UNIVERSITY OF MINNESOTA

Research Note on Linking CPS ASEC Files

José Daniel Pacas
University of Minnesota

Renae Rodgers
University of Minnesota

September 2021

Working Paper No. 2021-05

DOI: https://doi.org/10.18128/MPC2021-05

Address correspondence to José Daniel Pacas, Institute for Social Research and Data Innovation, University of Minnesota, 50 Willey Hall, 225 19th Ave S., Minneapolis, MN 55455 (email: pacas002@umn.edu or Renae Rodgers email rodge103@umn.edu). Acknowledgments: The authors would like to acknowledge the helpful comments of Joe Ritter and Elizabeth Wrigley-Field. We also thank the IPUMS CPS team and Sarah Flood. Funding was provided by the Minnesota Population Center through a grant (P2CHD041023) from the Eunice Kennedy Shriver National Institute for Child Health and Human Development NICHD.

Research Note on Linking CPS ASEC Files

Abstract

Measuring change over time in areas such as family structure, employment, income, and poverty is of great interest to social scientists. The panel component of the Current Population Survey (CPS) affords the opportunity to observe short-term change in these areas. The Annual Social and Economic supplement (ASEC), with its wealth of information on income, health insurance coverage, benefits receipt, and many other topics, is a particularly popular resource for this purpose. However, commonly used methods for linking CPS ASEC files do not address how to link the ASEC oversample records across years, leading to smaller linked sample sizes and unreliable estimates for certain subpopulations. In this research note, we describe how to link individuals from the ASEC oversamples in the 2005-2020 data and the increase in sample size researchers can achieve by doing so. Finally, we illustrate the importance of including oversample records by examining poverty transitions over time.

Keywords: Measurement, Poverty, Current Population Survey

Introduction

The Current Population Survey (CPS) and its many supplements, particularly the Annual Social and Economic Supplement (ASEC), are vitally important for demographic research in the US. The rotation pattern of the CPS, in which a household is interviewed four consecutive months a year and then again those same months the following year, has allowed researchers to construct two-year panels and exploit the longitudinal aspect of the survey. Indeed, while the ability to link adjacent ASEC files has been documented and used for years (Katz et al. 1984; Pitts 1988; Madrian & Lefgren 2000; Feng 2001), this feature of the data set has been growing in popularity because of recent innovations in linking (Drew et al. 2014; Flood & Pacas 2017; Flood et al. 2020). Taking advantage of the CPS rotation pattern has allowed researchers to use the ASEC as a two-year panel and study a wide variety of issues such as income volatility, the impact of education on employment, changes in the social safety net, and poverty transitions (Ziliak et al. 2011; Riddell & Song 2011; Hardy et al. 2018; Pacas 2017; Pacas & Davis 2019). Given the importance of the ASEC as a source of official statistics and its wide topical coverage, there is untapped potential for research using two-year ASEC panels. In this note, we call attention to a feature of the ASEC that has been largely overlooked when using this data set as a panel: two distinct oversamples, the Hispanic oversample and the State Children's Health Insurance Program (SCHIP)-eligible oversample, that were introduced by the Census Bureau to improve the reliability of subpopulation estimates.

Currently-popular CPS ASEC linking methodologies do not properly account for an idiosyncrasy of these records, resulting in the exclusion of the vast majority of the ASEC oversample respondents from linked files. We show that the oversamples are linkable and, more importantly, that their inclusion has important ramifications for small subpopulation estimates. Specifically, we demonstrate how to recover the linkable oversample cases in the 2005-2020 ASEC,

resulting in about 150,000 more linked records (between 13,000 and 19,000 yearly) which represents a 30% increase in the overall linked sample size. In addition to yielding larger linked sample sizes, including ASEC oversample records when linking results in more accurate denominators.

To illustrate the importance of including the ASEC oversamples in the linked data, we compare rates of poverty transitions (people moving in and out of poverty) in the U.S. population as a whole, and among Hispanics in particular, with and without the oversamples. We find that poverty transitions in the U.S. population are consistently underestimated when the oversample records are not properly included in the ASEC panels. Among Hispanics, excluding ASEC oversample records results in erratic estimates of poverty transitions.

Background

Linking Methodologies

Linking the CPS relies on an understanding of the CPS rotation pattern. The CPS is a monthly survey of roughly 73,000 households. These households are interviewed over a period of 16 months in a 4-8-4 rotation pattern; a household is interviewed for four consecutive months, rotates out of the sample for 8 months, and is then interviewed again for four consecutive months. A household's place in the CPS rotation pattern is indicated by a "month-in-sample" (MIS) variable - which takes on values one through eight. Given the CPS rotating panel structure, it is commonly understood that a household is interviewed, at most, eight times over a period of two years. This section highlights how ASEC oversamples are excluded from linking methodologies because of the validation criteria that a household follows a sequential pattern of MIS values.

Census Bureau documentation for linking CPS files is sparse. CPS panels have been created by independent researchers and various efforts have been made to lower the barrier to creating panels from the CPS. These methodologies include linking the ASEC in two ways: 1) as part of full CPS panels that link all basic monthly samples and 2) linking adjacent ASECs directly. The differences between the two methodologies are subtle but the similarity between them highlights how both exclude all ASEC oversample respondents.

