underlying dataset after successive sample restrictions, beginning with our initial merge to the Census Numident. N sizes and the sample demographic composition are reported for 2010, a year in the middle of our data series, along with the corresponding values in the CPS ASEC. We do not expect linkage to race/ethnicity information to change substantially in decennial census years, because of the prioritization of race/ethnicity data sources described above.
Source: IDDA and IPUMS CPS. Release authorization CBDRB-FY24-0131.

Table 2: IDDA Prime-aged Working Sample Composition (2010)

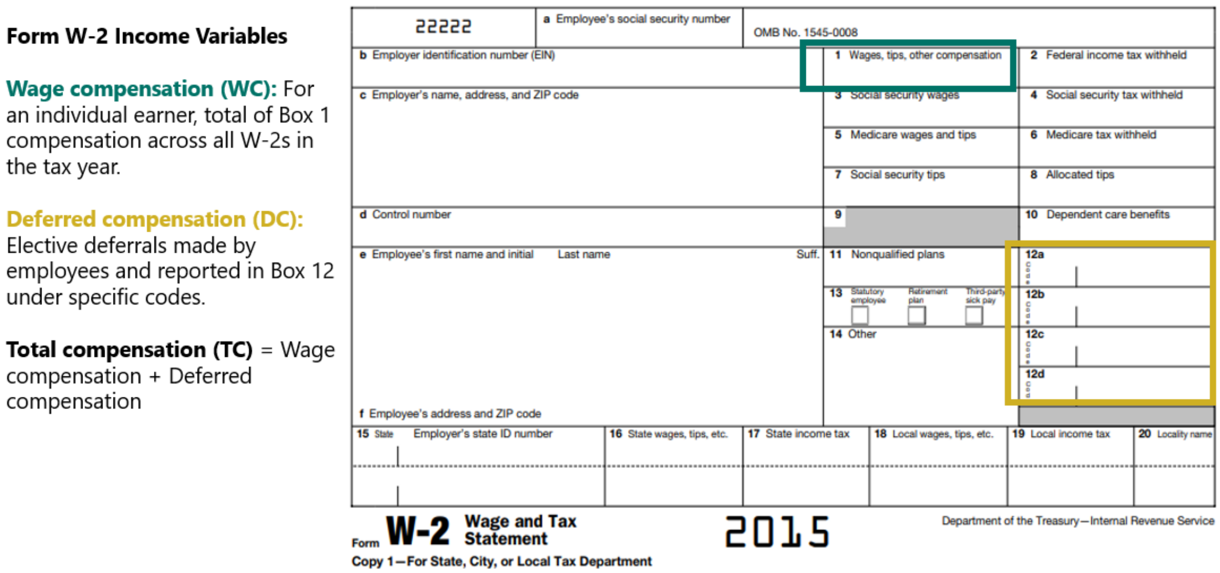
Demographic Composition	Prime-age Working W2	Prime-aged Working CPS
Female	49.0%	47.4%
Male	51.0%	52.6%
Hispanic	13.5%	16.3%
Non-Hispanic AIAN	0.9%	0.6%
Non-Hispanic Asian	5.2%	5.7%
Non-Hispanic Black	12.1%	11.2%
Non-Hispanic NHOPI	0.2%	0.3%
Non-Hispanic Other	1.4%	1.2%
Non-Hispanic White	66.6%	64.7%
Foreign born	15.2%	18.0%
Not Foreign born	84.8%	82.0%
25-34	33.0%	33.1%
35-44	32.1%	32.5%
45-54	34.9%	34.4%

Source: IDDA and IPUMS CPS.
Note: The IDDA Prime-age working-W2 sample includes individuals aged 25-54 with earnings at least equivalent to working 20 hours a week for 13 weeks at the federal minimum wage. In 2010, this was $1,885. Individuals in the 25-54 age bracket made up 63 percent of the overall W-2 sample in that year. The earnings threshold is much less restrictive: $1,885 fell substantially below the 10th percentile of individual earnings, which ranged from $4,174 for the 25-34 year-old group to $7,601 for the 45-54 year-old group. Earnings are measured using wage compensation reported on form W-2. Release authorization CBDRB-FY24-0131.

3 Income Variables

3.1 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, and is positive for all earners included in the W-2 sample. Deferred compensation includes elective deferrals recorded in Box 12 using specific codes–e.g. D, E, F, G, H, AA, BB, or EE–and may be positive or zero. Elective deferrals include contributions made out of an employee's own wages, not by an employer. While a worker's total compensation includes other components such as employer-sponsored health care, we only observe these two components and refer to their sum as a pseudo "total compensation." The actual total compensation includes items we do not observe, such as the cost of employer-sponsored health coverage recorded in Box 12 using code DD. Total compensation is positive for all earners included in the W-2 sample.



