F

Nonresponse Bias Analysis of the 2018-2020 Consumer Expenditure Interview Survey using Census Administrative Records

March 15, 2024

Stephen Ash and Jonathan Rothbaum

Consumer Expenditure Surveys Program Report Series

Contents

1. Introduction

The objective of this project is to identify the impact of the COVID-19 pandemic and resulting data collection changes on Interview Survey estimates. Particular attention was given to expenditures estimates at the start of the COVID-19 pandemic in 2020. As with all aspects of our lives, the COVID-19 pandemic impacted the Consumer Expenditure Surveys (CE) and particularly the response rates for CE. For example, the COVID-19 pandemic caused changes to survey collection methods. During the second and third quarters of 2020 (April through September) most of the Interview survey's data were collected by telephone interviewing. The "maximum telephone interviewing" protocol began on March 19, 2020, and was gradually phased out over the summer and fall of 2020. (Knappenberger et al. 2021; Armstrong et al. 2022).[1]

Nonresponse bias results from the omission of sampled units in the final measured sample, when key variable values associated with missed units differ from that of measured units. Nonresponse adjustments attempt to correct for that potential error, but do not cause it. No survey can completely account for nonresponse with weighting and therefore all surveys include a certain degree of nonresponse bias. Problems arise when the weighting does not adequately mitigate nonresponse and it results in a considerable bias. The COVID-19 pandemic induced change in data collection procedures may have affected how well CE addresses nonresponse bias because the current weighting methods were developed and tailored for the regular survey operations before the COVID-19 pandemic. It may be the case that the current methods do not adequately mitigate nonresponse bias in 2020.

Literature Review. This paper is a continuation of the Bureau of Labor Statistics' (BLS) mission to improve the accounting for non-interviews. Sabelhaus et al. (2013) examined response rates by combining the CE sample with publicly available Internal Revenue Service (IRS) Income at the ZIP code level. They concluded that high-income households were under-represented and low-income households were over-represented in the CE Interview Survey and suggested that CE's non-interview adjustment should account for the differential response propensities by income. Following Sabelhaus et al. (2013), Dumbacher et al. (2012) conducted research on the variables used in the non-interview adjustment. Their research suggested keeping some old variables and adding new variables including an income variable based on publicly available IRS income data at the ZIP code level that was defined as in Sabelhaus et al. (2013). Krieger et al. (2019) also compared the set of nonresponse variables used before and after 2014 and found that the newer variables (which included ZIP code income) made some improvement in stratifying the sample by response propensity, but they did not have a significant impact on expenditure estimates.

Brummet et al. (2018) conducted an initial investigation of the use of Census administrative records in conjunction with CE and concluded that "[i]ncorporating the linkage of administrative records into these production processes has the potential to improve the accuracy and quality of statistics produced from

[1] The telephone interviewing rate has still not returned to its pre-pandemic level.

the CE." Steinberg et al. (2020) and Steinberg, Ash, Voorheis (2022) examined replacing the publicly available ZIP code income used in the non-interview adjustment with IRS income from the Census Administrative records and found that it did not produce a substantial change in the expenditure estimates.

More recently, Ash, Nix, and Steinberg (2022a) and Ash, Nix, and Steinberg (2022b) examined nonresponse bias during the COVID-19 pandemic with the Interview and Diary Surveys, respectively. They compared alternative non-interview adjustments that used variables from the Census Planning Database (for example, quartiles of percent of population within a Census tract aged 65 years or older and quartiles of percent of population within Census tract below the poverty level) and found that they did not produce a substantial change in the expenditure estimates. [2]

Prior comprehensive studies of nonresponse bias for the CE surveys include Chopova et al. (2009) and Steinberg et al. (2022). They both found that that nonresponse bias is relatively small, around one percent of the survey's published expenditure estimates. For more background on the CE surveys, see the *Handbook of Method*s (U.S. Bureau of Labor Statistics 2018).

In this paper, we apply the same general approach that has been used by Steinberg et al. (2020), Steinberg, Ash, and Voorheis (2022), and Ash, Nix, and Steinberg (2022a) with the CE Interview Survey; Ash, Nix, and Steinberg (2022b) with the CE Diary Survey; Rothbaum and Bee (2021) with the Annual Social and Economic Supplement of the Current Population Survey (CPS ASEC) and Rothbaum et al. (2021) with the American Community Survey. In this approach, the survey's production estimates are compared with estimates derived from an alternative non-interview adjustment, where the alternative is assumed to be better than the production methodology at accounting for nonresponse bias because it is tailored for a specific year by using alternative methods, alternative variables used to explain nonresponse, or both. The difference between estimates produced by the production weight and the improved non-interview adjustment is then considered a measure of nonresponse bias.

Working from the assumption that the Census administrative records used by Rothbaum and Bee (2021) and Rothbaum et al. (2021) could provide variables with stronger associations to either response or expenditures than the variables used in our current non-interview adjustment, we incorporated the administrative records into non-interview adjustments with the goal of reducing nonresponse bias. We did this by producing three alternative weights that used different sets of variables and different methodologies to account for non-interviews. We refer to the first two weights as the "response" and "expenditure" weights because their non-interview adjustments modeled the probability of response using variables that were associated with either "response" or "expenditures." The third alternative weight was based on the entropy weight, as defined by Hainmueller (2011) and used in Rothbaum and Bee (2021). After applying the alternative non-interview adjustments to the weights, we compared the

[2] See the appendix for the definition of a Census tract. See also Census glossary at: https://www.census.gov/programs-surveys/geography/about/glossary.html (accessed April 7, 2023)

estimates of expenditures produced by the three non-interview adjustments with the estimates produced by the current production weights.

The remainder of the paper is organized as follows. We begin by describing the four sources of data used in the research. Next, we describe the methodology used to calculate the production weight and the three alternative weights. This is followed by the results and a conclusion.

2. Description of Data Sources

The strength of the non-interview adjustment is dependent on the strength of the relationship between nonresponse and the variables used by the non-interview adjustment. As Vartivarian and Little (2002) explain, effective non-interview adjustments use variables that are associated with both (1) the probability that a sample unit will respond to the survey and (2) the variable of interest for the sample unit. Since the variable of interest for CE is total expenditures, we should form non-interview adjustments for CE using either variables that are associated with response, total expenditures, or both. For all three strategies, the variables used in a non-interview adjustment are limited because they must be known for both the interviews and non-interviews.

A unique feature of our research is the large pool of variables that we were able to draw upon for our non-interview adjustment. We were able to leverage variables from the following four sources of data:

1) CE data including the variables used in the current non-interview adjustment.
2) ZIP code income data derived from publicly available IRS files.
3) The administrative records available from the Census.
4) The Census tract-level estimates that were derived from the Census Planning database by Ash et al. (2022)

The next four sections provide more information about the four sources of data for our research.

2.1. CE Interview Data

We used the eligible sample (interviews and non-interviews) from the CE Interview Survey from years 2018, 2019, and the quarters Q1, Q2, and Q3 of 2020.[3] In CE, a sample unit is eligible if contains a consumer unit (CU).[4] When we started work on the research, 2020 Q4 data were not available;

---

[3] We divide sample units into two groups: eligible and ineligible units. A sample unit is eligible if it is in the universe of interest and is ineligible otherwise. Eligible units are further divided into completed interviews and non-interviews. For brevity, we refer to completed interviews as simply interviews and note that the terms respondents and nonrespondents can be used interchangeably with interviews and non-interviews, respectively.
[4] A consumer unit is a group of people living together who are (1) related by blood, marriage, adoption, or some other legal arrangement such as foster children; (2) unrelated but who pool their incomes to make joint expenditure decisions; or (3) is a person living alone or sharing a housing unit with other people but who is

therefore, estimates of 2020 only represented Q1 to Q3. The variables available from the CE Interview Data for the alternative non-interview adjustments included the variables used in the current non-interview adjustment and variables related to the sampling frame and the administration of the survey.

Our expenditure estimates will not match the production estimates found at BLS (2020). The values of the survey expenditure data that we used are preliminary values and only include Interview Survey results. The official estimates use values that include the integrated data from the Interview and the Diary Surveys.

2.2. ZIP Code Income

Currently, the non-interview adjustment uses a variable that represents income, which BLS derives from publicly available IRS ZIP code level data. This variable is derived from an IRS dataset that the IRS generates every year and includes summary-level information about the individual income tax returns filed in nearly every ZIP code of the United States. The datasets are publicly available on the IRS's website and contain information such as the average adjusted gross income per tax return by ZIP code. We use that information to stratify the ZIP codes into the same three categories as researched by Sabelhaus et al. (2013): ZIP codes whose average adjusted gross income is in the top 10 percent of the distribution; ZIP codes whose average adjusted gross income is in the middle 80 percent of the distribution; and ZIP codes whose average adjusted gross income is in the bottom 10 percent of the distribution. These results are then merged with the CE data by the CUs' ZIP codes. For brevity, we will refer to this variable as *ZIP code income*.

The inclusion of ZIP code income in the non-interview adjustment started in 2014 and was motivated by the research of Sabelhaus et al. (2013) and Dumbacher et al. (2012).

2.3. Census Administrative Records

To calculate non-interview adjustments, we need information on both interviewed and non-interviewed CUs. However, by definition, surveys contain limited information on non-interviews. We use Census administrative data linked to the address of the surveyed housing unit, which therefore is available for all CUs, independent of response type.[5] To construct an administrative "roster" of individuals in each address (which is *not* informed by survey responses), we use the Information Returns Master File (IRMF). The IRMF links individuals to addresses for those that receive information returns (including Forms W-2s, 1099-G, 1099-INT, 1099-DIV, 1098, etc.). The IRMF does not contain any income information, only a flag for each form filed.

---

financially independent of the other people. Approximately 99 percent of all occupied housing units have one CU, hence the terms "household" and "consumer unit" are often used interchangeably.

[5] The linking methods we exploit here were developed independently by Census Bureau researchers. See Brummet (2014) for a description of linkage by addresses and Wagner and Layne (2014) for a description of the Person Identification Validation System (PVS) used to assign link individuals to administrative records.

With these administrative rosters, we link the individuals to additional data on their characteristics. From the 2010 decennial census, we observe information on race and Hispanic origin. From the Numident, we observe date of birth, gender, and some citizenship information.[6] We link to W-2, 1099-R, and 1040 tax filing information from the IRS. From the W-2 records, we used information from boxes 1 and 12 on earnings and deferred compensation by job. From the 1099-R records, we observed withdrawals from Defined Contribution retirement plans (such as 401(k)s) and payments from Defined Benefit pension plans. From the 1040 filings, we gathered additional information on the household roster (such as dependent children) and income (including interest, dividends, and taxable social security).

The match rates for the 2018 to 2020 CE samples to the Census administrative records are in Table C1 and are generally consistent to those in Table 4 of Rothbaum and Bee (2021) and Brummet, et al. (2018). Missing values due to nonmatches were treated as another value of each variable in the analyses.

Comparison of Income Variables

Given the access to IRS income variables at the address (household (HH) level) and the use of ZIP code income in the current non-interview adjustment of the production weighting, we were interested in how the HH IRS Income compares to the ZIP code income described in section 1.2. To make the comparisons simple, we derived a new variable based on the percentages of the HH IRS income within the categories of top 10 percent, middle 80 percent, and bottom 10 percent as in ZIP code income. For brevity, we will refer this this variable as *HH income*.

Table 1 shows the differences between the ZIP code income variable used in the production weighting and the HH income derived from the Census administrative records.

Table 1: 2018-2020 Sample Counts by ZIP Code Income and HH Income

HH Income		ZIP Code Income			Total
		Top 10 percent	Middle 80 percent	Bottom 10 percent	
Nonmatches		30	500	100	630
Matches	Top 10 percent	3,300	17,500	1,600	22,400
	Middle 80 percent	7,500	64,000	6,400	77,900
	Bottom 10 percent	300	7,700	3,200	11,200
Total without Nonmatches		11,100	89,200	11,200	111,500
Total with Nonmatches		11,130	89,700	11,300	112,130

[6] The Numident is the Social Security Administration's Numerical Identification System file with a record for each individual that has received a Social Security Number.

