SOI Migration Data: A New Approach
Methodological Improvements for SOI's United States Population Migration Data, Calendar Years 2011-2012

The Statistics of Income Division (SOI) of the Internal Revenue Service (IRS), in conjunction with the U.S. Census Bureau, has released migration data for the United States for several decades. These migration data are an important source of information detailing the movement of individuals from one location to another. SOI bases these data on the year-to-year changes of the addresses reported on individual income tax returns filed and processed with the IRS during two consecutive calendar years. Migration data prior to calendar years 2011-2012, were prepared by the Census Bureau for SOI. Beginning with the 2011-2012 migration data, the SOI has assumed the production of the migration tabulations. As part of this transition, SOI has undertaken a number of enhancements intended to improve the data's overall quality, as well as to provide a new series of information. This paper discusses those improvements and highlights some of the differences between the new migration data, starting in 2011-2012, with previous versions[1].

Three major improvements of the 2011-2012 SOI migration data:

- Migration data based on a full year of data, as opposed to a partial year of data
- Improved year-to-year return matching increased the number of matched records by 5 percent
- New tabulations that show migration flows at the State level, by size of adjusted gross income (AGI), and age of the primary taxpayer

Full-Year Data

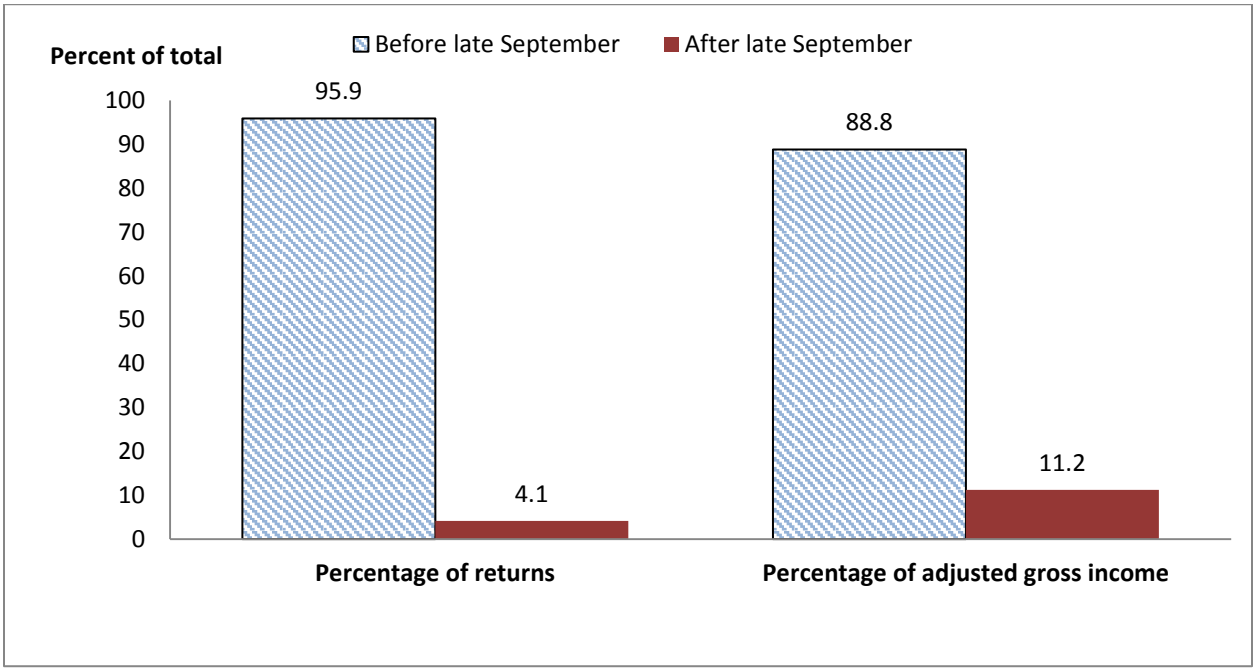
A major improvement made to the migration data was the switch from partial-year data to full-year data. Due to their internal production deadlines, migration data produced by Census for SOI only included individual income tax returns filed roughly by the end of September. Since SOI is not bound by those same deadlines, it can construct the migration tabulations using all individual income tax returns received during the full calendar year (January 1 to December 31)[2].