Form W-2 Income Variables

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

Deferred compensation (DC): Elective deferrals made by employees and reported in Box 12 under specific codes.

Total compensation (TC) = Wage compensation + Deferred compensation

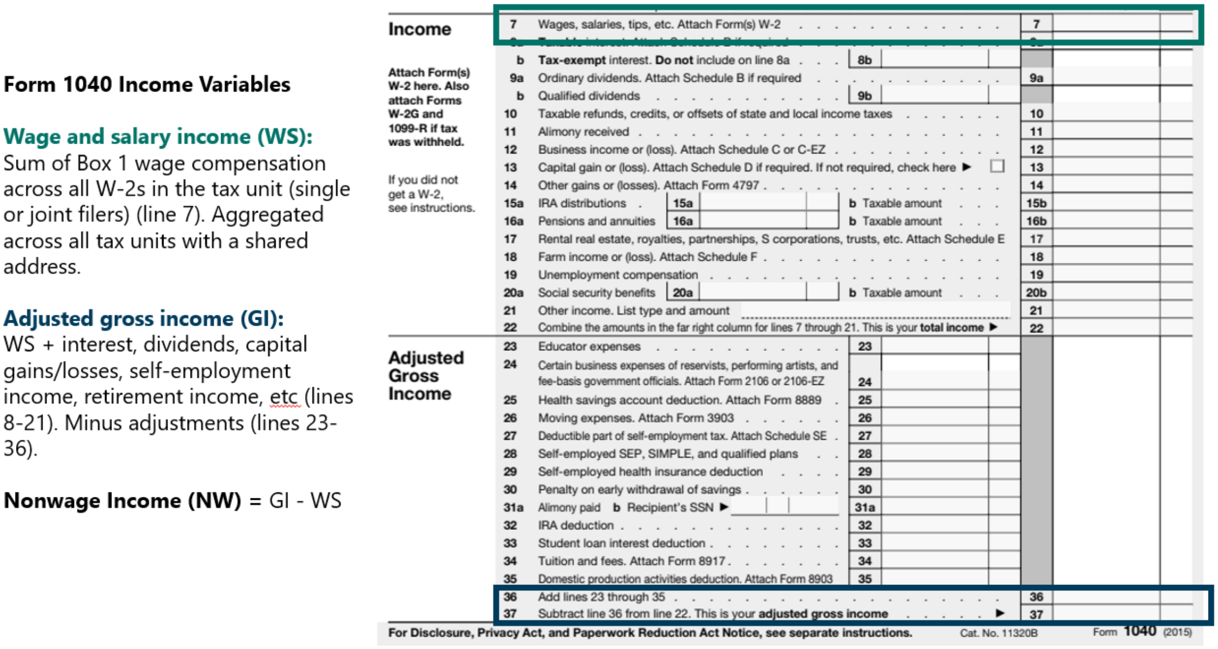
22222
a Employee's social security number
OMB No. 1545-0008
b Employer identification number (EIN)
1 Wages, tips, other compensation
2 Federal income tax withheld
c Employer's name, address, and ZIP code
3 Social security wages
4 Social security tax withheld
5 Medicare wages and tips
6 Medicare tax withheld
7 Social security tips
8 Allocated tips
d Control number
9
10 Dependent care benefits
e Employee's first name and initial Last name Suff.
11 Nonqualified plans
12a Code
13 Statutory employee □ Retirement plan □ Third-party sick pay □
12b Code
14 Other
12c Code
12d Code
f Employee's address and ZIP code
15 State Employer's state ID number
16 State wages, tips, etc.
17 State income tax
18 Local wages, tips, etc.
19 Local income tax
20 Locality name

Form **W-2** **Wage and Tax Statement**
2015
Department of the Treasury—Internal Revenue Service
Copy 1—For State, City, or Local Tax Department

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

The 1040 sample includes household-level adjusted gross income (GI), wage and salary income (WS), and non-wage 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).

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



Form 1040 Income Variables

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

Attach Form(s) W-2 here. Also attach Forms W-2G and 1099-R if tax was withheld.

If you did not get a W-2, see instructions.

7 Wages, salaries, tips, etc. Attach Form(s) W-2 7
8a Taxable interest. Attach Schedule B if required 8a
b **Tax-exempt** interest. **Do not** include on line 8a . . . 8b
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
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)

Figure 2: Household Income Variables Derived from Form 1040

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.

Non-wage income is the difference between gross income and wage and salary income. Wage and salary income can be positive or zero, while non-wage income and gross income can be positive, negative, or zero.