We see in Table 1 that about 63 percent [= (3,300 + 64,000 + 3,200)/111,500] of the matched sample units have the same value for both HH Income and ZIP code Income. Additionally, only 2 percent [= (300+1,600)/111,500] of the sample units moved from the Top 10 percent to the Bottom 10 percent or vice versa. This shows that some information is lost when using the ZIP code Income instead of HH Income, but ZIP code income is still informative.

2.4. Census Planning Database

As done in the related research of Ash et al. (2022a, 2022b), we expanded the list of potential variables for our non-interview adjustment by using Census tract-level estimates from the Census Planning Database (U.S. Census Bureau 2015). We merged the Census tract-level estimates from the Planning Database to the CE Interview sample CUs. Then for each variable from the Planning Database, we found the quartiles of the variable using the sample CUs. The value of the quartiles was then assigned to the CUs of the sample for each variable. Table 2 lists the variables that were derived from the Census Planning Database.

Table 2: Census-Tract-Level Quartiles Derived from the Census Planning Database

Description of Variable	**Variable Name on Census Planning Database**
Quartiles of percent of population aged 65 years or older	Pop_65plus_ACS_09_13
Quartiles of percent of population not a high school graduate	Pct_Not_HS_Grad_ACS_09_13
Quartiles of percent of population below the poverty level	Pct_Prs_Blw_Pov_Lev_ACS_09_13
Quartiles of median income	Med_HHD_Inc_ACS_09_13
Quartiles of percent of population Black Alone	Pct_NH_Blk_alone_ACS_09_13
Quartiles of percent of population Hispanic or Latino	Pct_Hispanic_ACS_09_13

The result is that each variable from the Planning Database defines a new variable with four values (1,2,3, and 4). These four values identify the four quartiles or the relative ranking of the CE sample Census tracts with respect to each variable. For example, a value of 1 for the Hispanic variable means that the sample CU is in a Census tract that is in the bottom quartile of all CE sample Census tracts with respect to the proportion of the tract's Hispanic population. Therefore, CUs with a value of 1 have the smallest proportion of Hispanic population, and Census tracts with a value of 4 have the largest proportion of Hispanic population when considering the Census tracts of our sample.

Less than 2 percent of the Interview sample did not match the Census Planning Database. Since the percentage of mismatches was so small, we included them in quartile 2 because we assumed that their average was similar to the overall average or to one of the two middle quartiles.

For more information on the ZIP code-level IRS data, see IRS (2020).

3. Description of the Weights

Our research compares the estimates of mean expenditures using a total of nine weights. This includes the base weight (see Section 3.1), the production weight (see Section 3.2), and the following three alternative weights:

- Entropy Weight (see Section 3.3)
- “Response” Weight (see Section 3.4)
- “Expenditure” Weight (see Section 3.4)

With the production weight and each of the three alternative weights there are two weights: a non-interview weight and a final weight. The *non-interview weight* is an intermediate weight that is produced within the weighting process and is the product of the base weight and the non-interview adjustment. Note that the base weight is defined in section 3.1 and the non-interview adjustment adjusts the base weights to account for non-interviews during the interviewing process. After the application of the non-interview adjustment, the interviews can produce estimates that represent all eligible CUs (both interviews and the non-interviews).

The *final weight* is the final weight produced by the weighting process and is the product of the base weight, the non-interview adjustment, and an adjustment to demographic totals. We note that our new term *adjustment to demographic totals* is referred to as the “calibration adjustment” by CE. We use a more general term *adjustment to demographic totals* to avoid confusion – the methodology of the entropy weight uses calibration in its non-interview adjustment and adjustment to demographic totals.

Borrowing from Haziza and Beaumont (2017), the goals of the adjustments for demographic totals are threefold:

- to force consistency of certain survey estimates to known population quantities;
- to reduce nonsampling errors such as coverage errors;
- to improve the precision of estimates.

The adjustments for demographic totals should not have much impact on expenditure estimates, if coverage errors are minimum, or coverage errors are either constant or differential in a way that is not related to expenditures.

In total, the research considers nine weights: one base weight and eight other weights – (non-interview weights and final weights) x (production, entropy, “response”, and “expenditure weights”).

Comparing the estimates derived from the non-interview weights and final weights provides useful insight because it allows us to examine the cumulative impact of the separate weighting adjustments. For example, if the estimate of expenditures produced with the non-interview weight is larger than the estimate produced with the base weight, we can say that the non-interviews have expenditures that are generally higher than the interviews. In order for the estimate of expenditures to decrease, interviews with smaller expenditures had to have their weights increased more than interviews with larger expenditures.

The next five sections review the base weight, production weight and the three alternative weights. For the production weight and alternative weights, we review both their non-interview adjustments and adjustment to demographic totals. The last section reviews the replicate weights that are produced for all three alternative weights and that are then used to estimate sample variances.

3.1. Base Weight

Each CU in the sample has a *base weight* associated with it, which is the number of CUs in the population it represents. The sum of the base weights of all the CUs in the sample is the total number of CUs in the nation, which is around 132 million for collection year 2020. The alternative weights all used the same base weight as their starting point.

Note that a mean of a variable of interest calculated with only the base weight represents the mean for completed interviews only since it does not account for non-interviews.

3.2. Production Weight

We define the *production weight* as the weight that is used to produce the official expenditure estimates that are provided to the public. The estimates produced using the production weights serve as the baseline in our comparisons.

3.2.1. Production Weight – Non-Interview Adjustment

The non-interview adjustment process uses the cell adjustment method, where the complete sample of CUs is partitioned into 192 cells according to the region of the country in which they live; the number of people in their CUs; the number of contact attempts made by the survey's field representatives when trying to collect their data; and the average income in their ZIP code according to the IRS. The probability of a CU in the sample participating in the survey is estimated for each of the 192 cells by dividing the sum of the base weights from the interviews by the sum of the base weights from all CUs in the sample (interviews plus non-interviews) within each cell. Then, the weights of the interview CUs are increased to account for the non-interview CUs by multiplying them by the inverse of their cell's estimated probability of participating in the survey. For technical background on the cell adjustment method, see also Oh and Scheuren (1983), Little (1986), and Brick (2013).

The non-interview cells are formed by crossing the following five variables of Table 3:

Table 3: Variables Used to Define the Production Non-Interview Adjustment

Variable	Values
Month of interview	1 to 12
Census Region	Northeast, Midwest, South, West
Number of contacts made during interviewing	1, 2, 3-4, and 5+ contacts
Consumer Units Size	1, 2, 3-4, and 5+ persons
ZIP code level IRS income of selected quintiles	Top 10 percent, Middle 80 percent, Bottom 10 percent

We collapsed the non-interview adjustment to mitigate extreme values of the non-interview factor that could result in influential sample units and increased variances. Non-interview adjustment cells are collapsed for maximum values of the non-interview adjustment factor. Between 2018 and 2020, the maximum value of any production non-interview adjustment factor was 2.6.

After the application of the weighting cell non-interview adjustment, the sum of the non-interview adjusted weights over the interviews is equal to the sum of the base weights for the interviews and non-interviews and this true for all of the collapsed cells formed by the variables in Table 3.

3.2.2. Production Weight – Adjustment to Demographic Totals

The final step in deriving the production weight is to apply an *adjustment to demographic totals*. The adjustment ensures that demographic estimates derived from the final weights are consistent with known demographic totals. The goal of the adjustment to demographic totals is to improve the differential coverage by demographic characteristics and reduce the variances of estimates. The CE production weight uses calibration with a restricted Generalized Least Squares distance function based on Jayasuriya and Valliant (1996). This methodology finds new weights that produce consistent estimates of the demographic totals but change the least amount from the non-interview weight. Additionally, weights are constrained for extreme values in an effort to reduce extreme weights.

What we refer to as the *adjustment to demographic totals* is usually referred to as the *calibration adjustment* by CE. In this paper, we are using a new and more general term because, in our research, the calibration method is sometimes used by the non-interview adjustment and is also used with the adjustment to demographic totals. So, our new term *adjustment to demographic totals* refers to the general step of the weighting that ensures that demographic estimates derived from the final weights are consistent with known demographic totals and *calibration* is a method that can be used by either the non-interview adjustment or the adjustment to demographic totals.

Table 4 summarizes the demographic totals used with adjustment to demographic totals of the production weight.

Table 4: Demographic Totals of Production Weight

Number of Totals	Description of Totals
1	U.S. households
14	U.S. population for each of 7 age categories x 2 race (black/nonblack) categories
9	U.S. population by Census Division
9	U.S. Urban population by Census Division
1	U.S. Hispanic population
1	U.S. homeowner households
35	Number of totals

Attachment A summaries how the 14 Age and Race totals and Census Divisions of Table 4 are defined.

3.3. Entropy Weight

The *entropy weight* uses an approach that is similar to the methodology that Rothbaum and Bee (2021) used to examine nonresponse bias during the COVID-19 pandemic with the CPS ASEC. Their research compared estimates from CPS ASEC with estimates produced from their entropy weight and Census Administrative records. Their entropy weight is based on Hainmueller (2012), who applied it to observational studies. Note that the general calibration method of Hainmueller (2012) is also described by Deville and Särndal (1992) for surveys. Appendix B provides a technical description of the weights and the variables used in the non-interview adjustment.

In comparison with the other alternative weights, the entropy weight is unique because it used more variables in its non-interview adjustment and its adjustment for demographic totals used a much different methodology as compared to the production weight and the other two alternative weights.

3.4. Response and Expenditure Weights

As mentioned previously, Vartivarian and Little (2002) explain that effective non-interview adjustments use variables that are either associated with (1) the probability that a sample unit will respond to the survey, (2) the variable of interest for the sample unit, or (3) associated with both (1) and (2). For our research, we produced an alternative "response" weight that uses variables that are associated with the probability of response and a second alternative "expenditure" weight that uses variables that are associated with the variable of interest: expenditures. We tried both types of non-interview adjustments because one of the two approaches would be better than the other and this tells us which set of associations – "response" or "expenditures" – was better at reducing nonresponse bias. We cannot do this with the entropy weight because it uses variables associated with both "response" and "expenditures."

3.4.1. “Response” and “Expenditure” Weights – Non-interview Adjustment

The two parts of a non-interview adjustment are the methodology used to implement the adjustment and the variables used by the methodology. We now review the non-interview adjustment for the “response” and “expenditures” weights in terms of the two parts.

Methodology of the “Response” and “Expenditure” Non-interview Adjustments. We calculated both the “response” and “expenditure” non-interview adjustments with the same calibration methodology as used by the production adjustment for demographic totals. With the calibration for nonresponse, our “known totals” are the sum of the base weights for both the interviews and non-interviews for each of the 40 variables used in the calibration.

With both the weighting cell and calibration method, the sum of the non-interview adjusted weights across the interviews is equal to the sum of the base weights for the interviews and non-interviews. The difference between the two methods is that the weighting cell approach keeps the combinations of the variables equal, while the calibration method does not. For example, the variables representing number of persons in the CU and number of contacts both have four values. With the weighting cell method, the weighted totals for all 16 combinations of number of persons and number of contacts would be consistent, whereas with the calibration method, the four totals of each variable, for a total of eight, would be consistent.

We used calibration instead of the weighting cell method because it had the following advantages: ability to handle a large number of variables easily; ease of implementation; flexibility with respect to changing variables; and the lack of a need to collapse cells.

Variables used by the “Response” and “Expenditure” Non-interview Adjustments.
We used a logistic regression model to evaluate binary variables for the “response” non-interview adjustment and a linear regression model of continuous variable expenditures to evaluate variables for the “expenditures” non-interview adjustment. The candidate variables for both models came from the three sources described in section 2. With both the “response” and “expenditure” non-interview adjustments, we also used a combination of forward and backward model selection methods along with significance testing of model parameters to select the variables. Table 5 lists the variables selected for both non-interview adjustments.