One set of linking methodologies aims to create full panels of all CPS responses (see Nekarda 2009, Rivera Drew, Flood, and Warren 2014). The approach is to link as many BMS responses for an individual as possible. As demonstrated in Flood & Pacas (2017), the ASEC can be used as a part of these panels. Specifically, as long as a respondent's set of BMS responses include the month of March, then it is possible to link their ASEC responses to the March BMS (see Figure 1 for a visual representation of this linking setup). Importantly, because the central purpose of these methodologies is to create full panels from the BMS responses, the ASEC oversamples are necessarily excluded from the March BMS to ASEC link.[1] In sum, linking methodologies that focus on linking BMS responses allow for linking ASECs across years but only for March BMS responses and therefore exclude ASEC oversample responses by design.

The second set of methodologies explicitly try to link adjacent ASEC samples. The most popular methodology is Madrian and Lefgren (2000), which has been used for many different analyses (see Ziliak et al. 2011; Riddell & Song 2011; Liu & Trefler 2011; Humensky et al. 2013, Hokayem & Heggeness 2014; Elsby et al. 2016; Hardy et al. 2018).[2] The methodology used by Madrian and Lefgren (2000) relies on the explicit instructions outlined in Census Bureau documentation

[1] Although the oversample respondents have non-March BMS responses, the Census Bureau does not provide identifiers to link the oversample responses to their BMS responses.
[2] Feng 2001 also creates ASEC panels using a Bayesian approach for validating links. However, the methodology cannot be extended after 2002 and so is not covered here.

which provides the linking keys needed for linking ASECs. The key instruction given by the Census Bureau that leads to the exclusion of all ASEC oversample respondents is the following:

> The first step in matching year t with year t+1 is to select from year-t those housing units with a ‘month in sample’ value of 1 through 4, and from year t+1 those units with a ‘month in sample’ value of 5 through 8. This will identify the sample subset eligible for matching. Within this subset, housing units in year t, month 1 will match only with units in year t+1, month 5, etc. (U.S. Census Bureau 2020)

These instructions implicitly ignore the MIS assignment for ASEC oversample respondents and methodologies based on these instructions will not link these respondents.

The ASEC and Its Oversamples

The ASEC is conducted primarily in March of each year. In addition to the demographic and labor force information contained in the basic monthly survey (BMS), the ASEC includes information on health insurance coverage, income, noncash benefits, and poverty. Because of its importance in government statistics and for socio-demographic research, the ASEC sample has been increased over the years. Today, the ASEC file contains all of the March BMS respondents plus two oversamples. Figure 1 outlines the specific sampling methodology used for selecting the ASEC oversample and their potential links across years. The Hispanic oversample was first introduced in 1976 to improve the reliability of estimates for this subpopulation and contains roughly 2,500 households (about 5,000 individuals) each year (Flood & Pacas 2017). These households are selected from those households who received the November BMS and had at least one Hispanic person in the household. Importantly, these households would have not been part of the March BMS. As a result, the Census Bureau will field the ASEC in February or March as a completely separate interview and therefore results in Hispanic households being interviewed a ninth and potentially tenth time.

The second oversample, known as the SCHIP oversample, was first introduced in 2001 to improve estimates of children without health insurance and contains about 12,000 households (about 24,000 individuals) (Flood & Pacas 2017). The SCHIP oversample households include non-Hispanic non-Whites or non-Hispanic Whites with children 18 years old or younger. These households are chosen from various different BMS and receive a separate ninth or tenth interview in February, March, or April.

Importantly, a subset of the oversample records can also be linked across years. Indeed, half of the Hispanic oversample respondents and 2/7 of the SCHIP oversample respondents (i.e. 2 MIS groups link forward out of a total of 7 MIS groups selected for the SCHIP oversample) from 2005 onward (Flood & Pacas 2017) can be linked across years. Many ASEC linking methodologies overlook that the ASEC oversample respondents consist of households that have been interviewed a ninth and tenth time by the CPS and have been assigned MIS values between one and eight. In other words, even though oversample households are being interviewed for a ninth or tenth time, MIS values of 9 or 10 are never found in the CPS public-use data. In processing the ASEC, the Census Bureau assigns all oversample households a month-in-sample value between one and eight (Flood & Pacas 2017) such that MIS assignments are evenly distributed across all 8 MIS groups. The oversample MIS values are irrelevant for linking purposes.

Methodology

For data from 2002 and onward, the Census Bureau instructs users to use household and person identifiers only to link ASEC files across years (U.S. Census Bureau). Beginning in 2005, the only linking key required to link ASEC files across years is the CPS variable PERIDNUM (U.S. Census Bureau). Using PERIDNUM as the sole linking key, we link all ASECs from 2005

through 2020.[3,4] Importantly, the key insight of our methodology is that MIS is not used in any capacity for linking ASEC oversamples. This is contrary to what the Census Bureau suggests in their documentation but our research shows that these MIS criteria do not apply to ASEC oversamples. After achieving naive links using PERIDNUM, we validate those links based on age, sex, and race. The validation criteria are adopted from Madrian and Lefgren (2000). Validation on sex and race require that responses be identical across years. Age values in the first year must be between -1 and 3 years in the following year.

