The CE Source Selection Process

Brett J. Creech
Branch of Information and Analysis
2019 Microdata Users Workshop
July 18, 2019

Overview

- Purpose
- Background
- Methodology
- Decision Criteria
- Next steps

What is Source Selection

- Methods used to select the appropriate survey for publication table estimates.
  - Interview
  - Diary
- For items that are unique to one survey or another, the choice is obvious.
- For items that overlap in coverage between the surveys, source selection methods are used to determine which source to select for publishing the Integrated data.

Purpose

- Primarily used for publication tables.
  - Identifies the more reliable source of survey data to use in estimation.

- PUMD data users
  - Provides a means for users to integrate survey estimates and closely replicate the publication tables.

Background

- The previous Source Selection method was developed in 1997 using data from 1993-1995. This method relied primarily on the Coefficient of Variance (CV) and in some cases the Mean Squared Error (MSE).
- In 2001, meetings were held with CE and CPI to look at differences in source selection using 1999 data. It was recommended that CPI adopt the CE source decision in all cases with greater than 50 reports of expenditures at the UCC level.

Background

- In 2006, when incorporating a few new UCCs with 2005 data, source selection was coordinated so that CE and CPI were in agreement on the newly introduced UCCs.

- In 2007, CE and CPI formed a team to evaluate and come up with a new methodology for Source Selection to be used for 2007 publication.

Methodology – Overall Goal

- Over 200 UCCs are processed using both the Diary and the Interview.
- A determination is made to which source is used for the integrated tables.
- The overall methodology selects the higher mean given two decision criteria with exceptions from the CPI.

Methodology

- Preliminary steps:
  - Calculating counts, sample means, and sample variances.
  - Data are top coded and bottom coded
    - This is done to minimize the impact of outliers

Methodology

- The counts (representing a reported expenditure for that UCC) and Z-scores are weighted for the three most recent collection years:
  - 1st collection year by 1/6 (For the 2017 data, use 2014)
  - 2nd collection year by 2/6 (For the 2017 data, use 2015)
  - 3rd collection year by 3/6 (For the 2017 data, use 2016)

Methodology

- If a new UCC was created in the past 2 years (for example, a new UCC created in 2015), then the following weights are used:
  - 1st collection year by 2/5 (For 2017, use 2015 data)
  - 2nd collection year by 3/5 (For 2017, use 2016 data)

Decision Criteria

- There are two criterions that are used in determining source selection decisions:
  - Criterion 1: Counts Sufficiency
  - Criterion 2: Statistical Significance

Criterion 1: Counts Sufficiency

- For each UCC and each survey (Interview or Diary), the number of consumer units with at least one expenditure is counted for each of the 3 most recent data years.
  - Yields 6 counts for each UCC
    - Three yearly counts for Interview
    - Three yearly counts for Diary

Criterion 1: Counts Sufficiency

- A sufficient amount of data exists when the count for each of the 3 years is greater than or equal to 60.
- If both surveys have sufficient data then proceed to the next Criterion.
- If both surveys lack sufficient data, then keep existing source.

Criterion 1: Counts Sufficiency

- If one survey has sufficient data, but the other has insufficient data, then a weighted average of the three yearly counts for the survey having an insufficient amount of data is computed: $n^* = (3/6)n_{t-1} + (2/6)n_{t-2} + (1/6)n_{t-3}$

Criterion 1: Counts Sufficiency

- If the weighted average n* from the insufficient survey is greater than or equal to 60, then proceed to the next Criterion.

- If the weighted average n* from the insufficient survey is still less than 60, then use the survey with sufficient data as the source.

Criterion 2: Statistical Significance (Z-scores)

- If the value of the weighted Z-Score, $z^* = (3/6)z_{t-1} + (2/6)z_{t-2} + (1/6)z_{t-3}$, is greater than or equal to 1.645, or less than or equal to -1.645 then select the source based on the following:
  - Greater than or equal to 1.645 – Interview Survey
  - Less than or equal to -1.645 – Diary Survey

Criterion 2: Statistical Significance (Z-scores)

- If the weighted Z-Score is between –1.000 and 1.000, then the current source will continue to be used.