Table 5: Variables Used in the "Response" and "Expenditure" Non-interview Adjustments

Source	Variable	Values	Non-interview Adjustment	
			"Response"	"Expenditure"
CE Interview Data	Number of Persons in CU	1, 2, 3-4, 5+	X	X
	Number of Contacts	1, 2, 3-4, 5+	X	
	Census Division	9 Categories	X	X
	Interview Number	1, 2, 3, 4	X	
	Quarter within year	1, 2, 3, 4	X	
	Type of Primary Sample Unit	SR, NSR/Urban, NSR/Rural	X	X
	Group Quarters?	0/1	X	X
IRS ZIP Code Data	Quartiles of 65+	4 Quartiles	X	
	Quartiles of Adjusted Gross Income	4 Quartiles		X
	Quartiles of High School Graduate	4 Quartiles		X
Census Administrative Records	Total Wages	Continuous	X	X
	1099R Retirement Income	Continuous	X	
	Presence of Schedules A	0/1	X	
	Presence of Schedules D	0/1	X	X
	Presence of Schedules E	0/1	X	X
	Presence of Schedules F	0/1	X	X
	Presence of Schedules S	0/1		X
	Presence of 1099 DIV	0/1	X	X
	Presence of 1099 S	0/1		X
	Presence of 1098	0/1	X	X
	Presence of SSA 1099	0/1		X
	Presence of any AGI	0/1		X
	Presence of any WSI	0/1		X
	Presence of anyone aged 18-25	0/1		X
	Presence of anyone aged 25-35	0/1		X
	Presence of anyone aged 35-45	0/1		X
	Presence of anyone aged 45-55	0/1		X
	Presence of anyone aged 65+	0/1	X	

X – indicates the variable was included in the given model.
0/1 – indicates the variable was binary with values 0 and 1.

Next, we make two observations on the variables selected for the "response" or "expenditures" non-interview adjustments.

First, the "response" or "expenditures" non-interview adjustments had more variables than the production weight but less than the entropy weight. We included the 40 variables with the largest associations to either "response" or "expenditures." Note that the 40 variables included groups of variables that represented one categorical variable. For example, the categorical variable Census

Division was counted as nine of the 40 variables. We limited the model to 40 variables. Note that this yields a smaller range than 40 or more would.

Second, several variables were selected by both non-interview adjustments. This means that some variables selected by each non-interview adjustment were associated with both “response” and “expenditures.”

3.4.2. “Response” and “Expenditure” Weights – Adjustment to Demographic Totals

We used the same methodology and the same demographic totals as described for the production weight in section 3.1.2. This means that the only difference between the “response”, “expenditure”, and production final weights is the non-interview adjustment.

3.5. Replicate Weights

CE uses balanced repeated replication (BRR) [McCarthy 1966], a replication method for estimating variances, to estimate the variance of estimates due to CE’s sample design. In our research, we also used BRR to estimate the sample variances. To use BRR, we produced replicate weights for each of the three alternative weights. All replicate weights used the same set of replicate factors that are used with the production of CE variance estimates. With all three alternative replicate weights, the non-interview adjustment and the adjustment to demographic totals was recalculated for each replicate with the aim of accounting for the impact of the weighting adjustments on the sample variances. See Swanson (2017) for additional information about variance estimation for CE.

4. Results

This section presents the results of our analyses. It presents the magnitudes of the non-interview adjustments, final weights for each year’s production weights, and the three alternative weights. Comparisons of expenditure estimates produced from all weights considered are shown.

4.1. Magnitudes of the Non-interview Adjustments and Final Weights

We begin by examining some descriptive statistics regarding the magnitude of the alternative non-interview adjustments and final weights. Table 6 compares the current non-interview adjustment with the alternatives in terms of their unit-level differences and provides the mean, minimum, and maximum of the absolute value of the differences. Table 6 also shows the minimum and maximum of each of the alternative non-interview adjustments.

Table 6: Non-interview Adjustment Descriptive Statistics

Year	Method	Absolute Value of Difference between the Production and the Alternative Non-interview Adjustments	Non-interview Adjustment

		Mean	Minimum	Maximum	Minimum	Maximum
2018	0 - Production	n/a	n/a	n/a	1.00	2.00
	1 - Response	0.33	0.00	1.25	1.00	2.81
	2 - Expenditure	0.23	0.00	1.06	1.00	2.35
	3 - Entropy	0.60	0.00	22.55	0.00	24.34
2019	0 - Production	0.00	0.00	0.00	1.00	2.00
	1 - Response	0.39	0.00	1.49	1.00	3.30
	2 - Expenditure	0.25	0.00	1.40	1.00	2.83
	3 - Entropy	0.61	0.00	15.02	0.48	17.02
2020	0 - Production	0.00	0.00	0.00	1.00	2.60
	1 - Response	0.63	0.00	2.40	1.00	4.64
	2 - Expenditure	0.42	0.00	1.97	1.00	3.37
	3 - Entropy	1.38	0.00	40.30	0.03	41.50

In Table 6, we see that the entropy non-interview adjustment, relative to the other weights, produces both very small (values less than 1.0) and large values (maximum of 41.50) for the non-interview factor. The magnitudes of “response” and “expenditure” non-interview adjustments were more consistent with the production non-interview adjustments because we only included 40 variables (adding more variables, which is equivalent to adding more constraints, would have caused the non-interview adjustments to become more extreme). Although we want the non-interview adjustment factors to vary so that they account for the differential nonresponse, we generally want to avoid extreme adjustments because they have the potential to increase variances.

Table 7 compares the current final weights with the alternative methods in terms of their unit level differences and provides the mean, minimum, and maximum of the absolute value of these differences. Table 7 also shows the minimum and maximum of each of the alternative final weights.

Table 7: Final Weight Descriptive Statistics

Year	Method	Absolute Value of Difference between the Production and Alternative Final Weights			Final Weight	
		Mean	Minimum	Maximum	Minimum	Maximum
2018	0 - Production	n/a	n/a	n/a	1,249	97,370
	1 - Response	4,039	2	38,810	1,475	90,140
	2 - Expenditure	2,907	0	33,570	2,101	64,340
	3 - Entropy	7,823	1	824,600	0	884,800
2019	0 - Production	n/a	n/a	n/a	1,252	120,400
	1 - Response	4,820	0	48,950	1,387	129,500
	2 - Expenditure	2,974	0	49,410	2,003	113,500
	3 - Entropy	7,615	0	146,500	2,031	165,900
2020	0 - Production	n/a	n/a	n/a	1,103	133,800
	1 - Response	7,455	0	63,650	1,134	154,400
	2 - Expenditure	4,249	0	63,500	1,627	148,600
	3 - Entropy	13,550	0	558,700	2	658,700

In Table 7, we see that the entropy final weights had both very small (values less than 1.0 in 2018) and large values in every year when we compare them to the production weight. The smallest entropy weight was less than 1.0 in 2018 and the largest weight was 884,800 in 2018. The “response” and “expenditure” final weights were more consistent with the production final weights with respect to magnitude.

4.2. Comparisons of Expenditure Estimates

4.2.1 National Estimates of Total Expenditures

The next set of the results compare the national expenditure estimates using the nine different weights. A limitation of these comparisons is that we do not know the actual value for mean total expenditures and therefore we cannot say which of the alternative methods is the best. However, the comparisons do shed light on the non-interview adjustments and their impact on the estimates.

Figure 1 shows the 2018 expenditure estimates and their 95 percent confidence intervals for the nine weights of our research. Note that the expenditure estimates and their standard errors represented in Figures 1 to 4 can be found in Table C1 of the appendix.

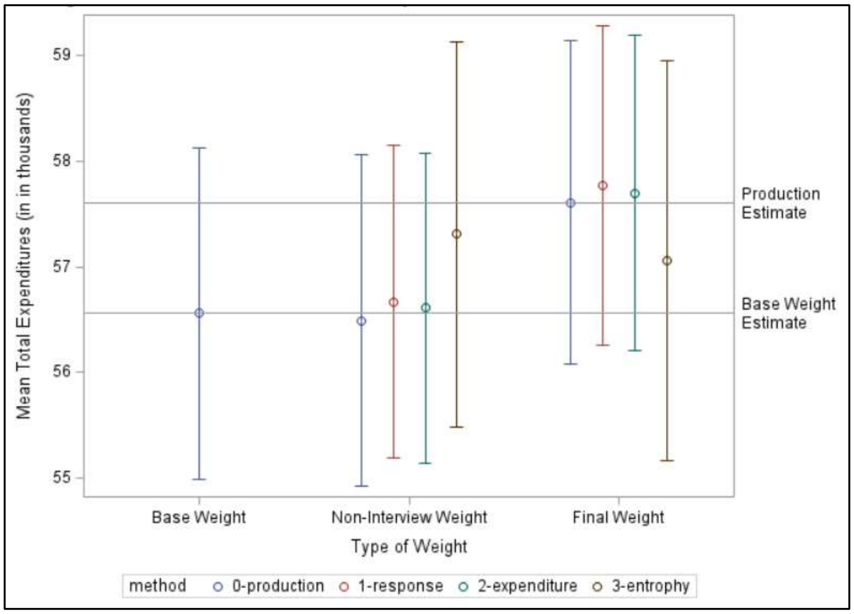

Figure 1: 2018 National Mean Total Expenditures – 95 percent Confidence Intervals

Several conclusions can be drawn from the expenditure comparisons of Figure 1. The first and primary conclusion that directly addresses the main objective of nonresponse bias is listed first below and is followed by several secondary conclusions, identified as "notes" that we also discuss.

Conclusion: The 2018 production expenditure estimates were not substantially different than the estimates derived from the alternative non-interview adjustments: there is no evidence of nonresponse bias. In Figure 1, all three alternative final weights produced 2018 expenditure estimates that were not substantially different than the production estimate of $57,610, when we consider the magnitude of estimates or the standard errors. The estimate of expenditures derived from the final "response" and "expenditure" weights were $160 and $90 larger than the final production weight, respectively. The estimate of expenditures derived from the entropy weight were $550 lower than the production estimate. All the differences are small when compared to the estimate of expenditures and the standard errors which ranged from $762 to $964. This suggests that the differences between the alternative weights and the production weights were not substantially different relative to the magnitude of the expenditure estimates or their standard errors.[7] More generally, it suggests that there is no evidence of nonresponse bias because the non-interview adjustments do not have much room for improvement.

Although the alternative weights did not make a substantial change in the 2018 expenditures estimates, there are some interesting differences between the methodologies and the separate weighting adjustments with respect to the estimates of Figure 1. We next provide some observations about some of the differences of the weights and methodologies; however, since all of the differences are small, we suggest considerable caution in using the conclusions from these observations.

[7] The length of each side of the 95 percent confidence intervals in Figure 1 is 1.96 times the standard error.

In addition to the primary conclusion, we note the following secondary conclusions.

Note 1: The 2018 entropy weight acted differently than the other 2018 weights. First, the non-interview entropy weight increased the estimate of expenditures from the base weight estimate of $57,610 to $57,310 or an increase of $750. The other weights made little change – differences of -$70, $110, and $50 for the production, "response," and "expenditure" non-interview weights, respectively. Second, the adjustment for demographic totals had the impact of lowering the estimate of expenditures by $250 when we compare the estimate derived from the final weight of $58,310 with the estimate derived from the non-interview weight of $57,060. With the other weights, the adjustment to demographic totals had the opposite impact of increasing the expenditure estimates – increases of $1,120 for the production weight, $1,100 for the "response" final weight, and $1,190 for the "expenditure" final weight.

Note 2: The non-interview adjustments had little impact on the 2018 estimate of expenditures compared with the estimate derived from the base weight. If there is nonresponse bias, we would expect that expenditure estimates would exhibit a significant change – either increase or decrease. However, the non-interview weights in Figure 1 either had no effect on the estimates of expenditure. The estimate of $56,560 produced by the base weight is not much different than the estimates for the production, "response", and "expenditure" non-interview final weights – differences of -$70, $110, and $50, respectively. The exception is the entropy final weight which produced the largest estimate of $57,310, or a difference of $750. The non-interview adjustments may not have much impact because, as mentioned previously, prior studies by Chopova et al. (2009) and Steinberg et al. (2022) found that nonresponse bias is relatively small, around one percent of the published expenditure estimates.

Note 3: The 2018 estimates produced by the "response" and "expenditure" weights were not much different. The point of having separate "response" and "expenditure" non-interview adjustments is to determine whether the variables associated with "response" or "expenditures" had more impact on the non-interview adjustments. However, the estimates for the two weights did not differ much. The 2018 estimate for total expenditures for the non-interview weight is $56,670 for the "response" weight and $56,610 for the "expenditure" weight for a difference of $60. Similarly, the final weights are not much different – $57,770 for the "response" weight and $57,700 for the "expenditure" weight for a difference of $70. We suspect that the weights produced similar results because there was considerable overlap in the variables used by the non-interview adjustment of each weight.