For 2012, some 4 percent of the returns the IRS received and processed came in after the September cutoff (Figure A). Also, the total amount of adjusted gross income (AGI) reported for returns received after September was equivalent to 11 percent of the total AGI in 2012. To the extent that the returns filed after September are somewhat different from the returns filed before September (more complex and higher income, for instance), including them in the migration data removes a potential bias. Another way of stating this is that by including the returns received after late September, the new SOI migration data represent the migration patterns for the full filing population of current tax year returns.

[1] For a full description of the 2011-12 migration data, see the Migration Data Users Guide available at http://www.irs.gov/uac/SOI-Tax-Stats-Migration-Data. For a description of the previous migration data, see Gross, Emily "U.S. Population Migration Data: Strengths and Limitations."

[2] The new migration data include all individual income tax returns processed from January 1 to December 31 that were for the current tax year. A small percentage (usually around 3 percent) of the returns that the IRS receives are for prior tax years. For matching purposes, SOI excludes prior year tax returns from the migration data.

Figure A: Percentage of Returns and Adjusted Gross Income Received Before and After Late September, Calendar Year 2012



Improved Year-to-Year Return Matching

Prior to the 2011-2012 migration data, returns were matched based on the primary filer's taxpayer identification number (TIN) only. One drawback of this approach was that the filing position for many individuals changes over time. For instance, individuals may be listed as a dependent on a parent or guardian's return in Year 1, but file their own return as a primary filer in Year 2. Similarly, an individual listed as a secondary filer on, for example, a married filing joint return in Year 1, could become a single filer (primary taxpayer) in Year 2 due to a divorce or death. Finally, some married couples switch their positions each year, the primary becoming the secondary and vice versa. Under the previous methodology, records from the migration data that had experienced such changes in their filing positions were excluded.

To address this limitation, the matching process, under the new methodology, uses the TINs of the primary, secondary, and dependent filers. The following represents the new matching process:

Figure B:

	Year 1		Year 2	Percent of the total matched returns
1.	Primary filer	→	Primary filer	94.6%
2.	Primary filer	→	Secondary filer	0.8%
3.	Secondary filer	→	Primary filer	1.7%
4.	Secondary filer	→	Secondary filer	less than 0.1%

5. Dependent filers → Primary filer 2.8%
6. Dependent filers → Secondary filer less than 0.1%

Under the new methodology, approximately 5 percent of the total matched records occur through additional matching beyond the primary-to-primary matching (Figure B). Dependent filers (in Year 1) that matched to primary filers (in Year 2) comprised the largest source of these additional records (around 3 percent).

Despite these improvements, a number of records can still be excluded from the final migration data. A non-matching return can occur if a TIN is recorded incorrectly between the two years; if a taxpayer switches from a temporary TIN to a permanent Social Security Number (SSN); or if a taxpayer filed a return in one year, but did not file a return timely in another year[3]. Overall, the additional merging of secondary and dependent filers increases the precision of the migration data by including a wider segment of the filing population.

New tabulations – The Gross-Migration File

The last major improvement introduces a new tabulation that shows aggregate migration flows at the State level, by the size of adjusted gross income (AGI) and the age of the primary taxpayer. The Gross-Migration File is a summary of the migration flows for each State, plus the District of Columbia, that shows the total number of matched returns, non-migrant returns, outflow returns, inflow returns, and same State returns.

Comparisons of the old and new methodologies

In general, SOI made every effort to preserve the comparability and continuity of the migration data between the previous version and the new one. The new migration data will include the same set of files: State-to-State outflow and inflow and county-to-county outflow and inflow; as well as the same codes and definitions used in previous years. However, due to the methodological changes between the two, the data are not directly comparable.

[3] Individuals can apply to the IRS for an Individual Taxpayer Identification Number (ITIN) for the purpose of filing a valid U.S. Federal income tax return. An ITIN is a special nine-digit tax processing number, beginning with the number "9." There are some instances where an individual will receive a valid Social Security Number (SSN) in place of their ITIN and must file their individual return using the SSN. Excluded from the data are returns that switch between an ITIN and a SSN between two migration years because of the non-matching TINs.