3.2 Aggregation from Tax Units to Households

According to the Census Bureau definition, "a household consists of all the people who occupy a housing unit." Income variables as reported on Form 1040 represent total income earned by members of a tax unit. 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 an equivalized version of gross income 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, Top Income Population Shares, Income Change Distributions, and Income Transition Matrices. The first three modules measure income distributions in a given year while the last two 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 granularity (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 4.1–4.5. Section 6 summarizes the dimensions of demographic granularity 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 set of 30 observations around the pth 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 4.6 titled "Suppression in the IDDA data" for a detailed discussion.

4.1 Percentiles of Income Module

The percentiles of income module contains group-percentile income values at the 10th, 25th, 50th, 75th, 90th, 95th, and 98th percentile (state and U.S.-level) and 99th, 99.9th, 99.99th, and 99.999th 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 90th percentile of total W-2 compensation among Hispanic men in Illinois was $88,360.

Percentiles of household adjusted gross income and household non-wage 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.

4.2 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 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.

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

4.3 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 pth percentile of income, regardless of demographic group. The numerator is the count of individuals at or above the pth 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.

4.4 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 75th percentile) is split into two smaller subsets at the national level: the 75-90th percentile of initial income, and above the 90th 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 an increase of $2000 and the 90th percentile was an increase of $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 a gain of $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 non-wage income) can take zero or negative values.

4.5 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. Geography is defined in the initial year. That is, an individual who moves from state A to B is included in the transition matrix for state A and their subsequent year income is placed within a quartile based on the distribution across all individuals initially in state A.

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.

4.6 Suppression in the IDDA data

The IDDA statistics are only published based on underlying samples and implicit samples that meet a minimum size of 30 observations. This threshold is implemented at the statistic level. For example, in order to disclose the share of income received by the top 10 percent of earners within a certain demographic group and state, IDDA requires at least 30 individuals be in the top 10 percent, implying that there are at least 300 individuals in the demographic group and state in total.

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,

Table 3: 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

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 percent. 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 3. Note that the suppression rules try to preserve smaller race/ethnicity groups whenever possible instead of systematically favoring the larger groups.

5 Coverage and Limitations

5.1 Coverage of Statistics in IDDA

IDDA statistics are not available for all feasible demographic groups or places. The levels of granularity included in IDDA vary by module, geography, and sample. Tables 4 and 5 present 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. Availability is lower at higher levels of granularity and in particular the intersection of age and race, which 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 "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 4: Availability of Statistics by Demographic Group: Form 1040 data (1998–2019)

IDDA Module	Defined	All	Age	BPL	Race	AgeXRace
US Household-1040						
Income Levels	154,176	100	100	100	90.8	83.9
Income Changes	51,300	100	100	100	100	
Transition Matrix	59,850	100	100	100	100	
State Household-1040						
Income Levels	587,928	100	100	100	94.6	
Income Changes	1,395,360	100	100	100	94.9	
Transition Matrix	1,395,360	100	100	100	92.2	

Note: Columns provide the total number of defined statistics in each IDDA sample and geography, along with the corresponding availability rate in percentages. Each statistic is defined by a geography, sample and income concept, year(s), and demographic group. For example, the total number of transition matrix statistics defined at the state level includes all possible transitions between income quartiles, multiplied by 50 states, multiplied by the total number of income concepts, demographic groups, and pairs of years for which that statistic is computed. Availability rates for statistics that are not disaggregated by demographic group are reported in the "All" column. BPL represents place of birth (U.S. or foreign-born). Release authorization CBDRB-FY23-0277.
Source: IDDA

Table 6 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 Tables 4 and 5, coverage is lower when more demographic intersections are included and where statistics are reported further into the tail of the income distribution.

However, availability is quite high at the national level, even for small race and ethnicity groups. In addition, it is rare that a group/state combination is excluded from the IDDA data entirely. For example, using 2010 as an example year, percentiles of income from Form W-2 are not available for the Native Hawaiian or other Pacific Islander group in Delaware, Rhode Island, Vermont, or the

Table 5: Availability of Statistics by Demographic Group: Form W-2 Data (2005-2019)

IDDA Module	Defined	All	Age	BPL	Race	Sex	AgeXRace	AgeXSex	RaceXSex
US Individual-W2									
Income Levels	121,680	100	97	100	91.2	100	80.6	91.9	86.7
Income Changes	110,880	100	100	100	100	100	99.3	100	100
Transition Matrix	129,360	100	100	100	100	100	94.8	100	99.3
US PAW-W2									
Income Levels	80,370	100	99.5	100	90.5	100	85.1	94.6	84.8
Income Changes	72,000	100	100	100	100	100	100	100	100
Transition Matrix	120,816	100	100.0	100.0	99.9	100	95.4	100.0	98.7
State Individual-W2									
Income Levels	1,054,170	100	97.9	100	95.8	100	74.9	93.2	89.8
Income Changes	499,392	100	100	100	95.9	100			
Transition Matrix	499,392	100	98.8	99.6	86.2	100			