Note 4: The adjustment to demographic totals had more impact on the 2018 estimates than the non-interview adjustment. This is perhaps the most surprising finding of the research. With the production weight, the adjustment for demographic totals increased the estimate of total expenditures from $56,490 for the non-interview weight to $57,610 for the final weight for an increase of $1,220. The "response" and "expenditure" weights exhibited similar increases between the estimates derived from the interview weight and the final weight – increases of $1,100 for the "response" weight and $1,090 for the "expenditure" weight.

We find the large impact of the adjustment for demographic totals surprising because it suggests that coverage is differential in a way that is related to expenditures. If the impact of the adjustment for demographic totals was small or constant, there would be no difference between the estimates derived from the non-interview weights and the final weights. However, we see that the adjustment for demographic totals has the impact of increasing the expenditure estimates which means that coverage was differential in a way that was related to expenditures. Specifically, the weights of sample CUs with larger expenditures were increased more than sample units with smaller expenditures.

Figure 2 shows the 2019 expenditure estimates and their 95 percent confidence intervals for the nine weights of our research.

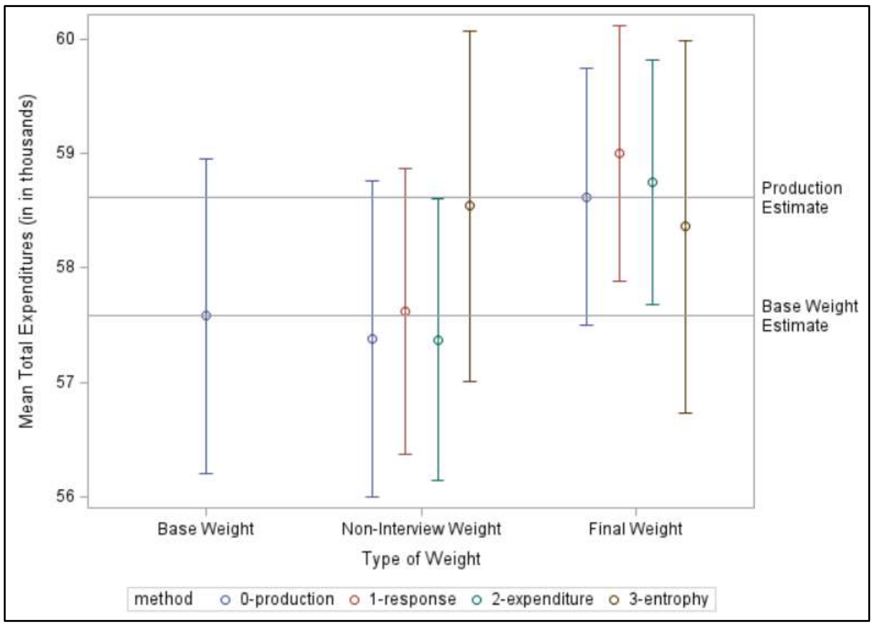

Figure 2: 2019 National Mean Total Expenditures – 95 Percent Confidence Intervals

The first thing we notice about Figure 2 is how similar it is to Figure 1. This is not unexpected because prior to the COVID-19 pandemic, the weighting and the estimates produced by the weights had been consistent. Because Figure 2 is so similar to Figure 1, we attribute to Figure 2 the conclusion and notes 1 to 4 that we attributed to Figure 1 without repeating them.

Figure 3 shows the 2020 expenditure estimates and their 95 percent confidence intervals for the nine weights of our research. Note that the estimates represented in Figure 3 include completed interviews from the start of the COVID-19 pandemic.

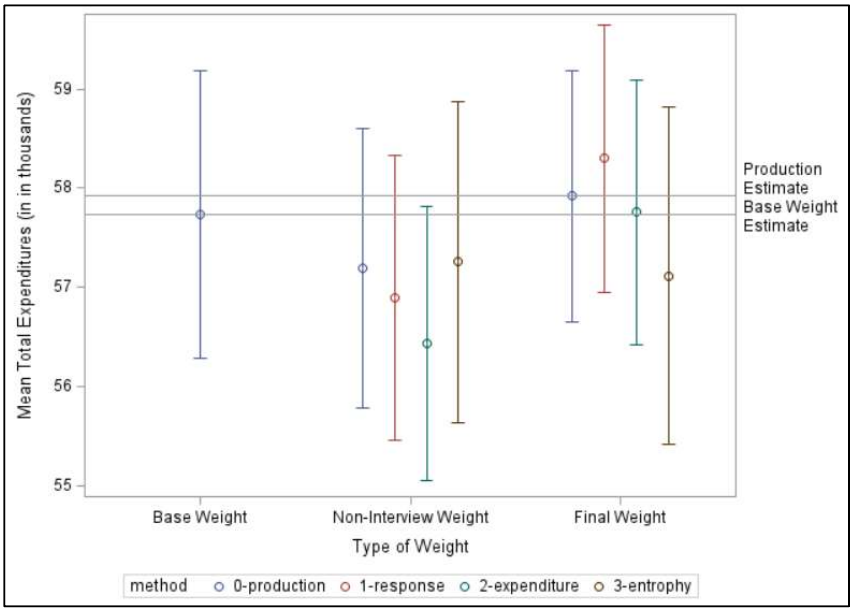

Figure 3: 2020 National Mean Total Expenditures – 95 Percent Confidence Intervals

Conclusion: The 2020 production expenditure estimates were not substantially different than the estimates derived from the alternative non-interview adjustments: therefore, there is no evidence of nonresponse bias. In Figure 3, all three alternative final weights produced estimates of 2020 expenditures that were not substantially different than the production estimate of $57,920 when we consider the magnitude of estimates or the standard errors. The estimate of expenditures derived from the final "response" weight was $380 larger than the final production weight. The estimate of expenditures derived from the "expenditure" and entropy weights were $160 and $490 lower than the production estimate, respectively. The differences are small when compared to the estimate of expenditures and the standard errors which ranged from $646 to $867. As with 2018 and 2019, the small differences in expenditure estimates suggest that there is no evidence of nonresponse bias because the non-interview adjustments do not have much room for improvement.

Next, we provide one observation about the differences between the 2020 estimates of Figure 3 and the 2018 and 2019 estimates of Figure 1 and Figure 2. We again mention that all of the differences are small and suggest considerable caution in using the conclusions from these observations.

Note 5: The non-interview adjustments reduced the 2020 expenditure estimates compared with the estimate derived from the base weight. This is different than with 2018 and 2019, where the non-interview adjustment had little impact. In 2020, the production non-interview adjustment decreased the base weight estimate of $57,470 to $57,200 or a difference of $540. We also see the same decrease with the alternative weights – decreases of $840, $1,300, and $480 for the "response," "expenditure," and entropy non-interview weights, respectively. This suggests that the 2020 non-interview adjustment increased the weights of CUs with lower levels of expenditures more than the CUs with higher levels of expenditures as compared to the non-interview adjustments of 2018 and 2019. This does not mean that the 2020 expenditure estimates are in any way faulty, but it does suggest that the way in which the sample responded to CE before the COVID-19 pandemic was different than during the start of the COVID-19 pandemic.

4.2.2. Regional Estimates of Total Expenditures

Figure 4 shows the 2020 national and regional expenditure estimates and their 95 percent confidence intervals for the nine weights of our research.

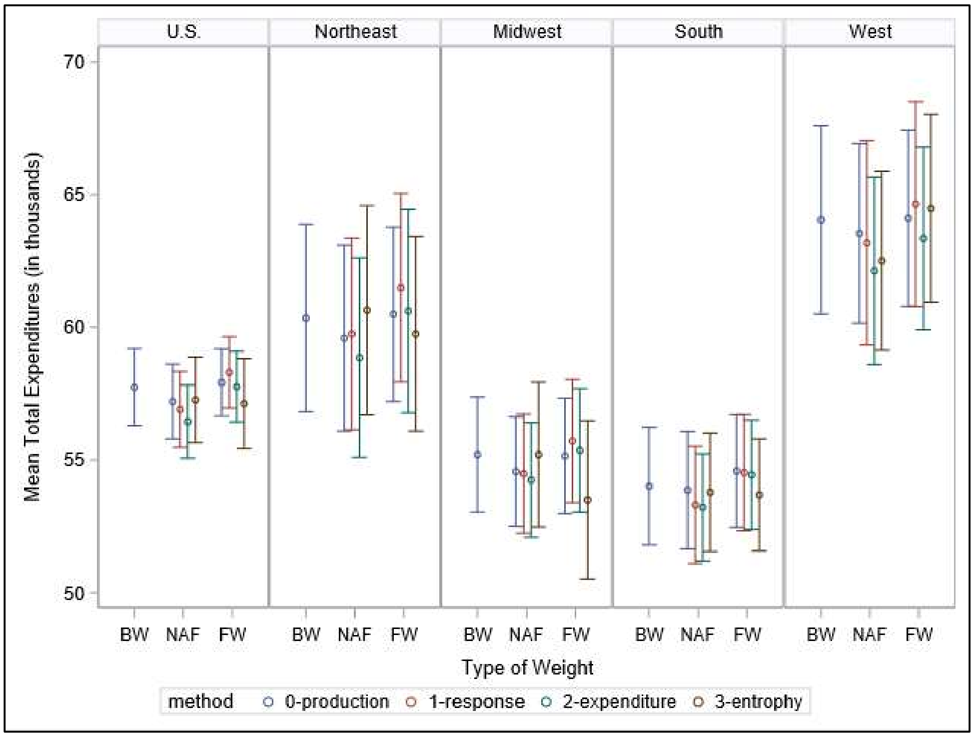

Figure 4: 2020 National Mean Total Expenditures – 95 Percent Confidence Intervals

The labels in Figure 4 are BW for base weight, NAF for non-interview adjustment factor, and FW for final weight.

Figure 4 compares the 2020 national expenditure estimates with the expenditure estimates of the four Census Regions to determine whether the regional estimates differ in any way with the national estimates. Generally, the regional expenditure estimates of Figure 4 show similar patterns of increases and decreases as compared to the national estimates, relative to their respective regional means. We note that the confidence intervals for the regional estimates are larger than those for the national estimates due to the smaller sample sizes. Given how similar the regional estimates are to the national estimates, we suggest that the conclusions that we made about the national estimates also apply to the regional estimates.

Other expenditure estimates. We choose not to review the results for the other expenditure estimates in detail because they generally repeat what we have seen with total expenditures. This means that all of the conclusions we have made about total expenditures transfer to the other five types of expenditures that we considered. The graphs of the other expenditures are included in Appendix C.

5. Conclusion

In this work, we examined nonresponse bias in CE expenditure estimates by producing three different weights that were specifically tailored to 2018 to 2020, where alternative methods and alternative variables were used to produce the non-interview adjustments. The “response” and “expenditure” weights used calibration in its non-interview adjustment, and the entropy weight applied calibration methodology as used in Rothbaum and Bee (2021). The expenditure estimates produced with the alternative weights were then compared with the production estimates.

Our research found no evidence of nonresponse bias with the expenditure estimates for the first three quarters of 2020 (which includes the start of the COVID-19 pandemic) or with the 2018 and 2019 estimates. Although we produced specialized non-interview adjustments and weights for the period of 2018 to 2020, none of our efforts changed the expenditure estimates substantially.

Census Disclaimer. Given that this paper is a joint effort of BLS and the Census Bureau using administrative data from the Census, we include the following Census disclaimer:

> The U.S. Census Bureau has not reviewed the paper for accuracy or reliability and does not endorse its contents. Any conclusions expressed herein are those of the authors and do not necessarily represent the views of the U.S. Census Bureau. All results were approved for release by the U.S. Census Bureau, authorization numbers **CBDRB-FY22-SEHSD003-011**.

References

Armstrong, G., Grayson, G., Miller, T. and Pham, S. (2022). "Reference Guide CE Data Quality Profile," Consumer Expenditure Surveys Program Report Series, April 29, 2021. Available at: https://www.bls.gov/cex/research_papers/pdf/dqp-reference-guide-midyear-2021.pdf (accessed March 15, 2024).