Figure C:

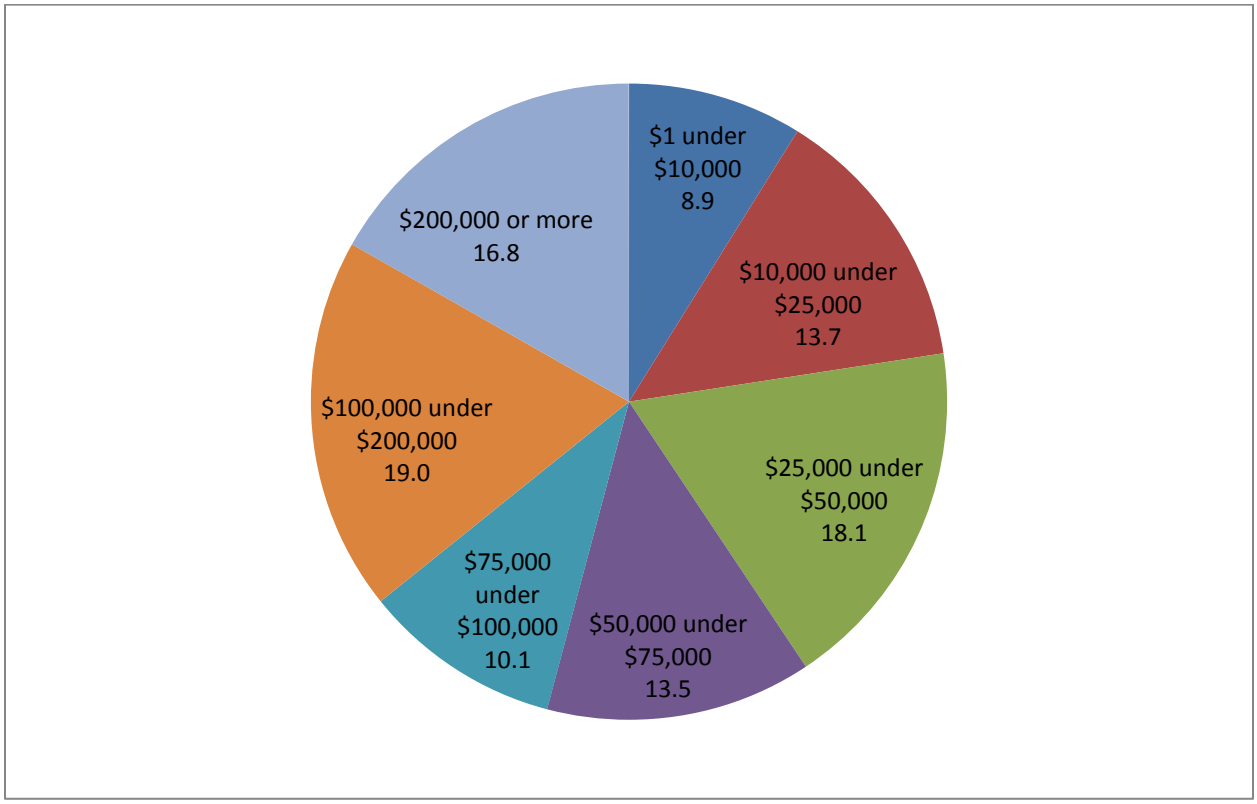
Total Number of Matched Returns, by Size of Adjusted Gross Income, Calendar Years 2011-2012

Size of adjusted gross income	Number of returns		
	Previous migration data	New migration data	Percentage change
Total	**111,505,981**	**116,764,589**	**4.7**
$1 under $10,000	10,380,065	10,846,031	4.5
$10,000 under $25,000	26,886,965	27,607,234	2.7
$25,000 under $50,000	29,804,233	30,755,044	3.2
$50,000 under $75,000	17,007,772	17,717,602	4.2
$75,000 under $100,000	10,723,397	11,253,887	4.9
$100,000 under $200,000	13,109,345	14,107,846	7.6
$200,000 or more	3,594,204	4,476,945	24.6

For the 2011-2012 migration data, SOI created a test file that mimicked the same procedures the Census Bureau employed to create the previous migration data.[4] This test migration file resulted in a total match of 111.5 million returns (Figure C). The new version of migration data had 116.8 million returns, an increase of 5.3 million returns or 4.7 percent of the previous version. Returns in the $200,000-or-more AGI category experienced the greatest increase in the new migration data, increasing about 25 percent, from 3.6 million returns to 4.5 million returns.