Note: Columns provide the total number of defined statistics in each IDDA sample and geography, along with the corresponding availability rate in percentages. Each statistic is defined by a geography, sample and income concept, year(s), and demographic group. For example, the total number of transition matrix statistics defined at the state level includes all possible transitions between income quartiles, multiplied by 50 states, multiplied by the total number of income concepts, demographic groups, and pairs of years for which that statistic is computed. Availability rates for statistics that are not disaggregated by demographic group are reported in the "All" column. BPL represents place of birth (U.S. or foreign-born). Release authorization CBDRB-FY23-0277.
Source: IDDA.

District of Columbia. However, the population represented by these statistics was 0.38 percent of the total working NHOPI population. Looking at all combinations of race and age for which IDDA did not include any percentiles of income in 2010, these accounted for only 0.015 percent of the total working U.S. population. Higher in the distribution, suppression is more common, but even here the states where statistics are suppressed represent small fractions of the total population. The 98th percentile of income was suppressed for NHOPI earners in 21 states and AIAN earners in 2 states, but those states represented only 4.57 percent of the total working NHOPI population and 0.11 percent of the total working AIAN population, per ACS estimates.

5.2 Limitations

IDDA users should note the following limitations of the IDDA statistics:

1. IDDA contains only pre-tax and taxable incomes from tax returns. IDDA income measures therefore do not reflect informal and non-taxable incomes.

2. IDDA income measures do not contain public transfers to low-income households, in particular.

3. IDDA does not have a breakdown of non-wage income by type (for instance, isolating Schedule C

Table 6: IDDA Availability by Race/Ethnicity

Module	Intersections	Hispanic	AIAN	Asian	Black	NHOPI	White
US-Level							
Household-1040							
Income Levels	age	91.7	76.4	89.4	88.4	68.9	95.4
Income Changes	-	100	100	100	100	100	100
Transition Matrix	-	100	100	100	100	100	100
Individual-W2							
Income Levels	age, sex	89.9	74.1	87.7	90.6	66.4	92.3
Income Changes	age, sex	100	99.6	100	100	97.5	100
Transition Matrix	age, sex	100	95.0	99.9	100	83.5	100
PAW-W2							
Income Levels	age, sex	93.8	77.0	90.5	93.2	70.9	94.0
Income Changes	age, sex	100	100	100	100	100	100
Transition Matrix	age, sex	100	96.2	100	99.6	87.4	100
State-Level							
Household-1040							
Income Levels	-	99.9	96.8	99.9	98.2	72.5	100
Income Changes	-	100	97.1	99.9	98.8	73.5	100
Transition Matrix	-	99.8	95.1	98.9	97.7	61.8	100
Individual-W2							
Income Levels	age, sex	91.3	76.5	88.8	87.8	44.0	95.0
Income Changes	-	100	98.2	100	99.9	77.5	100
Transition Matrix	-	97.5	83.6	94.5	94.4	47.1	100

Source: IDDA.
Note: Table 5 reports availability rates for individual race/ethnicity groups in IDDA. The column “intersections” indicates whether statistics in the geography and sample are provided for the intersection of age and race, race and sex, or neither. In general, statistics are provided further into the tail of the income distribution of the U.S.-level data than that of the state-level data. Income levels are reported up to the 98th percentile at the state level and 99.999th percentile nationally. Transition matrices and conditional income changes are calculated within quartiles of the state-level income distributions. At the U.S. level, the top income quartile is divided into two smaller brackets (the 75th-90th percentile and above the 90th percentile). Release authorization CBDRB-FY23-0277.

self-employment income from other non-wage sources). Given the large variety of taxable non-wage incomes and the accounting rules involved, caution should be exercised when interpreting non-wage income statistics.

4. The IDDA statistics cover only individuals who filed a Form 1040 or received a Form W-2 in a given year. The statistics are not weighted to be representative of the full population. In other words, the statistics are representative of tax filers nationally and within each state, but they are not representative of all U.S. or state residents.

5. IDDA statistics are not available for all feasible demographic groups or places. For a list of the levels of demographic granularity available in IDDA, see section 6.

6. IDDA does not contain any statistics on wealth.

6 Levels of Demographic Granularity in IDDA

Tables 7, 8, 9, and 10 specify the categories defined along the k, l, g dimensions of IDDA statistics. For more information on variable definitions and abbreviations in the IDDA datasets, please see the IDDA codebook.