Ash, S., Nix, B., and, Steinberg, B. (2022). Report on Nonresponse Bias during the COVID-19 Period for the Consumer Expenditures Interview Survey, dated March 2022. Available at: https://www.bls.gov/cex/research_papers/pdf/ash-nix-steinberg-covid-nonresponse-diary-report.pdf (accessed March 15, 2024).

Brick, J.M. (2013). "Unit Nonresponse and Weighting Adjustments: A Critical Review," Journal of Official Statistics, 29, 329–353. Available at: https://doi.org/10.2478/jos-2013-0026 (accessed March 15, 2024).

Brummet, Q. (2014). "Comparison of Survey, Federal, and Commercial Address Data Quality." U.S. Census Bureau CARRA Working Paper #2014-06. Available at: https://www.census.gov/library/working-papers/2014/adrm/carra-wp-2014-06.html. (accessed March 15, 2024).

Brummet, Q., Flanagan-Doyle, D., Mitchel, J., Voorheis, J., Erhard, L., and McBride, B. (2018). Investigating the Use of Administrative Records in the Consumer Expenditure Survey, CARRA Working Paper 2018-01. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2018/adrm/carra-wp-2018.pdf (accessed March 15, 2024).

Bureau of Labor Statistics. (2018). Handbook of Methods. Available at: BLS Handbook of Methods, Chapter 17. The Consumer Price Index (Updated 2-14-2018) (accessed March 15, 2024).

Bureau of Labor Statistics. (2022, March 25). Consumer Expenditure Survey: CE Tables, Bureau of Labor Statistics. Available at: https://www.bls.gov/cex/tables.htm (accessed March 15, 2024).

Chopova, B., Edgar, J., Gonzalez, J., King, S., McGrath, D.E., and Tan, L. (2009). U.S. Bureau of Labor Statistics Report. "Assessing Nonresponse Bias in the CE Interview survey," dated 2008. Available at: https://www.bls.gov/osmr/research-papers/2009/st090220.htm (accessed March 15, 2024).

Deville, J-C. and Särndal, C.-E. (1992). Calibration Estimators in Survey Sampling. Journal of the American Statistical Association 418, 376-382. Available at: https://doi.org/10.2307/2290268 (accessed March 15, 2024).

Dumbacher, B., Swanson, D., Reyes-Morales, S., Steinberg, B., Ash, S., and Castelo, D. (2012). U.S. Census Bureau memorandum entitled "Consumer Expenditure Surveys 2010 Redesign Research: Noninterview and Calibration Weighting Adjustments for the Consumer Expenditure Surveys," from Ruth Ann Killion to Jay Ryan, dated September 26, 2012.

Hainmueller, J. (2011). Entropy Balancing for Causal Effects: A Multivariate Reweighting Method to Produce Balanced Samples in Observational Studies, Political Analysis, 20, 25-46. Available at: http://dx.doi.org/10.2139/ssrn.1904869 (accessed March 15, 2024).

Haziza, D. and Beaumont, J.-F. (2017). Construction of Weights in Surveys: A Review, Statistical Science, 30, 206-226. Available at: https://projecteuclid.org/journals/statistical-science/volume-32/issue-2/Construction-of-Weights-in-Surveys-A-Review/10.1214/16-STS608.pdf (accessed March 15, 2024).

Internal Revenue Service. (2020, March 25). SOI Tax Stats - Individual Income Tax Statistics - ZIP Code Data (SOI), IRS website. Available at: https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-ZIP code-data-soi (accessed March 15, 2024).

Jayasuriya, B.R. and Valliant, R. (1996). An application of Restricted Regression Estimation in Household Surveys, Survey Methodology, 22, 127-137. Available at: https://www150.statcan.gc.ca/n1/en/pub/12-001-x/1996002/article/2979-eng.pdf?st=utZCzQ48 (accessed March 15, 2024).

Knappenberger, C., Lee, Y., Pham, S., and Armstrong, G. (2021). "The Midyear CE Data Quality Profile - 2020," Consumer Expenditure Surveys Program Report Series, April 29, 2021. Available at: https://www.bls.gov/cex/dqreport-midyear-2020.pdf (accessed March 15, 2024).

Lee, Y., and Biagas, D. (2021). "COVID-19 Response and Data Quality Impacts in the CE Interview Survey," Consumer Expenditure Surveys Program Report Series, October 2021. Available at: https://www.bls.gov/cex/research_papers/pdf/lee-biagas-covid-19-response-and-data-quality-impacts-in-the-ce-interview-survey.pdf (accessed March 15, 2024).

Little, R.J.A. (1986). Survey Nonresponse Adjustments for Estimates of Means. International Statistical Review, 54, 139–157. Available at: https://doi.org/10.2307/1403140 (accessed March 15, 2024).

Oh, H.L., and Scheuren, F.S. (1983). Weighting adjustments for unit nonresponse. In *Incomplete Data in Sample Surveys*, 2, Theory and Bibliographies, (Eds. W.G. Madow).

Rothbaum, J. and Bee, A. (2021). Coronavirus Infects Surveys, Too: Survey Nonresponse Bias and the Coronavirus Pandemic, dated May 3, 2021. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2020/demo/sehsd-wp2020-10.pdf (accessed March 15, 2024).

Rothbaum, J., Eggleston, J., Bee, A., Klee, M., and Mendez-Smith, B. Addressing Nonresponse Bias in the American Community Survey During the Pandemic Using Administrative Data, dated November 30,

2021. Available at: https://www.census.gov/content/dam/Census/library/working-papers/2021/acs/2021_Rothbaum_01.pdf (accessed November 1, 2023).

Sabelhaus J., Johnson, D., Ash, S., Swanson, D., Garner, T., and Henderson, S. (2013). "*Is the Consumer Expenditure Survey Representative by Income?*," (NBER Working Paper No. 19589, October 2013). Available at: https://www.nber.org/system/files/working_papers/w19589/w19589.pdf (accessed March 15, 2024).

Steinberg, B., Krieger, S., McBride, B., Nix, B., Sverchkov, M., and Yang, D. (2022). U.S. Bureau of Labor Statistics Report "A Nonresponse Bias Study of the Consumer Expenditure Survey for the Ten-Year Period 2007-2016," Consumer Expenditure Surveys Program Report Series, May 9, 2021. Available at: https://www.bls.gov/cex/research_papers/pdf/steinberg-et-al-a-nonresponse-bias-study-of-the-consumer-expenditure-survey-for-the-ten-year-period-2010-2019.pdf (accessed March 15, 2024).

Steinberg, B., Voorheis, J., Reyes-Morales, S., and McBride, B. (2020). "Using Administrative Data to Improve Nonresponse Weighting Procedures in the Consumer Expenditure Survey," JSM, Section on Government Statistics. Available at: https://www.bls.gov/osmr/research-papers/2020/st200050.htm (accessed March 15, 2024).

Steinberg, B., Ash, S., and Voorheis, J. (2022). "Using Administrative Data to Improve Nonresponse Weighting Procedures in the Consumer Expenditure Survey: Follow Up Analysis," forthcoming.

Swanson, D. (2017). Consumer Expenditure Surveys Program Series, "Standard Errors in the 2016 Consumer Expenditure Survey," September 2017. Available at: https://www.bls.gov/cex/research_papers/pdf/ce-standard-error-article.pdf (accessed March 15, 2024).

U.S. Census Bureau. (2015). "Planning Database with 2010 Census and 2009 – 2013 American Community Survey Data, At the Tract Level," July 25, 2015. Available at: https://www.census.gov/data/developers/data-sets/planning-database.2015.html (accessed March 15, 2024).

U.S. Census Bureau. (2020). "Frequently asked questions: The impact of the coronavirus (COVID-19) pandemic on the Current Population Survey/Housing Vacancy Survey (CPS/HVS)." Available at: https://www.census.gov/housing/hvs/files/qtr220/impact_coronavirus_20q2.pdf (Accessed March 15, 2024).

Vartivarian, S. and Little, R.J.A. (2002). Does Weighting for Nonresponse Increase the Variance of Survey Means?. Survey Methodology, 31, 161-168. Available at: https://www150.statcan.gc.ca/n1/pub/12-001-x/2005002/article/9046-eng.pdf (accessed March 15, 2024).

Wagner, D., and Layne, M. (2014). "The Person Identification Validation System (PVS): Applying the Center for Administrative Records and Research and Applications' record linkage software." U.S. Census

Bureau CARRA Report Series #2014-01. Available at: https://www.census.gov/library/working-papers/2014/adrm/carra-wp-2014-01.html. (accessed March 15, 2024).

Demographic Totals of the Production Weight

Table A1: Demographic Totals of the production Weight

Number of Totals	Description of Totals
1	U.S. households
14	U.S. population for each of 14 age x 2 race (black/nonblack) categories
9	U.S. Population by Census Division
9	U.S. Urban Populations by Census Division
1	U.S. Hispanic population
1	U.S. homeowner households
35	Number of Totals

Table A2: Age Categories

14-24
25-34
35-44
45-54
55-64
65-74
75+

Table A3: Census Divisions

New England
Middle Atlantic
East North Central
West North Central
South Atlantic
West South Central
Mountain
Pacific

Census Tracts are small, relatively permanent statistical subdivisions of a county or statistically equivalent entity that can be updated by local participants prior to each decennial census as part of the Census Bureau’s Participant Statistical Areas Program (PSAP). The Census Bureau delineates census tracts in situations where no local participant responded or where state, local, or tribal governments declined to participate. The primary purpose of census tracts is to provide a stable set of geographic units for the presentation of statistical data.

Census tracts generally have a population size between 1,200 and 8,000 people, with an optimum size of 4,000 people. A census tract usually covers a contiguous area; however, the spatial size of census tracts varies widely depending on the density of settlement. Census tract boundaries are delineated with the intention of being maintained over a long time so that statistical comparisons can be made from census to census. Census tracts occasionally are split due to population growth or merged as a result of substantial population decline.

Census tract boundaries generally follow visible and identifiable features. They may follow nonvisible legal boundaries, such as minor civil division (MCD) or incorporated place boundaries in some states and situations, to allow for census tract-to-governmental unit relationships where the governmental boundaries tend to remain unchanged between censuses. State and county boundaries always are census tract boundaries in the standard census geographic hierarchy. Tribal census tracts are a unique geographic entity defined within federally recognized American Indian reservations and off-reservation trust lands and can cross state and county boundaries. The tribal census tracts may be completely different from the standard county-based census tracts defined for the same area. (see “Tribal Census Tract”).

Technical Description of the Entropy Weights

This appendix describes the methodology of the research weights derived for the Consumer Expenditure Survey (CE). The methodology uses an approach that is similar to the calibration that was applied to CPS by Rothbaum and Bee (2021), which is based on Hainmueller (2012) who applied it to observational studies. Note that the general calibration method applied here is also described by Deville and Särndal (1992) for surveys. The methodology makes use of the administrative data available to the U.S. Census Bureau in order to reduce bias in the estimates due to nonresponse.

Notation

i	index on Consumer Units (CUs)
k	index on persons within the CUs
CU_i	the set of persons in CU *i*
r_{CUs}	sample CU interviews
nr_{CUs}	sample CU non-interviews
s_{CUs}	sample CU interviews and non-interviews ($s_{CUs} = r_{CUs} \cup nr_{CUs}$) – excludes ineligible sample units
$r_{persons}$	sample persons that are in r_{CUs}
d_i	design weight for CU *i* – base weight and any other subsampling factors
q_i	normalized design weight for CU *i*
$\mathbf{x}_i$	vector of non-interview adjustment variables for the CU *i*.
$\mathbf{z}_k$	vector of non-interview adjustment variables for person k. The values are all 0/1.

The values of $\mathbf{x}_i$ and $\mathbf{z}_k$ are all 0/1, where 1 indicates they have the characteristic and 0 indicates they do not have the characteristic.

The methodology calculates a final CU weight for all respondents in three steps. Each step calculates a different weight, which are listed below.