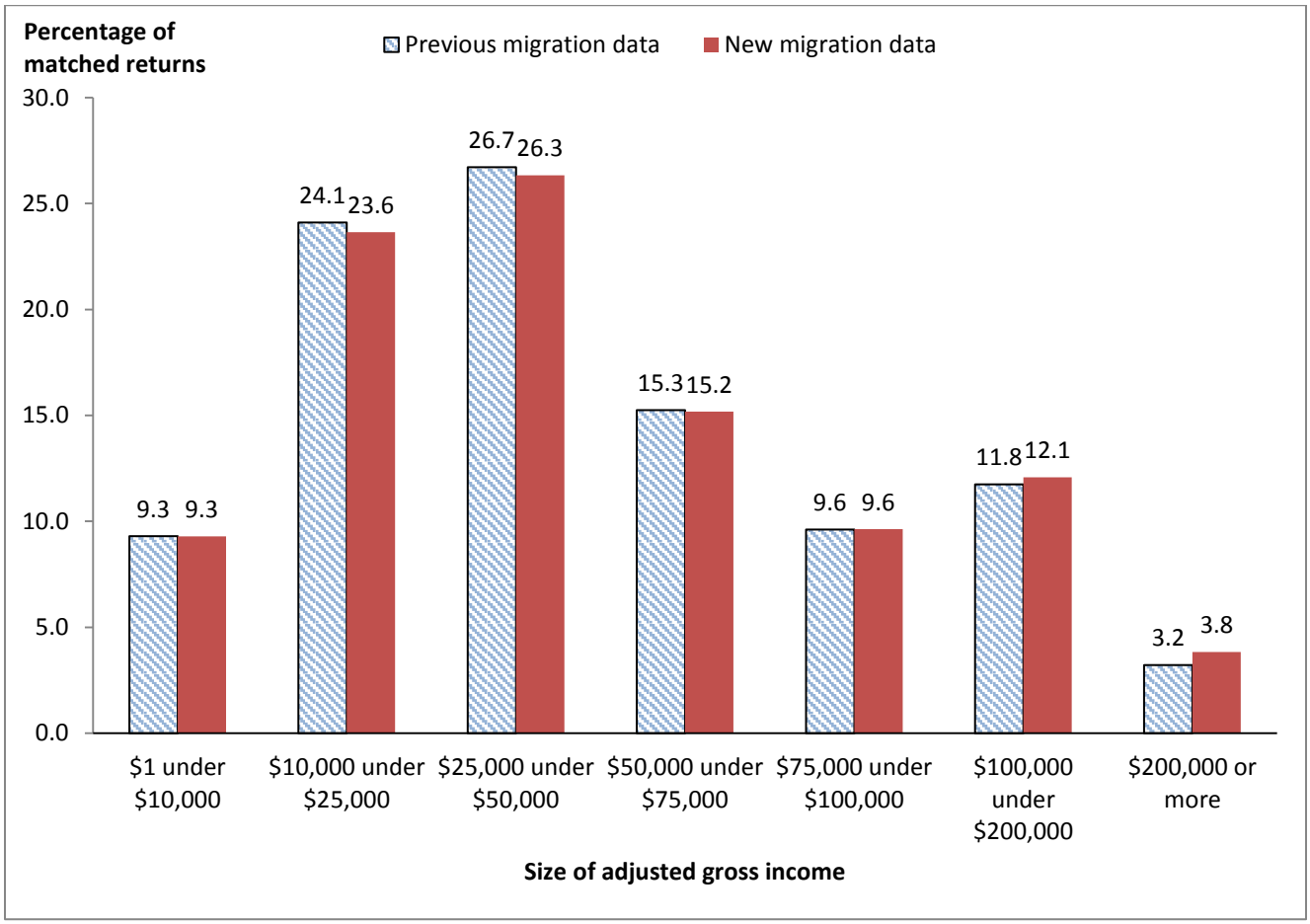
[4]SOI created a test file to compare the previous migration data with the new migration data. To avoid complementary disclosure, SOI is unable to release this test file to the public.