Table 7: Percentiles of Income Module

Geography (l)	Sample & income concept (k)		Demographic groups (g)	y0 quantiles
U.S.	Individual 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
	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 working 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

Note: The IDDA percentiles of income module contains selected percentiles (listed in y0 quantiles) of each group-specific income distribution at the levels of demographic granularity reported in the table. Statistics are cross-sectional and include each year from 2005 to 2019 (W-2 samples) and 1998-2019 (1040 samples).

Table 8: Top income shares and population shares

Geography (l)	Sample & income concept (k)		Demographic groups (g)	y0 quantiles
U.S.	Individual 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
	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 working 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

Note: The IDDA top income shares and top income population shares modules summarize the concentration and the demographic composition of incomes above a given percentile (listed in y0 quantiles). Statistics are cross-sectional and include each year from 2005 to 2019 (W-2 samples) and 1998-2019 (1040 samples).

Table 9: Income change distributions

Geography (l)	Sample & income concept (k)		Demographic groups (g)	y0 quantiles	y1 quantiles
U.S.	Individual W-2	TC, DC	xall, xaged, xrea, xsex, xfb, xagedXrea, xagedXsex, xredXsex	lt25, 25t50, 50t75, 75t90, gt90	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 working 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

Note: The IDDA income change distributions module reports selected percentiles of the income growth distribution (listed in y1 quantiles) for individuals in an initial year income bin (listed in y0 quantiles). Statistics are longitudinal and include 1- and 5-year time horizons from 2005 to 2019 (W-2 samples) and 1998-2019 (1040 samples).

Table 10: Transition Matrices

Geography (l)	Sample & income concept (k)		Demographic groups (g)	y0 quantiles	y1 quantiles
U.S.	Individual W-2	TC, DC	xall, xaged, xrea, xsex, xfb, xagedXrea, xagedXsex, xredXsex	miss, lt25, 25t50, 50t75, 75t90, gt90	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 working 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

Note: The IDDA transition matrix module reports the probability that an individual moves from a given initial year income bin (listed in y0 quantiles) to a different income bin (listed in y1 quantiles). Statistics are longitudinal and include 1- and 5-year time horizons from 2005 to 2019 (W-2 samples) and 1998-2019 (1040 samples).

7 Native Areas Geography and Data Sources

IDDA includes over 70,000 statistics on income distributions, income mobility, and migration for both Native and non-Native populations living in Native areas. This document provides additional information on the definition and construction of these statistics. However, the administrative sources IDDA is built from do not fully capture the complexity and fluidity of Native land areas or Native identity. Data users can learn more about understanding and interpreting Native incomes in IDDA using the resources in Huff et al. (2023).

The IDDA Native areas geography provides statistics derived from IRS Forms W-2 and 1040 within an aggregate geography that includes all American Indian, Alaska Native, and Native Hawaiian (AIANNH) areas as delineated by the Census Bureau in 2017. The Census Bureau's Master Address File identifies whether a given address (MAFID) falls within a Census block that is designated with a 4-character AIANNH area census code (aiannhce code). The aiannhce code specifies individual non-overlapping reservations, trust lands, and other legal and statistical entities as defined by the Census Bureau (Table 11).

The aiannhce code takes a value of 9999 if the Census block is not designated as a Native area. We convert this information to a 0/1 variable which flags whether individuals in the tax sample resided in a Native area in a given year based on their MAFID in that year.

Demographic variables are drawn from the same sources as in the main IDDA dataset with the exception of race/ethnicity. The main dataset reports income statistics for individuals who identify as non-Hispanic American Indian or Alaska Native only and for individuals who identify as Native Hawaiian or other Pacific Islander only. The Native areas geography uses a more expansive notion of race and Native identity that includes all individuals who identify one of their races as American Indian, Alaska Native, Native Hawaiian, or other Pacific Islander, regardless of Hispanic ethnicity. We use self-reported primary and secondary races from the most recent decennial census (2010 and 2000) or American Community Survey a record appears in. If a record does not appear in either the American Community Survey or decennial census, we use the Census Best Race and Ethnicity Administrative Records Composite File which draws information from a variety of administrative sources to determine a race and Hispanic ethnicity for individuals in the Census Bureau data system.