$w_i^{(1)}$	calibrated weight adjusted for nonrespondents for CU *i*
$w_k^{(2)}$	calibrated weight for person *k* of CU *i*
$w_i^{(3)}$	final weight for CU *i* – sums to 1.0
$w_i^{(4)}$	final weight for CU *i*– sums to total number of CUs

Step 1: Apply a Non-interview Adjustment to Consumer Unit Weights

This step adjusts the sample interviews for non-interviews by calculating new weights that have three goals:

(1) The new weight $w_i^{(1)}$ is "close" or changed as little as possible from $q_i$the normalized base weight **.**

(1A) Using variables *x* that are known for both interviews and non-interviews, the mean of *x* using the new weight $w_i^{(1)}$ across the interviews r_{CUs} is equal to the mean of *x* using the normalized base weights d_i across the interviews and non-interviews s_{CUs}.

(1B) The sum of the weights for interviews is equal to one.

The three goals can be mathematically stated as, find the weights $w_i^{(1)}$ that minimize

$$\sum_{i \in r_{CUs}} w_i^{(1)} log\left(\frac{w_i^{(1)}}{q_i}\right) \qquad (1)$$

subject to the constraints

$$\sum_{i \in r_{CUs}} w_i^{(1)} \mathbf{x}_i = \sum_{i \in s_{CUs}} q_i \mathbf{x}_i \qquad (1A)$$

and

$$\sum_{i \in r_{CUs}} w_i^{(1)} = 1 \qquad (1B)$$

where q_i is the normalized base weight d_i, i.e.,

$$q_i = \frac{d_i}{\sum_{i \in s_{CUs}} d_i}.$$

Table 1 lists of variables used in the calibration of Step 1.

Step 2: Calibrate the Person Weights to Known Totals

This step calibrates the person weights $w_k^{(2)}$ to known demographic totals. First, it assigns each person in the CU the non-interview adjusted weight $w_i^{(1)}$ from Step 1. Next, it applies calibration using two sets of totals.

This step adjusts the sample interviews for non-interviews by calculating new weights that have four goals:

(2) The new weight $w_k^{(2)}$ is "close" or changed as little as possible from $w_k^{(1)}$. – the CU weight from Step 1.

(2A) The mean of *x* using the new weight $w_i^{(2)}$ across the interviews $r_{persons}$ is equal to the mean of *x* using the normalized base weights $w_i^{(1)}$ across the interview CUs r_{CUs}.

(2B) Using variables *z* that are known for both interviews and correspond to known demographic totals Z, the mean of *z* using the new weight $w_i^{(2)}$ across the interview persons $r_{persons}$ is equal to the mean of *x* using the known total *Z*

(2C) The sum of the weights for persons equal to one.

The four goals can be mathematically stated as, find the person weights $w_k^{(2)}$ that minimize

$$\sum_{k \in r_{persons}} w_k^{(2)} log \left(\frac{w_k^{(2)}}{w_i^{(1)}} \right) \tag{2}$$

subject to the constraints of Table B1:

Table B1: Person Calibration Constraints for Step 2

Goal	Description of Constraints	Calibration Constraints
(2A)	Preserve distribution of housing unit characteristics	$\sum_{k \in r_{persons}} w_k^{(2)} \tilde{\mathbf{x}}_k = \sum_{i \in s_{CUs}} w_i^{(1)} \mathbf{x}_i$
(2B)	External Population targets	$\sum_{k \in r_{persons}} w_k^{(2)} \mathbf{z}_k = \bar{\mathbf{Z}}$
(2C)	Sum of the weights is equal to 1.0.	$\sum_{i \in r_{persons}} w_i^{(2)} = 1$

The CE weights do not include a set of constraints for "spousal equivalence". Ensuring that the two spouses of a household (HH) have the same HH weight is not needed for CE as it is with CPS in Rothbaum and Bee (2021).

For the (2A) constraints, the person-level variables $\tilde{\mathbf{x}}_k$ are the value of the CU $\mathbf{x}_k$ divided by the number of persons in the CU and we can express as:

$$\tilde{\mathbf{x}}_k = \left(1/\sum_{k\in CU_i} 1\right)\#\mathbf{x}_i.$$

Defining $\tilde{\mathbf{x}}_k$ in this way gives each person of the CU an equal share of each of the CU characteristics $\mathbf{x}_i$. The aim of using the variables of 2A within this step is to preserve the effects of Step 1.

The aim of using the calibration variables of 2C is to ensure that the weights are consistent with the Census population estimates ("POP totals") **Z**.

Step 3: Calculate the Final CU Weight

The final CU weight is calculated as the average of the person weights for CU *i*, which we can express as:

$$w_i^{(3)} = \sum_{k\in CU_i} w_i^{(2)} / \sum_{k\in CU_i} 1$$

We applied the next step to make the final weight consistent with the CE production weight. Given that the sum of the CE production weights sum to the number of CUs in the U.S., we adjusted $w_i^{(3)}$ so it also summed to the number of CUs in the U.S. This was necessary since the sum of the person weights $w_i^{(2)}$ is equal to 1.0,

$$w_i^{(4)} = \left(C/\sum_{i\in r_{CUs}} w_i^{(3)}\right) w_i^{(3)}$$

where *C* is the total number of CUs in the U.S.

Next, we list the variables used in the entropy balance weighting (all of which are address-level summaries):

- Linkage indicators – is a household member linked to each data set
- Number of linked individuals at the address
- Presence of a household member by:
    - Race and Hispanic origin
    - Citizenship and foreign-born status
    - Gender
    - Age (binned)
- 1040 Tax filing information
    - Indicators for various schedules (such as Schedule C: Profit or Loss from Business)
    - Marital status
    - Total income (continuous measures and binned)
- W-2
    - Earnings in the survey year (continuous measures and binned)
    - Change in earnings indicators (earnings declined by 50 percent or more, declined by 10-50 percent, etc. in arc percent changes)
- 1099-R
    - Income (sum of DC withdrawals and DB payments, continuous and binned)

- Race and Hispanic-origin indicators are interacted with many of the variables from 1040 tax filings, W-2s and 1099-R's to capture possible differential selection into nonresponse by income by race

All continuous income measures are transformed using inverse hyperbolic sine, which is nearly perfectly correlated with the natural log (in their common support) but is defined at zero and negative values.

The external population targets used in the entropy balance weights come from the CPS ASEC. Because the CPS ASEC is weighted to external population controls from the U.S. Census Bureau's Population Estimates Program,[8] they are indirectly estimated using the same external population controls. We matched the individual data in the CE to region x race and Hispanic-origin x age cells estimated from the CPS ASEC.

[8] See https://www.census.gov/programs-surveys/popest/about.html for more information on the PEP estimates.

Table C1: Match Rates of Census Administrative Records

		2018			2019			2020		
		Match Rate	SE of Match Rate	CV of Match Rate	Match Rate	SE of Match Rate	CV of Match Rate	Match Rate	SE of Match Rate	CV of Match Rate
1-1099 IRMF	0-Sample	87.09	0.59	0.67	87.48	0.65	0.74	87.77	0.55	0.62
	1-Interviews	87.27	0.69	0.79	87.88	0.74	0.84	89.66	0.53	0.60
	2-Non-interview	86.85	0.58	0.66	87.02	0.67	0.78	86.07	0.63	0.73
2-W2	0-Sample	63.10	0.57	0.90	62.81	0.63	1.00	62.40	0.55	0.88
	1-Interviews	63.08	0.71	1.13	62.31	0.75	1.20	62.73	0.63	1.00
	2-Non-interview	63.13	0.66	1.04	63.41	0.70	1.11	62.10	0.69	1.11
3-1099R	0-Sample	43.15	0.57	1.32	43.74	0.58	1.34	42.05	0.56	1.33
	1-Interviews	44.51	0.72	1.61	45.24	0.64	1.42	44.79	0.80	1.78
	2-Non-interview	41.31	0.64	1.54	41.98	0.70	1.67	39.57	0.63	1.59
4-1040	0-Sample	79.75	0.60	0.75	79.64	0.69	0.87	79.91	0.59	0.74
	1-Interviews	80.01	0.73	0.91	80.03	0.91	1.13	81.60	0.62	0.76
	2-Non-interview	79.40	0.59	0.74	79.18	0.69	0.87	78.38	0.70	0.90
5-2010 Census	0-Sample	85.97	0.58	0.68	86.27	0.66	0.77	86.80	0.55	0.63
	1-Interviews	86.06	0.69	0.80	86.59	0.76	0.88	88.65	0.53	0.60
	2-Non-interview	85.84	0.58	0.68	85.88	0.69	0.81	85.13	0.64	0.75
6-ACS	0-Sample	21.10	0.51	2.40	21.74	0.44	2.02	21.41	0.44	2.04
	1-Interviews	22.94	0.54	2.34	23.20	0.53	2.26	24.93	0.64	2.58
	2-Non-interview	18.61	0.70	3.75	20.02	0.56	2.78	18.23	0.48	2.66

Table C2: 2018 to 2021 Expenditure Estimates by Census Region and Alternative Nonresponse Adjustment

Year		Method	U.S.	Census Region			
				Northeast	Midwest	South	West
2018	Base Weight	Production	56,560 (802)	61,100 (2,767)	54,520 (1,432)	52,060 (998)	62,760 (1,774)
	Nonresponse Weight	Production	56,490 (800)	60,830 (2,683)	54,000 (1,425)	52,390 (997)	62,590 (1,783)
		Response	56,670 (756)	61,330 (2,655)	54,920 (1,407)	52,260 (970)	62,260 (1,817)
		Expenditures	56,610 (750)	61,130 (2,627)	54,710 (1,365)	52,300 (994)	62,270 (1,810)
		Entropy	57,310 (931)	62,040 (3,006)	55,530 (1,496)	52,870 (1,208)	62,910 (1,691)
	Final Weight	Production	57,610 (783)	62,360 (2,893)	54,590 (1,388)	53,520 (1,069)	63,840 (1,724)
		Response	57,770 (772)	62,890 (2,921)	55,510 (1,430)	53,330 (1,026)	63,550 (1,716)
		Expenditures	57,700 (762)	62,720 (2,862)	55,290 (1,362)	53,340 (1,062)	63,570 (1,725)
		Entropy	57,060 (964)	62,400 (3,135)	53,560 (1,429)	53,190 (1,232)	62,970 (1,750)
2019	Base Weight	Production	57,580 (702)	61,590 (2,452)	55,220 (832)	53,530 (1,166)	63,850 (1,275)
	Nonresponse Weight	Production	57,380 (705)	61,560 (2,411)	54,620 (830)	53,490 (1,170)	63,490 (1,244)
		Response	57,620 (636)	61,160 (2,146)	55,240 (858)	53,830 (1,061)	63,610 (1,349)
		Expenditures	57,370 (627)	61,100 (2,159)	54,970 (805)	53,540 (1,091)	63,300 (1,293)
		Entropy	58,540 (780)	62,560 (2,421)	56,360 (922)	54,470 (1,210)	64,400 (1,329)
	Final Weight	Production	58,620 (572)	62,980 (1,962)	55,390 (732)	54,740 (1,051)	64,960 (1,290)
		Response	59,000 (569)	62,670 (2,103)	56,170 (868)	55,110 (998)	65,480 (1,348)
		Expenditures	58,750 (544)	62,690 (2,059)	55,890 (798)	54,840 (1,030)	65,050 (1,253)
		Entropy	58,360 (828)	62,640 (2,316)	54,060 (1,043)	54,750 (1,199)	65,400 (1,445)
2020	Base Weight	Production	57,740 (740)	60,350 (1,804)	55,200 (1,107)	54,010 (1,132)	64,050 (1,814)
	Nonresponse Weight	Production	57,200 (718)	59,590 (1,787)	54,560 (1,053)	53,860 (1,128)	63,540 (1,725)
		Response	56,900 (730)	59,750 (1,843)	54,480 (1,146)	53,300 (1,129)	63,180 (1,967)
		Expenditures	56,440 (703)	58,860 (1,918)	54,240 (1,103)	53,210 (1,031)	62,130 (1,802)
		Entropy	57,260 (825)	60,640 (2,013)	55,200 (1,392)	53,780 (1,140)	62,510 (1,721)
	Final Weight	Production	57,920 (646)	60,490 (1,679)	55,150 (1,112)	54,580 (1,083)	64,110 (1,698)
		Response	58,300 (686)	61,490 (1,814)	55,710 (1,185)	54,520 (1,118)	64,640 (1,970)
		Expenditures	57,760 (681)	60,610 (1,956)	55,360 (1,190)	54,440 (1,051)	63,350 (1,754)
		Entropy	57,120 (867)	59,750 (1,872)	53,490 (1,526)	53,680 (1,078)	64,480 (1,806)