Figure D:
Percentage of Additional Matched Returns, by Size of Adjusted Gross Income, Calendar Years 2011-2012



Of the 5.3 million records added to the migration data, over a third (35.8 percent) came from the top two income categories (Figure D). Matched returns in the $100,000-under-$200,000 category comprised 19 percent of the additional records, while returns in the $200,000-or-more category made 16.8 percent. As discussed earlier, by including full-year data, the new migration data represent the movement or non-movement of high-income individuals and the income associated with the specific return to a greater degree.

Figure E: Distribution of Matched Returns, by Size of Adjusted Gross Income, Calendar Years 2011-2012



In spite of the additional records, the distribution of returns by the size of adjusted gross income is fairly consistent between the test data set and the new one (Figure E). The two largest AGI categories, in terms of percentage of matched returns, decreased slightly in the new migration data relative to the previous migration data. Matched returns in the $10,000-under-$25,000 category declined slightly from 26.7 percent to 26.3 percent. Returns in the $25,000-under-$50,000 category declined from 24.1 percent to 23.6 percent. In contrast, matched returns in the $100,000-under-$200,000 and $200,000-or-more categories experienced slight increases.

Net-Migration Rates

An alternate method of comparing the previous migration data with the new migration data is by examining the net-migration rate between the two data files. SOI calculates the net-migration rates by subtracting the number of out-migrant returns from the number of in-migrant returns and dividing this net amount by the sum of the non-migrant returns plus the out-migrant returns. Despite the differences between the old and new methodologies, a majority of states (86.2 percent) had a percentage difference of less than 5 percent (Figure F). A handful of states had modest differences between the old and new migration rates and one state had a percentage difference that was greater than 10 percent.

Figure F: Number of States, by Percentage Difference between the Previous and New Migration Data Net-Migration Rates, Calendar Years 2011-2012