Table 11: Definitions for AIANNH Areas Included in Aggregate Native Areas Geography

Area	Census Bureau Definition
American Indian reservations (Federal)	Areas that have been set aside by the United States for the use of tribes and whose boundaries are defined by tribal treaties, agreements, executive orders, federal statutes, secretarial orders, or judicial determinations. The Census Bureau recognizes federal reservations (and associated off-reservation trust lands) as territory over which American Indian tribes have primary governmental authority.
Off-reservation trust lands	Areas for which the United States holds title in trust for the benefit of a tribe (tribal trust land) or for an individual American Indian (individual trust land).
Hawaiian Home Lands	Areas held in trust for Native Hawaiians by the State of Hawaii.
Oklahoma Tribal Statistical Areas	Statistical areas identified and delineated by the Census Bureau in consultation with federally recognized American Indian tribes that had a former reservation in Oklahoma.
Alaska Native Village Statistical Areas	Statistical geographic entities representing permanent and/or seasonal residences of Alaska Natives who are members of, or receive governmental services from, the defining Alaska Native village. ANVSAs are intended to include only an area where Alaska Natives, especially members of the defining Alaska Native Village, represent a substantial proportion of the population during at least one season of the year.
Tribal Designated Statistical Areas	Statistical entities identified and delineated for the Census Bureau by federally recognized American Indian tribes that do not currently have a federally recognized land base (reservation or off-reservation trust land). Generally encompasses a compact and contiguous area that contains a concentration of individuals who identify with a federally recognized American Indian tribe and in which there is structured or organized tribal activity.
American Indian reservations (State)	Reservations established by some state governments for tribes recognized by the state.
State Designated Tribal Statistical Areas	Statistical entities for state-recognized American Indian tribes that do not have a state-recognized land base (reservation). Generally encompasses a compact and contiguous area that contains a concentration of individuals who identify with a state recognized American Indian tribe and in which there is structured or organized tribal activity

Note: Definitions are taken from the glossary of Census geographic programs and products, publicly available on the Census website: https://www.census.gov/programs-surveys/geography/about/glossary.html

7.1 Sample Selection

The Native areas statistics are produced from a subset of the final, valid W-2 and 1040 samples used to construct the main IDDA dataset. In the cross-sectional data, they are produced from the subset of W-2 or 1040 records whose address identifier links to an aiannhce code for a Native area in the given year. Records that do not link to a valid MAFID or whose MAFID is missing an aiannhce code are excluded from analysis.

Like in the main IDDA dataset, the Native areas geography includes income dynamics over 1- and 5-year time horizons. Records appear in the longitudinal files if the individual is in the valid W-2 (or 1040) sample in one of the two years and reside in a Native area in one of the two years. Thus, individuals can migrate in/out of the Native areas sample similarly to the way that individuals can age in/out of the prime-age working sample in the main dataset. Individuals are considered "out-of-sample" if they are in the relevant tax sample but do not reside in a Native area, and their income values in that year are set to missing. Individuals that are not in the relevant tax sample are considered "missing" in that year regardless of whether they live in a Native area. Statistics in the Native areas geography do not include a prime-age working subsample.

7.2 Income Variables

Three income concepts are available in the Native areas geography: household adjusted gross income and household nonwage income from Form 1040 and individual total compensation from Form W-2 (see section 3).

As in the overall W-2 and 1040 samples, some individuals in the Native areas sample are affiliated with a tribe and some are not. IRS Publication 5424 "Income Tax Guide for Native American Individuals and Sole Proprietors" provides detailed income reporting instructions for individuals with tribal affiliations. The following sources of income may be particularly relevant in the Native areas sample:

- Schedule C self-employment income is included in household nonwage income.
- General welfare payments made by tribal governments to individuals are not reported on Form 1040.

- Per-capita distributions (for example, dividends from trust lands or gaming activity) made by tribal governments are included in household nonwage income.
- The taxable portion of Alaska Permanent Fund dividends are included in household nonwage income.

The Native areas geography includes the same five statistics modules as the main dataset: Percentiles of Income, Top Income Shares, Top Income Population Shares, Income Change Distributions, and an augmented version of the Income Transition Matrix module that includes migration. Like in main IDDA, statistics in the Native areas geography exist at the $ktlg$ level, where

- k is a sample and corresponding income concept (in this case, the Native areas-household-1040 or Native areas-individual-W-2 sample),
- t is the time horizon (1998 to 2019 + 1- and 5-year horizons),
- l is the geography (Native areas aggregate geography), and
- g is a demographic group (overall, age, sex, Native identity, or an intersection).

All modules adhere to the rule that statistics are always calculated within k, t, and l. That means all income bins—for example, when population shares are reported for the top 10 percent of the income distribution—are defined within the Native areas geography and sample, not across the full U.S. population. The statistics included in the five modules are defined in section 4.