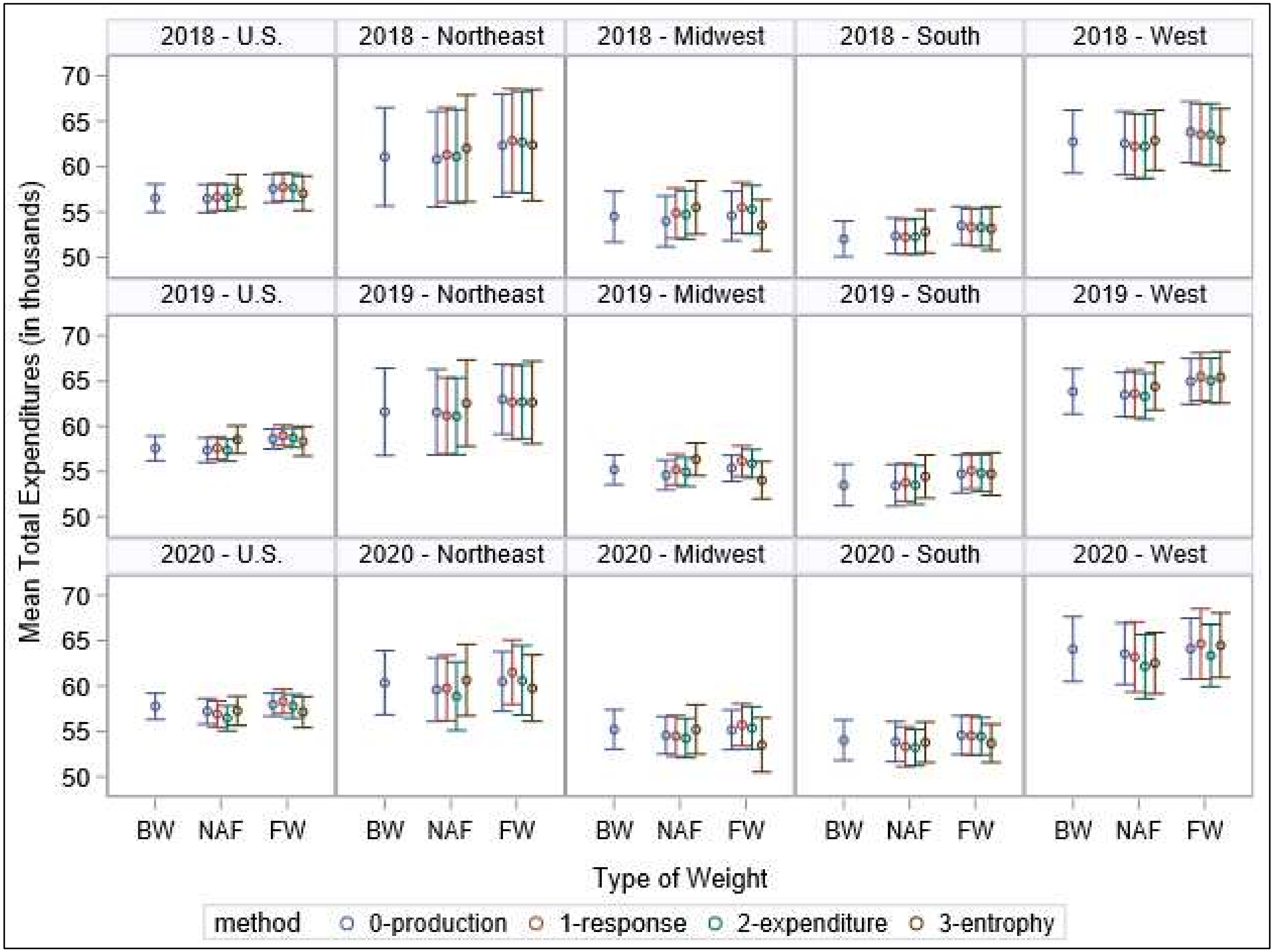

Figure A1: National and Regional Total Expenditures – 95 Percent Confidence Intervals

The labels in Figure A1 are BW for base weight, NAF for non-interview adjustment factor, and FW for final weight.

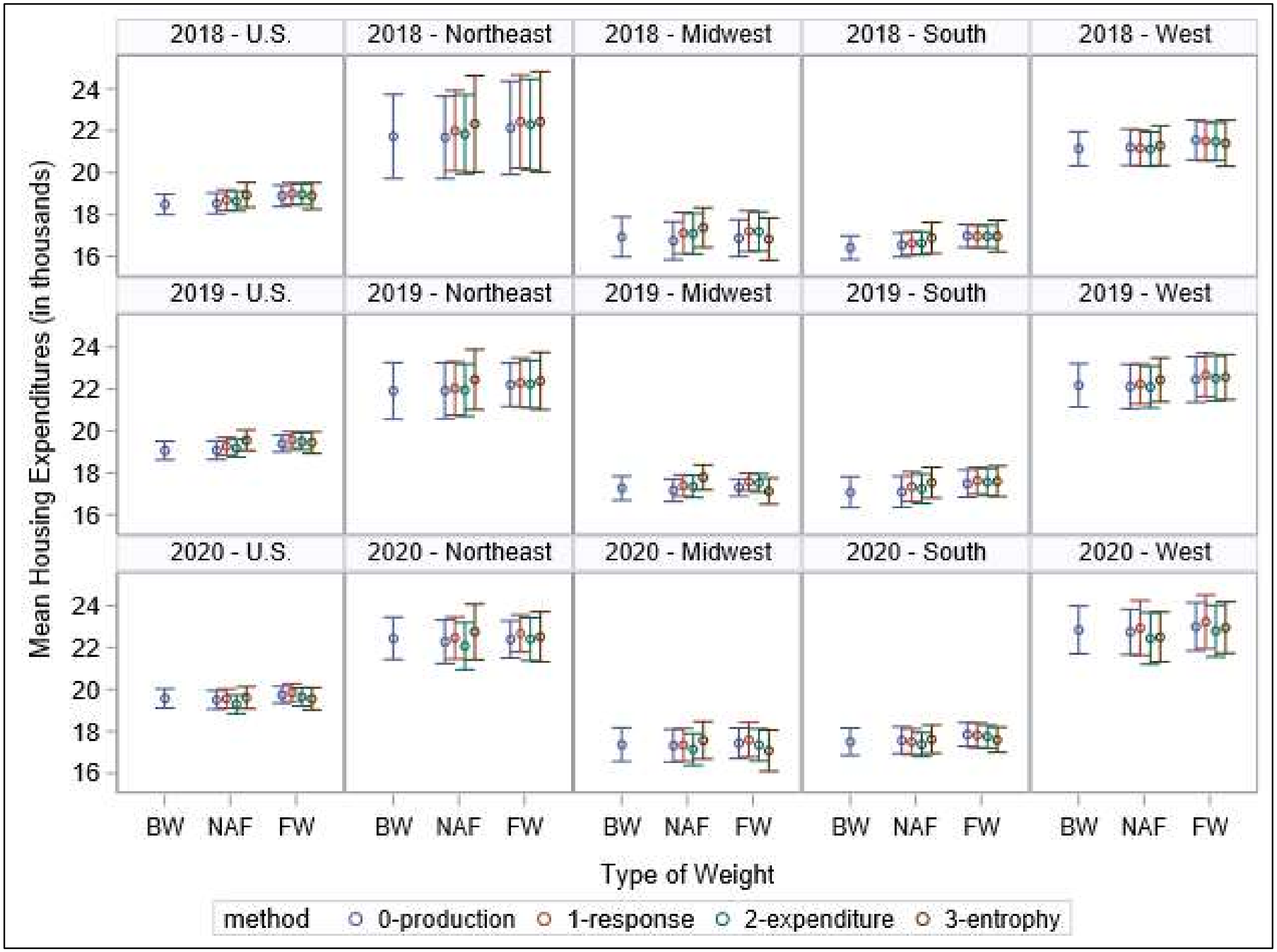

Figure A2. National and Regional Mean Housing Expenditures – 95 Percent Confidence Intervals

The labels in Figure A2 are BW for base weight, NAF for non-interview adjustment factor, and FW for final weight.

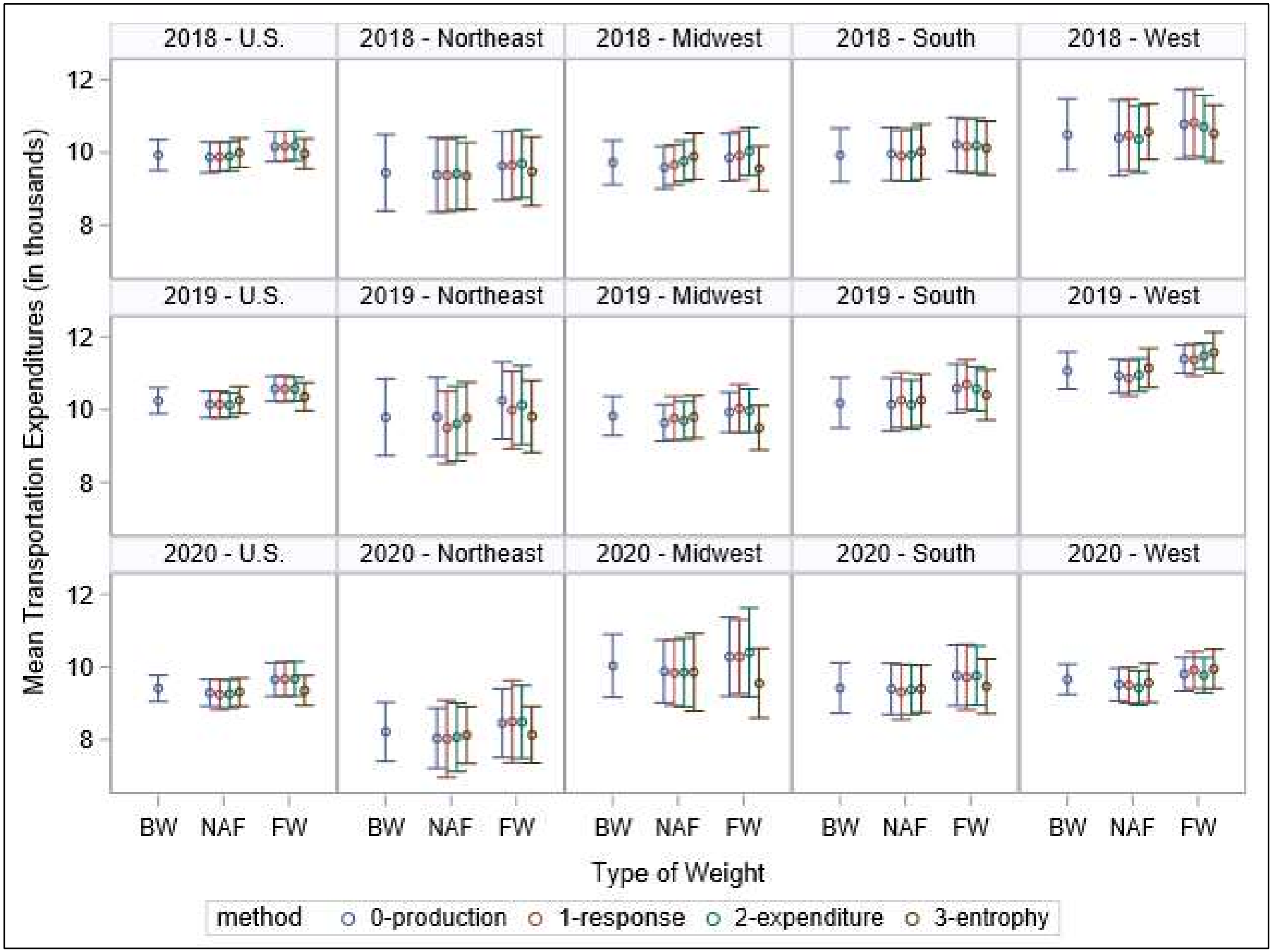

FIGURE A3: NATIONAL AND REGIONAL MEAN TRANSPORTATION EXPENDITURES – 95 PERCENT CONFIDENCE INTERVALS

The labels in Figure A3 are BW for base weight, NAF for non-interview adjustment factor, and FW for final weight.