Our methodology of using PERIDNUM and, more importantly, relaxing the MIS validation criteria, recovers thousands of validated responses each year and clearly demonstrates the problems of using MIS to link ASECs. Table 1 shows month-in-sample values for March BMS records from the 2005 and 2006 ASEC files linked using identifiers only. The rotation pattern is clearly visible. The MIS values for households in 2005 range from 1 to 4 and link to 2006 MIS values that range from 5 to 8. As required by the rotation pattern, the MIS values in 2006 should be exactly the 2005 MIS value plus 4. Table 1 clearly shows how this holds true for all March BMS records.

Unlike for the March BMS links, the month-in-sample values do not reflect the rotation pattern for oversample links. Table 2 shows MIS values for linked Hispanic and SCHIP oversample records between the 2005 and 2006 ASEC files; note that most successful links do not follow the expected MIS progression. Given that these records match on PERIDNUM and demographic characteristics, Table 2 confirms both that the assignment of MIS to ASEC

[3] As properly linking all components of the ASEC files prior to 2005 is much more complicated than in 2005 and after, we restrict ourselves to the later period for the purposes of demonstrating the importance of taking the ASEC oversamples into account when linking.

[4] MIS values in the original 2016, 2018, and 2020 ASEC files must be addressed to properly compare the outcomes of linking using only the Census Bureau-recommended linking keys and a method that incorporates month-in-sample. In 2016, 2018, and 2020, the original month-in-sample values for the March BMS respondents do not follow the expected rotation pattern. In these years, March BMS households actually in months-in-sample one through four have month-in-sample values of five through eight and households actually in months-in-sample five through eight have month-in-sample values of one through four.

oversample respondents does not maintain the expected rotation pattern and that, in spite of this, ASEC oversample records can be linked across years. As our subsequent analysis shows, our methodology for linking ASEC oversamples has important ramifications for analysis.

Illustration

By ignoring MIS when linking ASEC oversamples between 2005 and 2020, we recover between 13,000 and 19,000 linked oversample records for each pair of linked files between 2005 and 2020. Figure 1 shows unweighted counts of links across ASEC files broken down by oversample type. For comparison, we implement the Madrian and Lefgren methodology for the same time period. As Table 2 shows, there are some cases where the MIS pattern does hold and therefore does link and validate across years. However, the increase in sample size using our methodology is substantial.

Table 3 demonstrates that the ASEC oversample consists of respondents who are demographically different than those in the March BMS. One obvious implication is increased precision of estimates based on the linked sample. The increased precision is most pronounced for estimates of subpopulations targeted by the oversampling, but since these subpopulations are overrepresented among low-income households, the impact on overall precision is greater than it would be if the core ASEC sample were increased by the same number of households.

To illustrate the importance of the ASEC oversamples in longitudinal analysis, we look at poverty transitions. Pacas (2017) decomposed the churn of the official poverty rate across each year and found that there is about the same proportion of people exiting poverty each year and staying in poverty each year. This finding implies that about half of the 34 million people in poverty in 2019 were not in poverty in 2018 while the other half were poor in both years. Importantly, the analysis conducted in Pacas 2017 does not include the ASEC oversample.

Figure 2 demonstrates that the increase in sample size allows us to capture hundreds more transitions into poverty across years in the data, compared to the Madrian and Lefgren methodology.

As Figure 3 demonstrates, failing to account for the oversamples consistently underestimates poverty transitions in the U.S. population as a whole. We graph the unweighted proportion of ASEC respondents that fall into and out of poverty from year to the next and contrast our methodology to Madrian and Lefgren's. Recovering the ASEC oversample using our methodology leads to higher poverty transition rates. This result is expected given the higher poverty rates found in the ASEC oversample as demonstrated in Table 3.

Correctly linking the ASEC oversamples is also important for studying transitions among subpopulations, particularly Hispanics. Figure 4 shows that excluding the ASEC oversamples when linking leads to erratic estimates of poverty transitions among Hispanics. The fact that estimates of the percentage of Hispanic individuals experiencing poverty transitions fluctuates noticeably between the Madrian and Lefgren and PERIDNUM linking techniques highlights another pitfall of not properly linking oversamples: the denominator. Not including all linkable Hispanic oversample records in the linked data set has an impact not only on the number of Hispanics experiencing a poverty transition, but on the total number of Hispanics included in the data set. Therefore, the inclusion of ASEC oversample respondents is of vital importance in understanding the trends of poverty transitions for subpopulations, particularly of Hispanics.

Conclusion

Properly including the Hispanic and SCHIP oversample records when linking ASEC files across years results in larger linked sample sizes and more accurate population estimates. In this note, we demonstrate a methodology for properly including the ASEC oversample and

highlight how prior methodologies have overlooked this component of the ASEC. We bring attention to the fact that the ASEC oversample respondents are demographically different than the March BMS respondents. Importantly, we do not directly address the issue of weighting with linked CPS ASEC samples. Our paper is narrowly focused on the increased sample size rather than how to make the particular sample representative of the population. Certainly the choice of weights is extremely important and can be done by various techniques (i.e. raking, inverse probability weights) but ultimately the weighting decisions depends on a variety of factors: attrition across waves, validation criteria chosen by researchers that leads to fewer/more sample being included, the subsamples of interests and the particular analysis question. Our methodology allows researchers to fully utilize the ASEC and provides further documentation on an important aspect of using CPS panels.