Augmented Transition Matrix In the main IDDA dataset, the Income Transition Matrix Module gives the probability than an individual in a given demographic group and initial income bin moves to a different income bin, or out of the W-2 or 1040 sample, over 1 or 5 years (from y0 to y1). Geography is defined in the base year: an individual who moves from state A to B is included in the transition matrix for state A and their subsequent year income is placed within a quartile based on the distribution across all individuals initially in state A.

The Native areas transition matrices track migration as well as income mobility. Individuals are included in the transition matrix if they reside in a Native area in either the base or subsequent year, and their income quartile is considered "out-of-sample" in the year they do not reside in a Native area. These augmented transition matrices provide several probabilities:

- The probability that an individual in a particular demographic group and initial income quartile is in another quartile of the Native areas income distribution in y1
- The probability that an individual in a particular demographic group and initial income quartile is not in the relevant tax sample in y1.
- The probability that an individual starting in a given quartile of the Native areas income distribution does not live in the Native areas geography in y1.
- The distribution across y1 income quartiles of recent movers to the Native areas geography.

Table 12 summarizes the interpretation for each of these types of transitions, as well as transitions in and out of the W-2 and 1040 data.

7.3 Coverage

Table 13 provides availability rates for statistics in the IDDA Native areas geography by module and sample. Table 14 provides availability of statistics by race and ethnicity, including intersections of Native identity with age or sex. Statistics are suppressed using the same method and minimum cell size described in section 4. 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.

The IDDA Native areas geography includes statistics by age, sex, Native identity, the intersection of age and Native identity, and the intersection of sex and Native identity. Tables 15 and 16 shows the levels of demographic granularity included for each module and sample. The selected income concepts, demographic groups, and quantiles follow those included in the state-level IDDA data.

Table 12: Native Areas Transition Matrix Module: Income Mobility and Migration

pctl_y0	pctl_y1	In y0 Native areas sample?	In y1 Native areas sample?	Interpretation
miss	lt25-gt75	No	Yes	The individual is not in the relevant tax sample in y0, and may or may not live in a Native area.
				In y1, the individual is in the tax sample, lives in a Native area, and has income in quartile pctl_y1 of the income distribution in Native areas in y1.
out	lt25-gt75	No	Yes	The individual is in the relevant tax sample in y0, but does not live in a Native area. In y1, the individual is in the tax sample, lives in a Native area, and has income in quartile pctl_y1.
out	miss	No	No	The individual is in the relevant tax sample in y0, but does not live in a Native area. In y1, the individual lives in a Native area and is in the IRS dataset but not the relevant tax sample.
out	out	No	No	Not reported
miss	miss	No	No	Not reported
miss	out	No	No	The individual is not in the relevant tax sample in y0, but is in the IRS dataset and lives in a Native area. In y1, the individual is in the relevant tax sample but does not live in a Native area.
lt25-gt75	out	Yes	No	In y0, the individual is in the tax sample, lives in a Native area, and has income in quartile pctl_y0 of the income distribution in Native areas in y0. In y1, the individual is in the tax sample but not in a Native area.
lt25-gt75	miss	Yes	No	In y0, the individual is in the tax sample, lives in a Native area, and has income in quartile pctl_y0 of the income distribution in Native areas in y0. In y1, the individual is not in the tax sample and may or may not live in a Native area.
lt25-gt75	lt25-gt75	Yes	Yes	In y0, the individual is in the tax sample, lives in a Native area, and has income quartile pctl_y0 of the income distribution in Native areas in y0. In y1, the individual is in the tax sample, lives in a Native area, and has income quartile pctl_y1 of the income distribution in Native areas in y1.

Note: The transition matrix module in IDDA provides the probability that an individual moves between given quantiles of the income distribution or out of the tax sample between pairs of years. In the Native areas geography, the transition matrix also gives information on conditional migration in and out of Native areas. This table provides guidance on interpreting these additional types of transitions. For a detailed definition of the statistics in each IDDA module, see section 4. For complete record layout information, see the IDDA Native areas codebook.

Table 13: IDDA Availability: Native areas Geography

Module	Defined	All	Age	Native	Sex	Age X	Sex X
				Identity		Identity	Identity
Native areas-Household 1040							
Income Levels	16,280	100	100	100		100	
Income Changes	16,416	100	100	100			
Transition Matrix	23,256	100	100	100			
Native areas-Individual W-2							
Income Levels	7,170	100	100	100	100	96	100
Income Changes	6,336	100	100	100	100		
Transition Matrix	8,976	100	100	100	100		

Note: Columns provide the total number of defined statistics in each IDDA sample and geography, along with the corresponding availability rate in percentages. Each statistic is defined by a geography, sample and income concept, year(s), and demographic group. Availability rates for statistics that are not disaggregated by demographic group are reported in the "All" column. Release authorization CBDRB-FY23-0373.