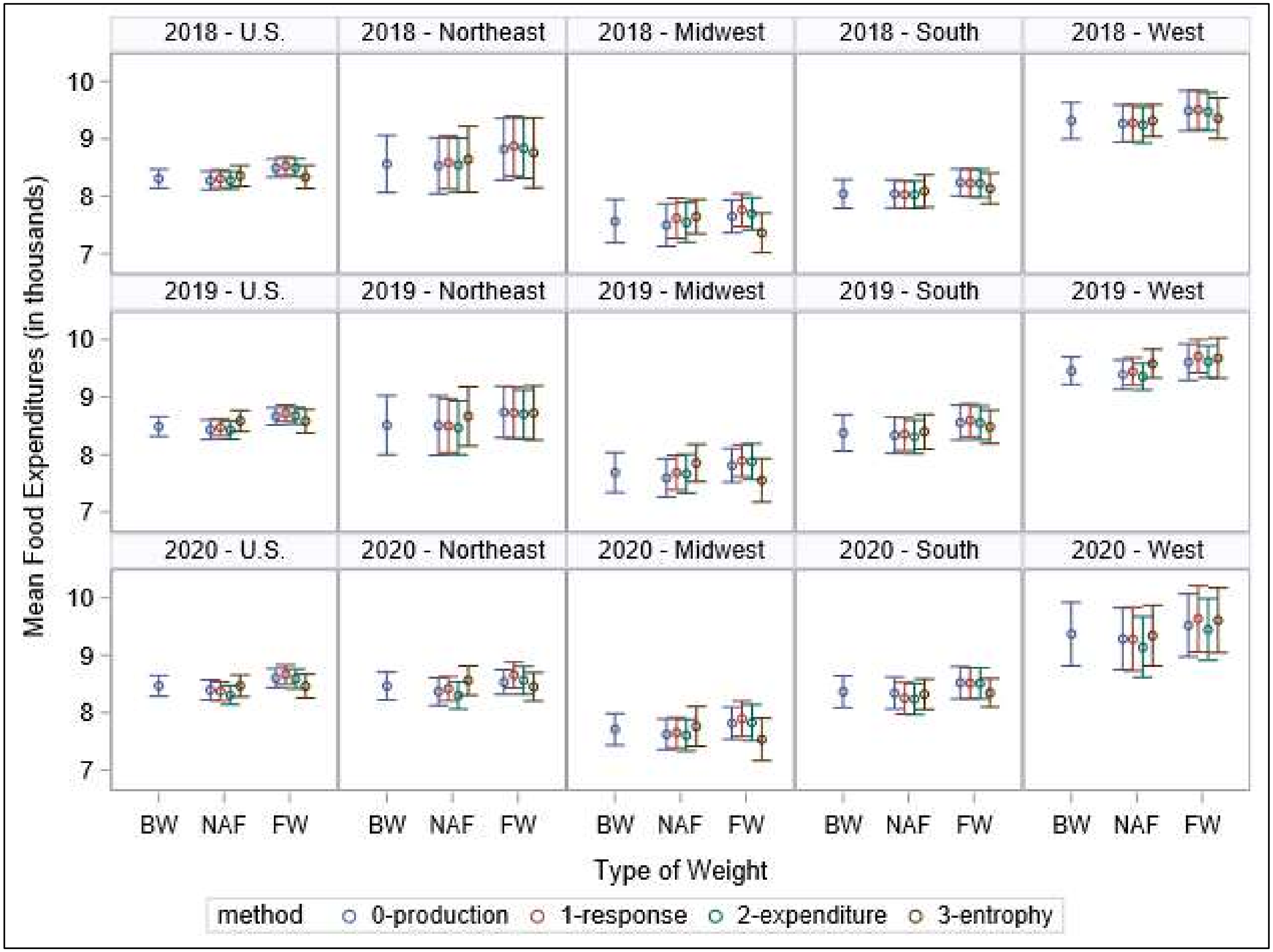

Figure A4: National Mean Food Expenditures – 95 percent Confidence Intervals

The labels in Figure A4 are BW for base weight, NAF for non-interview adjustment factor, and FW for final weight.

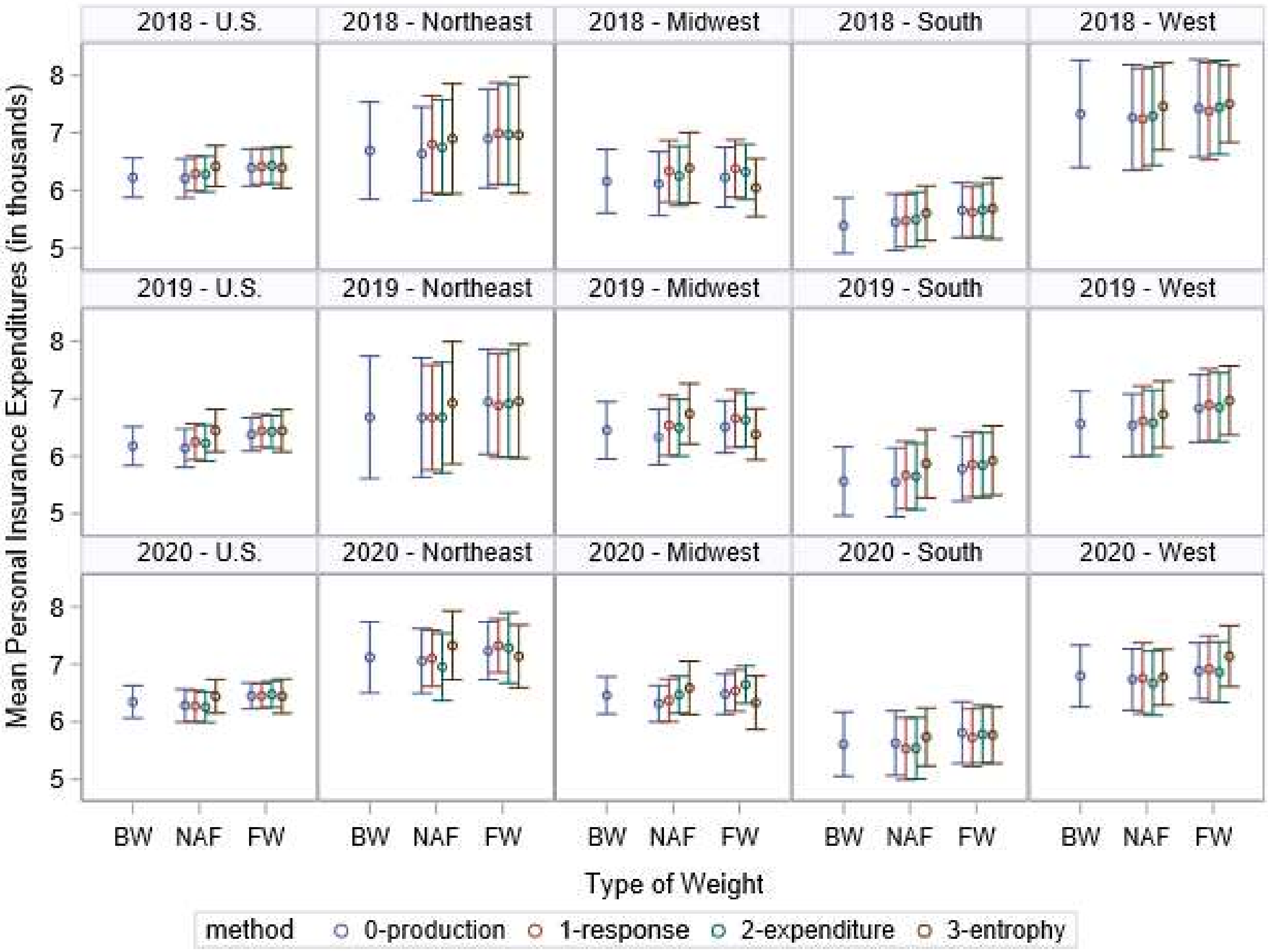

Figure A5: National Mean Personal Insurance Expenditures – 95 Percent Confidence Intervals

The labels in Figure A5 are BW for base weight, NAF for non-interview adjustment factor, and FW for final weight.

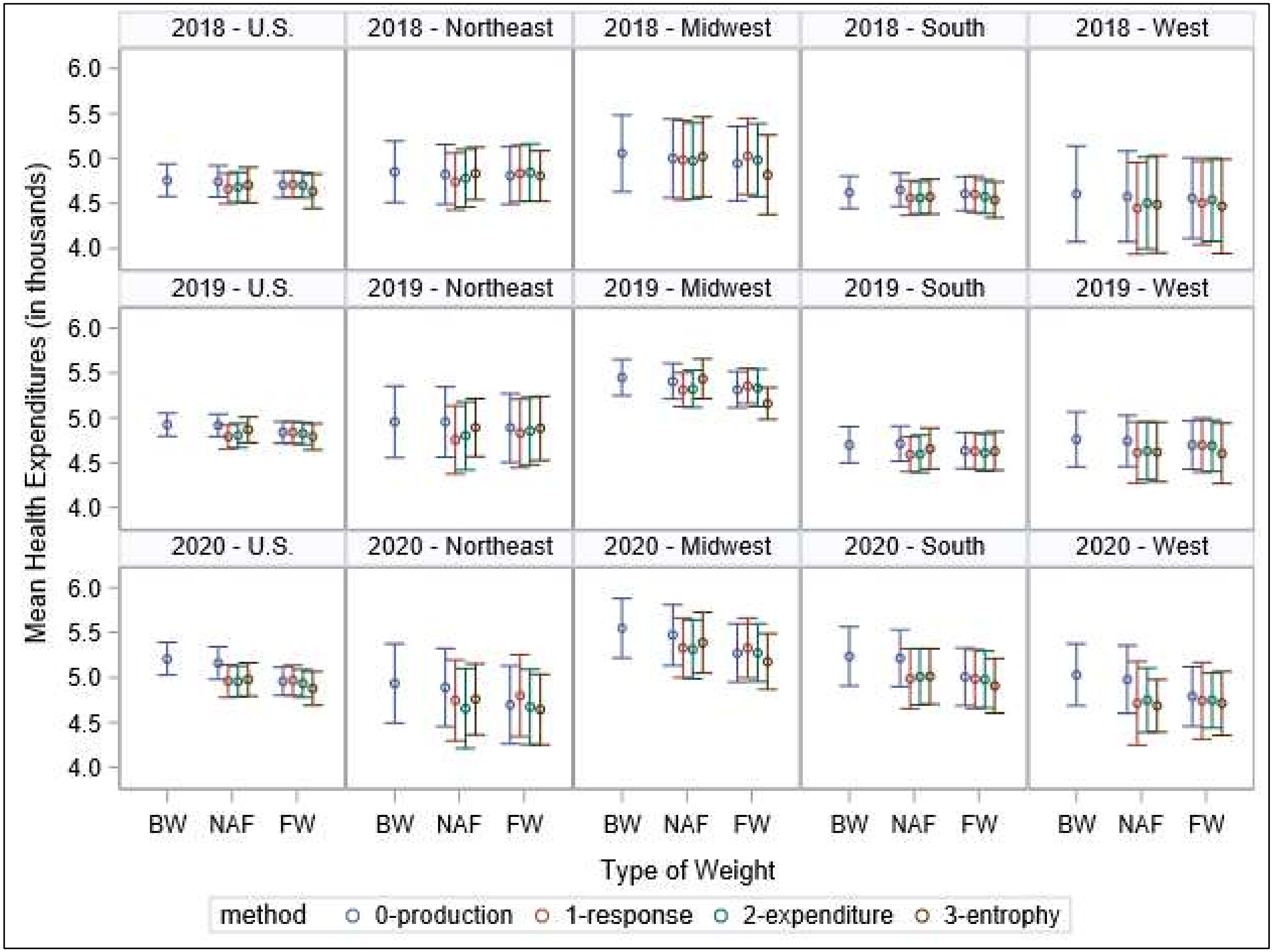

Figure A6: National Mean Health Expenditures – 95 Percent Confidence Intervals

The labels in Figure A6 are BW for base weight, NAF for non-interview adjustment factor, and FW for final weight.