References

Elsby, M. W. L., Shin, D., & Solon, G. (2016). Wage Adjustment in the Great Recession and Other Downturns: Evidence from the United States and Great Britain. *Journal of Labor Economics*, *34*(S1), S249–S291. https://doi.org/10.1086/682407

Feng, S. (2001). The longitudinal matching of current population surveys: A proposed algorithm. *Journal of Economic and Social Measurement*, *27*(1–2), 71–91. https://doi.org/10.3233/JEM-2003-0197

Sarah Flood, Miriam King, Renae Rodgers, Steven Ruggles and J. Robert Warren. Integrated Public Use Microdata Series, Current Population Survey: Version 8.0 [dataset]. Minneapolis, MN: IPUMS, 2020. https://doi.org/10.18128/D030.V8.0

Flood, S. M., & Pacas, J. D. (2017). Using the Annual Social and Economic Supplement as part of a Current Population Survey panel. *Journal of Economic and Social Measurement*, *42*(3–4), 225–248. https://doi.org/10.3233/JEM-180447

Flood, S. M., Rodgers, R., Pacas, J. D., Kristiansen, D., & Klass, B. (2021). *Extending Current Population Survey Linkages: Obstacles and Solutions for Linking Monthly Data from 1976 to 1988, Version 2*. (Working Paper No. 2021-02). Minneapolis:MN. Institute for Social Research and Data Innovation. https://doi.org/10.18128/IPUMS2020-02.v2

Hardy, B., Smeeding, T., & Ziliak, J. P. (2018). The Changing Safety Net for Low-Income Parents and Their Children: Structural or Cyclical Changes in Income Support Policy? *Demography*, *55*(1), 189–221. https://doi.org/10.1007/s13524-017-0642-7

Hokayem, C., & Heggeness, M. (2014). *Factors Influencing Transitions Into and Out of Near Poverty: 2004-2012.* (SEHSD Working Paper 2014-05).

Humensky, J. L., Jordan, N., Stroupe, K. T., & Hynes, D. M. (2013). How Are Iraq/Afghanistan-Era Veterans Faring in the Labor Market? *Armed Forces & Society*, *39*(1), 158–183. https://doi.org/10.1177/0095327X12449433

Katz, A., Teuter, K., & Sidel, P. (1984). Comparison of Alternative Ways of Deriving Panel Data from the Annual Demographic Files of the Current Population Survey. *Review of Public Data Use*, *12*, 35–44.

Liu, R., & Trefler, D. (2011). *A Sorted Tale of Globalization: White Collar Jobs and the Rise of Service Offshoring*. National Bureau of Economic Research. https://www.nber.org/system/files/working_papers/w17559/w17559.pdf

Madrian, B. C., & Lefgren, L. J. (2000). An approach to longitudinally matching Current Population Survey (CPS) respondents. *Journal of Economic and Social Measurement*, *26*(1), 31–62. https://doi.org/10.3233/JEM-2000-0165

Nekarda, C. J. (2009). A longitudinal analysis of the current population survey: Assessing the cyclical bias of geographic mobility. Federal Reserve Board of Governors.

Pacas Viscarra, J. (2017). Innovative Methods for Using Census Data to Study Poverty, Labor Markets, and Policy. Retrieved from the University of Minnesota Digital Conservancy, https://hdl.handle.net/11299/191475.

Pacas, J. & Davis, E. (2018). Moving into and out of rural poverty. *IRP Focus* 34(3): 4-12.

Pitts, A. (1988). *Matching Adjacent Years of the Current Population Survey*. Unicon Corporation.

Riddell, W. C., & Song, X. (2011). The impact of education on unemployment incidence and re-employment success: Evidence from the U.S. labour market. *Labour Economics*, *18*(4), 453–463. https://doi.org/10.1016/j.labeco.2011.01.003

Rivera Drew, J. A., Flood, S., & Warren, J. R. (2014). Making full use of the longitudinal design of the Current Population Survey: Methods for linking records across 16 months. *Journal of Economic and Social Measurement*, *39*(3), 121–144. https://doi.org/10.3233/JEM-140388

U.S. Bureau of Labor Statistics. (2006). Design and Methodology: Current Population Survey. Technical Paper 66. Washington, D.C.: U.S. Department of Labor, Bureau of the Census.

U.S. Census Bureau. (2020). Current Population Survey, 2020 Annual Social and Economic (ASEC) Supplement.

U.S. Census Bureau. (n.d.). *How To Link CPS Public Use Files*.

Ziliak, J. P., Hardy, B., & Bollinger, C. (2011). Earnings volatility in America: Evidence from matched CPS. *Labour Economics*, *18*(6), 742–754. https://doi.org/10.1016/j.labeco.2011.06.015

Table 1. Month-in-sample for linked March BMS records in the ASEC, 2005-2006

		2006 MIS values			
		5	6	7	8
2005 MIS values	1	12,476	0	0	0
	2	0	12,667	0	0
	3	0	0	12,903	0
	4	0	0	0	12,858

Table 2. Month-in-sample values for linked oversample records in the ASEC file, 2005-2006

		2006 MIS values							
		1	2	3	4	5	6	7	8
2005 MIS values	1	275	234	161	190	146	199	218	245
	2	255	215	224	220	206	143	210	212
	3	243	243	231	158	230	185	173	245
	4	161	206	287	207	233	162	179	137
	5	1,526	193	221	180	145	194	263	208
	6	193	1,241	172	222	197	184	136	254
	7	210	196	1,208	254	270	167	255	144
	8	199	207	231	1,245	183	244	240	191

Note: These links are validated on age, sex, and race such that only those whose sex and race are identical between years and whose age in 2006 is between -1 and 3 years larger than in 2005. 96.5% of oversample links are retained after validation.