Table 14: Availability of Statistics by Native Identity

Module	Intersections	Identifies as Native	Does not identify as Native
Native areas-Household 1040			
Income Levels	age	100	100
Income Changes	-	100	100
Transition Matrix	-	100	100
Native areas-Individual W-2			
Income Levels	age, sex	97.4	97.4
Income Changes	-	100	100
Transition Matrix	-	100	100

Source: IDDA.

Note: Table 5 reports availability rates for Native and non-Native identity groups in the IDDA Native areas geography. The column "intersections" indicates whether statistics in the geography and sample are provided for the intersection of age and race, race and sex, or neither. Income levels are reported up to the 98th percentile in the Native areas geography. Transition matrices and conditional income changes are calculated within quartiles of the Native areas income distribution. Release authorization CBDRB-FY23-0373.

Table 15: Static Measures

Module	Sample & income concept (k)		Demographic groups (g)	y0 quantiles
Percentiles of Income	Native areas-Individual W-2	TC	xall, xaged, xaiannh, xsex, xagedXaiannh, xaiannhXsex	p = 10, 25, 50, 75, 90, 95, 98
	Native areas-Household-1040	GI, NW	xall, xaged, xaiannh, xagedXaiannh	p = 10, 25, 50, 75, 90, 95, 98
Income and Population Shares	Native areas-Individual W-2	TC	xall, xaged, xaiannh, xsex, xagedXaiannh, xaiannhXsex	p = 0 (across only), 90, 95, 98
	Native areas-Household-1040	GI, NW	xall, xaged, xaiannh, xagedXaiannh	p = 0 (across only), 90, 95, 98

Note: The IDDA percentiles of income module contains selected percentiles (listed in y0 quantiles) of each group-specific income distribution at the levels of demographic granularity reported in the table. The IDDA top income shares and top income population shares modules summarize the concentration and the demographic composition of incomes above a given percentile (listed in y0 quantiles). Statistics are cross-sectional and include each year from 2005 to 2019 (W-2 samples) and 1998-2019 (1040 samples).

Table 16: Dynamic Statistics

Module	Sample & income concept (k)		Demographic groups (g)	y0 quantiles	y1 quantiles
Income Change Distributions	Native areas-Individual W-2	TC	xall, xaged, xaiannh, xsex	lt25, 25t50, 50t75, gt75	10, 25, 50, 75, 90, mean
	Native areas-Household-1040	GI, NW	xall, xaged, xaiannh	lt25, 25t50, 50t75, gt75	10, 25, 50, 75, 90, mean
Transition Matrix	Native areas-Individual W-2	TC	xall, xaged, xaiannh, xsex	miss, out, lt25, 25t50, 50t75, gt75	miss, out, lt25, 25t50, 50t75, gt75
	Native areas-Household-1040	GI, NW	xall, xaged, xaiannh	miss, out, lt25, 25t50, 50t75, gt75	miss, out, lt25, 25t50, 50t75, gt75

Note: The IDDA transition matrix module reports the probability that an individual moves from a given initial year income bin (listed in y0 quantiles) to a different income bin (listed in y1 quantiles). The IDDA income change distributions module reports selected percentiles of the income growth distribution (listed in y1 quantiles) for individuals in an initial year income bin (listed in y0 quantiles). Statistics are longitudinal and include 1- and 5-year time horizons from 2005 to 2019 (W-2 samples) and 1998-2019 (1040 samples).

References

Flood, Sarah, Miriam King, Renae Rodgers, Steven Ruggles, J. Robert Warren, and Michael Westberry, *Integrated Public Use Microdata Series, Current Population Survey: Version 11.0 [dataset]* Minneapolis, MN: IPUMS 2023. https://doi.org/10.18128/D030.V11.0.

Huff, Andrew, H Trostle, Natalie Gubbay, and Illenin O. Kondo, "Understanding and Interpreting Native Incomes in IDDA," Income Distributions and Dynamics in America (IDDA) Series, Federal Reserve Bank of Minneapolis, Minneapolis MN 2023. (Published on November 13, 2023). https://www.minneapolisfed.org/article/2023/understanding-and-interpreting-native-incomes-in-idda.

Layne, Mary, Deborah Wagner, and Cynthia Rothhaas, "Estimating Record Linkage False Match Rate for the Person Identification Validation System," CARRA Working Paper No. 2014-02, U.S. Census Bureau jul 2014.

Wagner, Deborah and Mary Layne, "The Person Identification Validation System (PVS): Applying the Center for Administrative Records Research and Applications (CARRA) Record Linkage Software," CARRA Working Paper No. 2014-01, U.S. Census Bureau jul 2014.