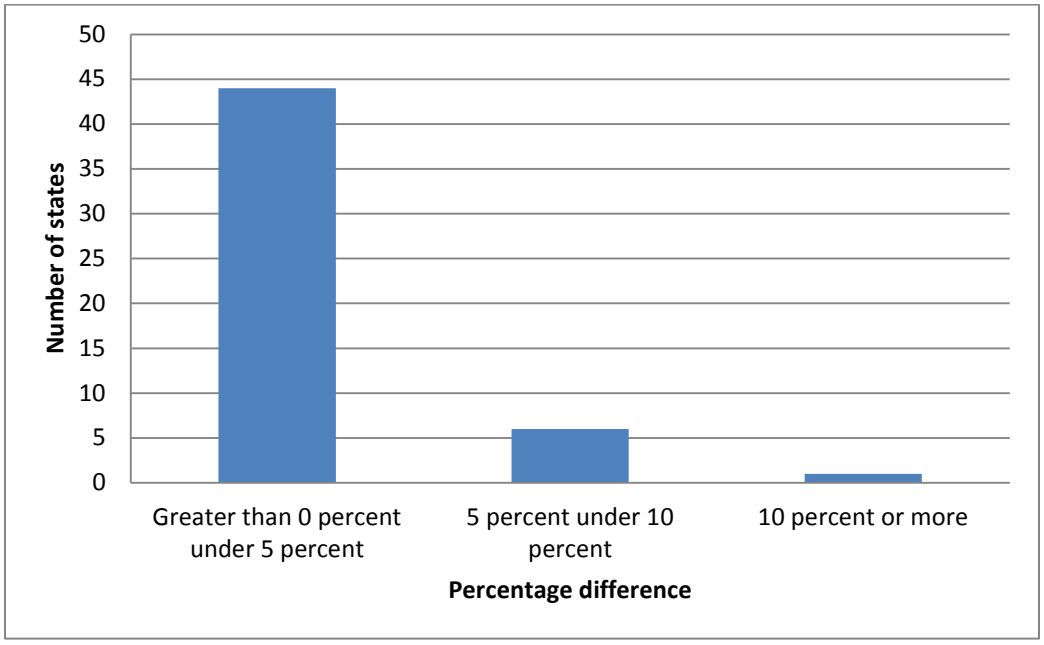
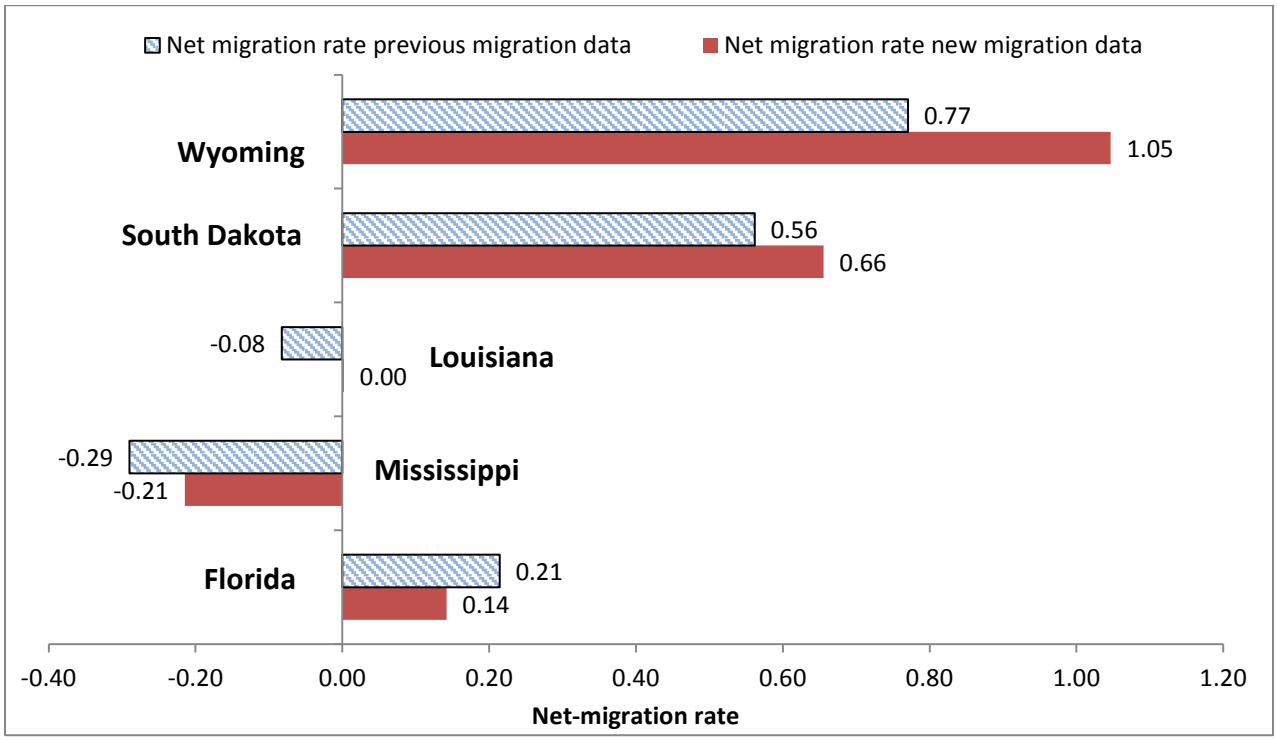


Figure G: Top 5 States with the Largest Net-Migration Rate Differential, Calendar Years 2011-2012



Wyoming had the largest percentage difference between the calculated net-migration rate of the previous migration data and the new version (Figure G). Wyoming's percentage difference of 27.6 percent was nearly three times as high as the second largest percentage difference of 9.4 percent for South Dakota.

For more information on SOI's migration data see: http://www.irs.gov/uac/SOI-Tax-Stats-Migration-Data.