Table 3. Comparison of Demographic Characteristics of Linked March BMS and ASEC oversample records, 2005-2020

	March BMS	ASEC oversamples	Difference*
Sex (Percents)			
Female	51.9	51.4	-0.5
Male	48.1	48.6	0.5
Age (Percents)			
Less than 18	22.5	35.2	12.8
18-64	60.8	58.3	-2.5
65+	16.8	6.5	-10.3
Race (Percents)			
White Non-Hispanic	71.5	45.4	-26.1
Black Non-Hispanic	8.9	10.8	1.9
Other single race Non-Hispanic	5.5	7.1	1.6
Other mutiple race Non-Hispanic	1.8	2.3	0.5
Hispanic (any race)	12.2	34.4	22.1
Percent In Poverty	10.4	13.4	3.1

*Differences are statistically different from zero at the 95 percent confidence level.

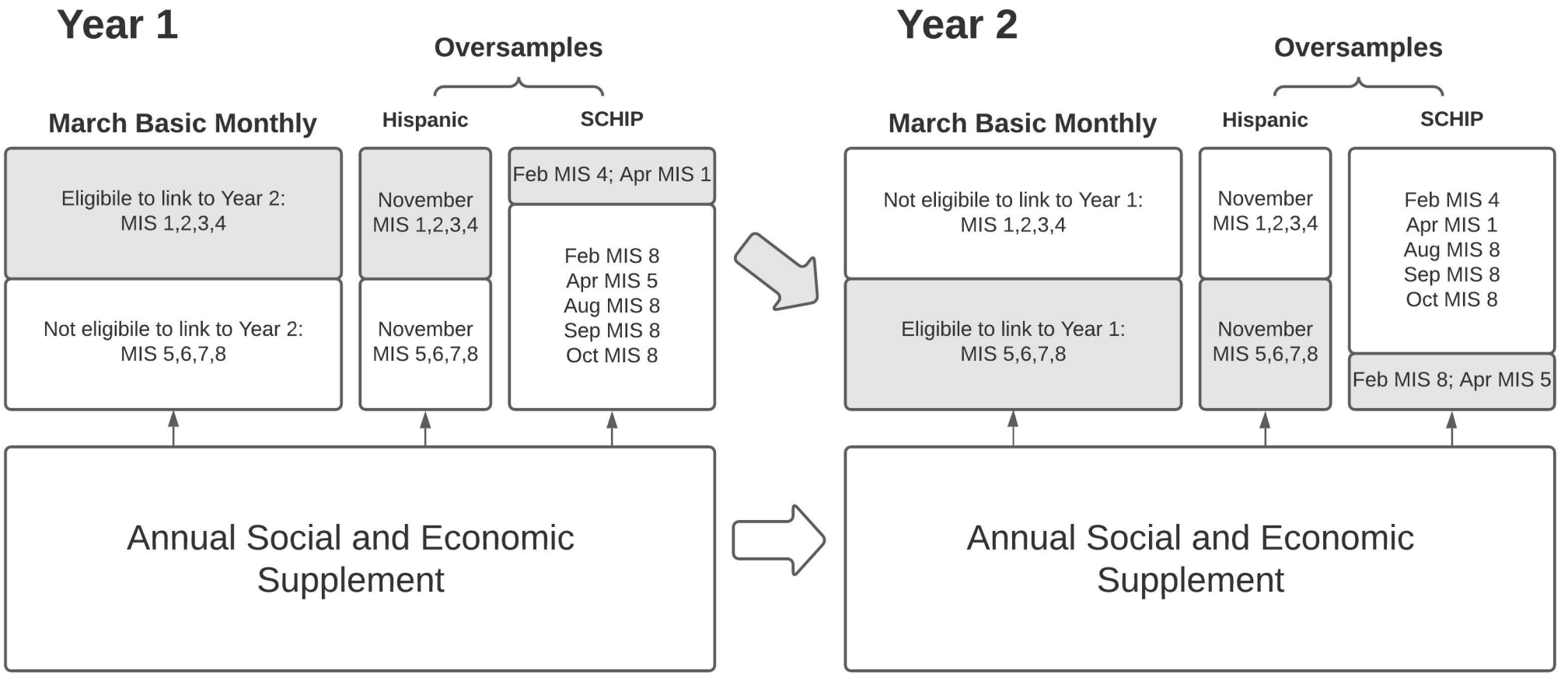

Figure 1. A Visual Description of ASEC Oversamples and Linking

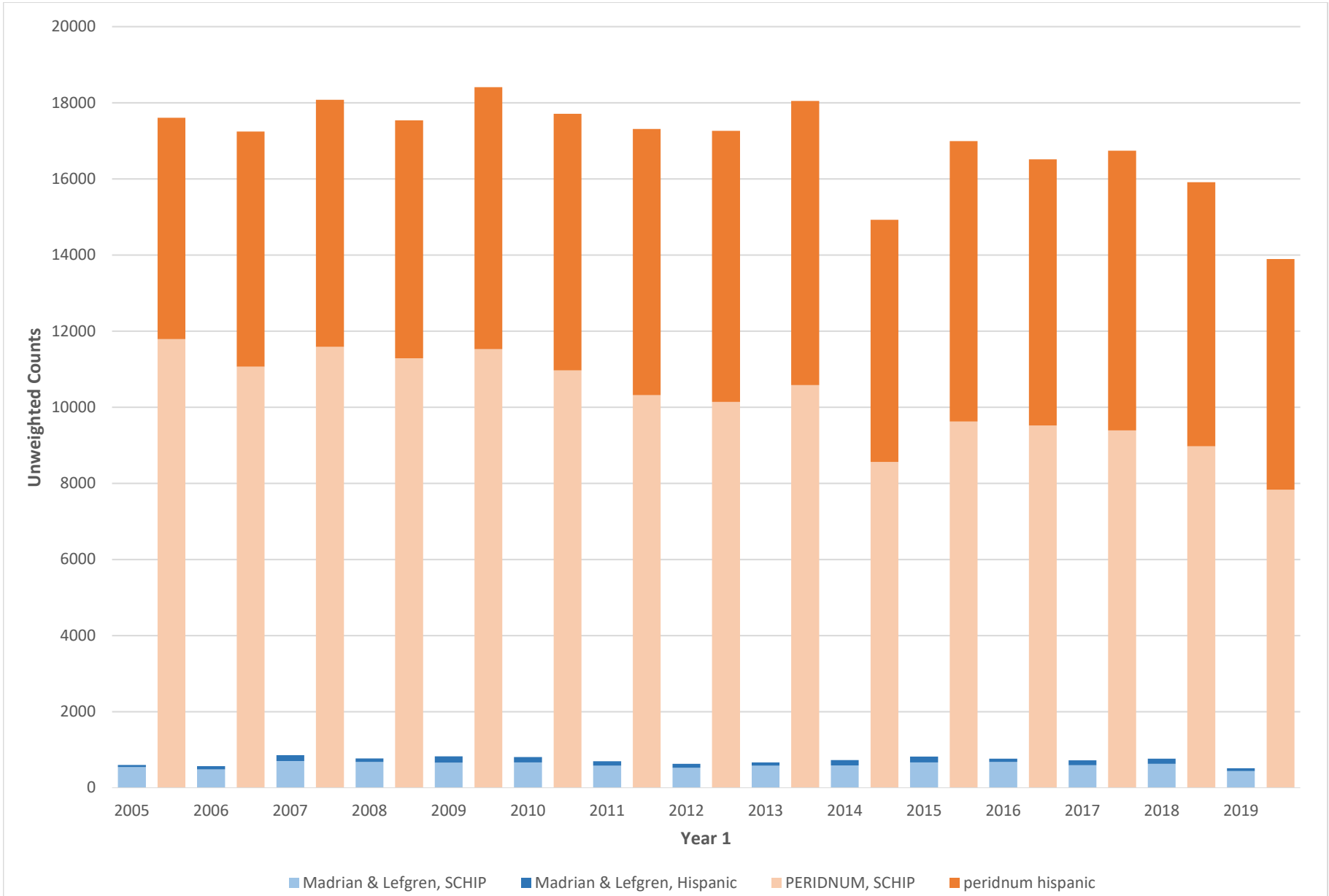

Figure 2. Validated Oversample Links by Type, Madrian & Lefgren vs PERIDNUM, 2005-2020

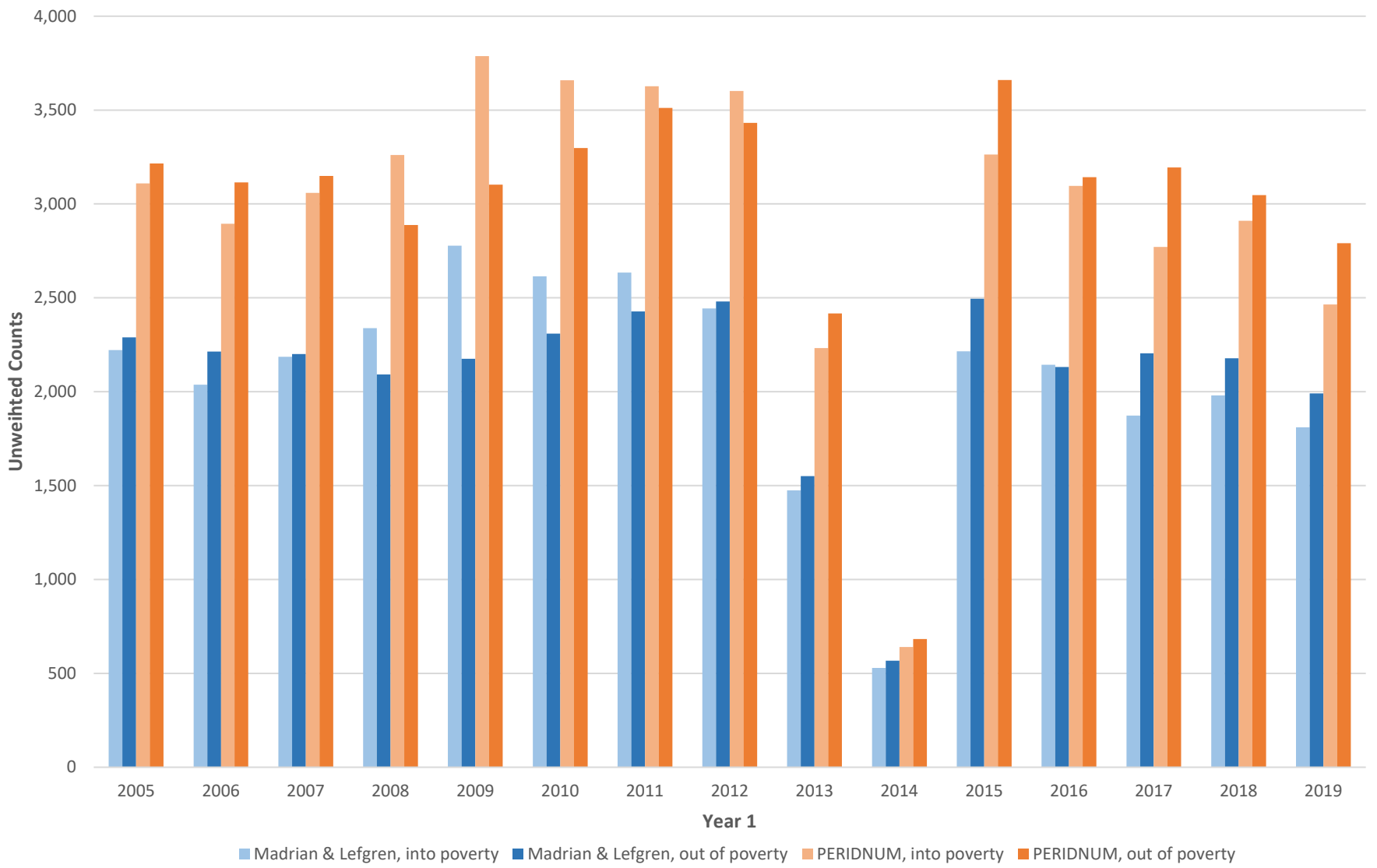

Figure 3. ASEC-ASEC Poverty Transitions, 2005-2020

Note: In 2014 the ASEC income questions were re-designed. To help assess the effect of the new questionnaire, the 2014 ASEC sample into two groups. 3/8 of the sample were given the new income questionnaire and 5/8 of the sample was given the original income questions. Because income figures into poverty calculations, we link 2013 to the 5/8 2014 sample and the 3/8 2014 sample to 2015.

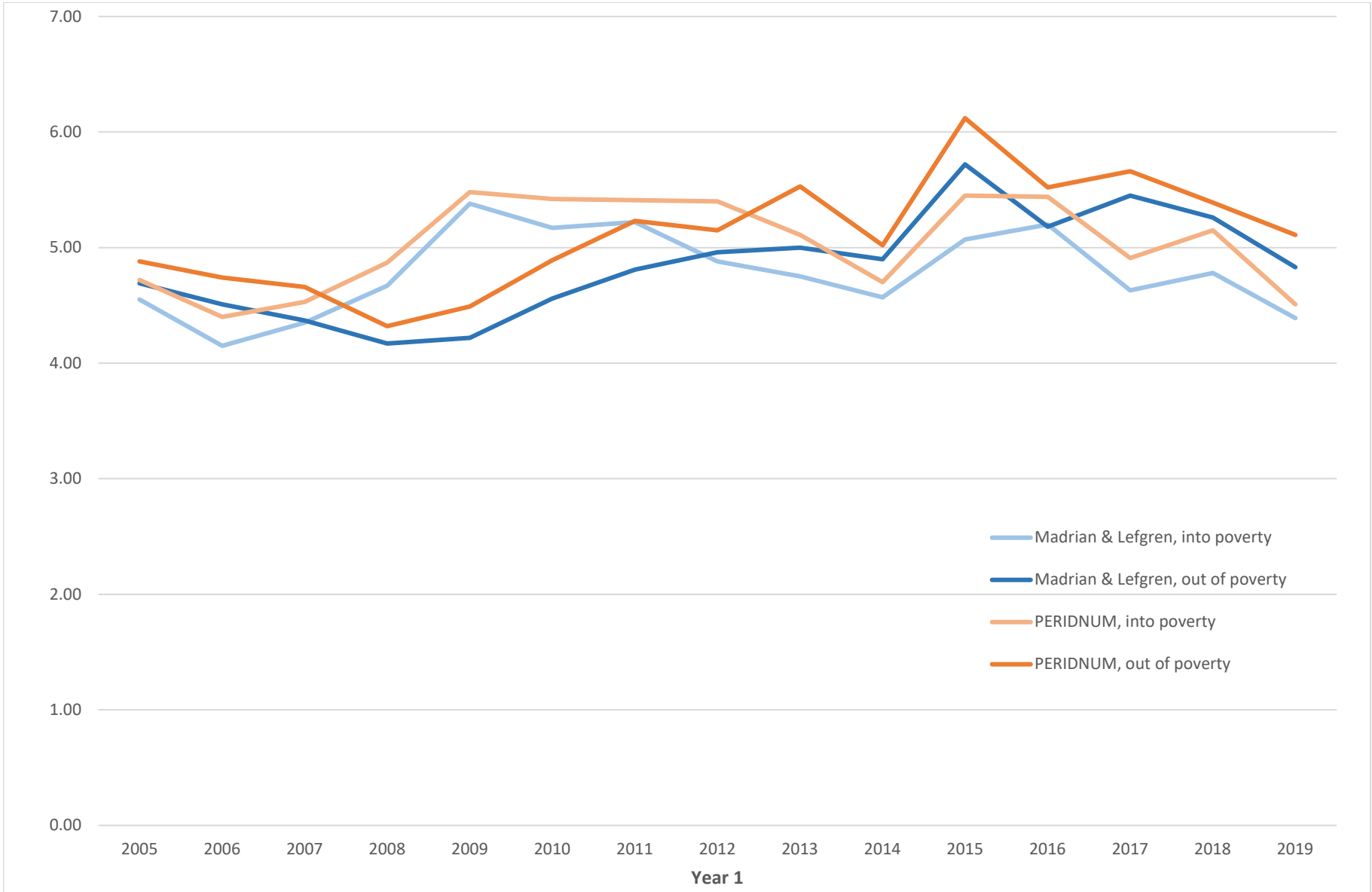

Figure 4. Percentage of Linked Individuals Experiencing Poverty Transitions, 2005-2020

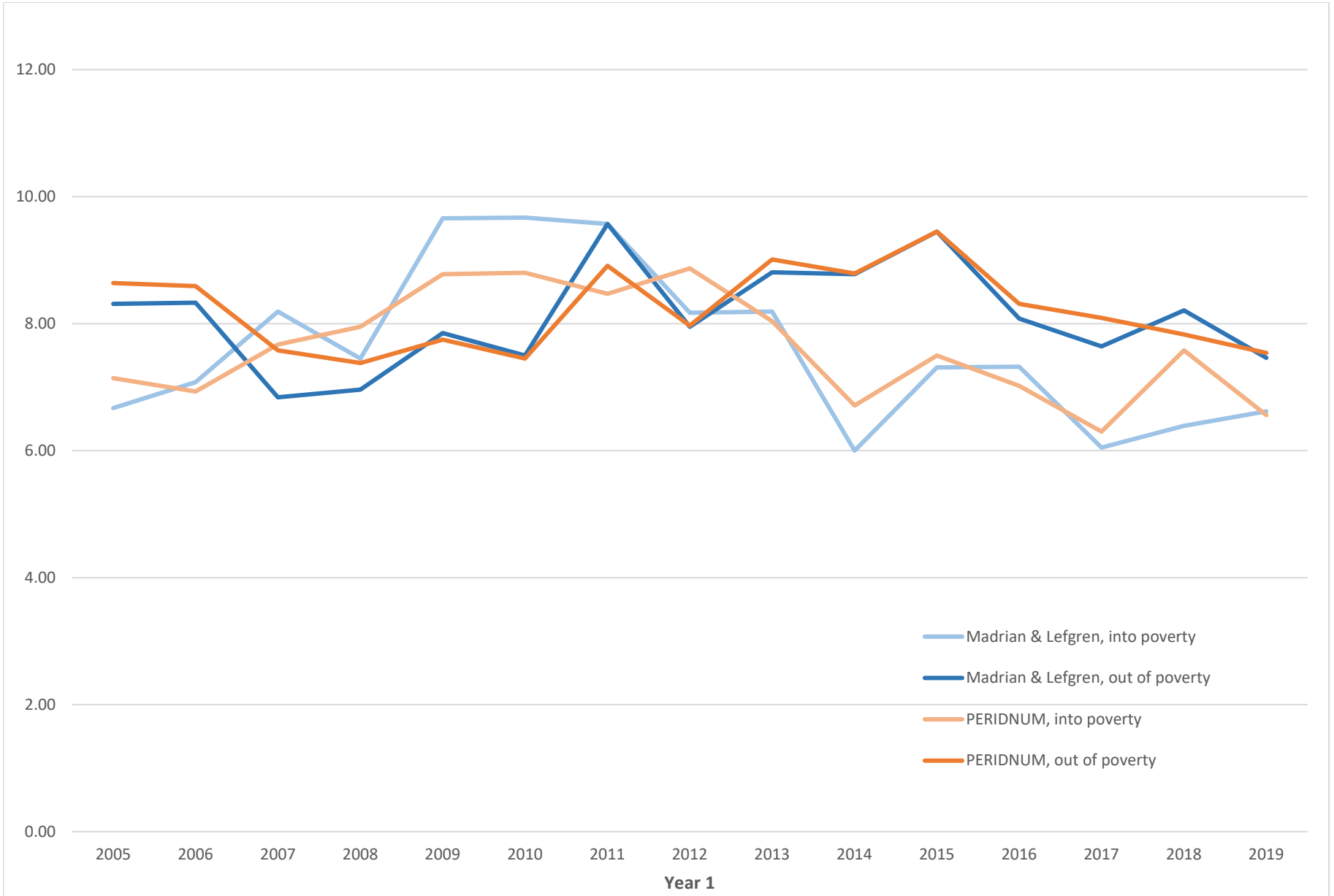

Figure 5. Percentage of Linked Hispanic Individuals Experiencing Poverty Transitions, 2005-2020