Criterion 2: Statistical Significance (Z-scores)

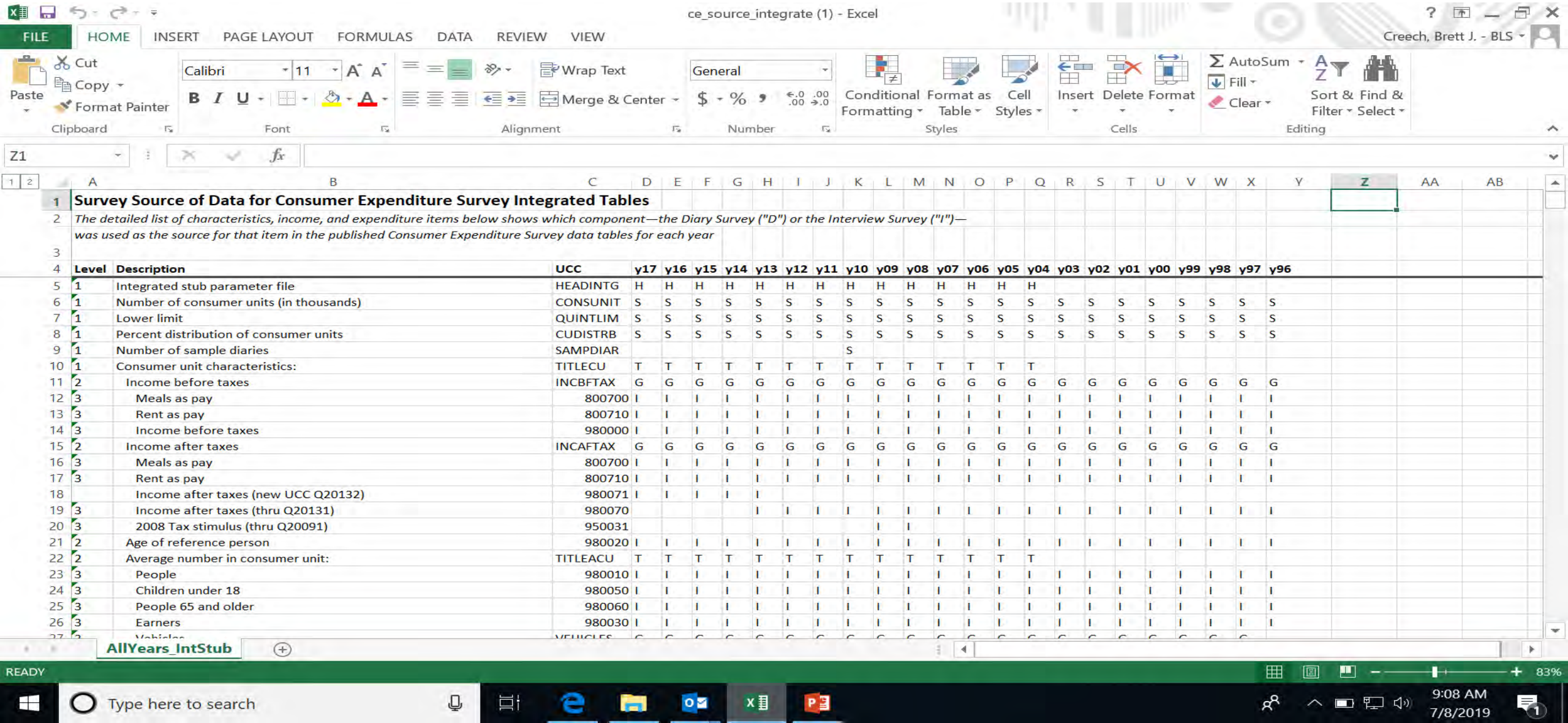
- If all three z-scores are 1.000 and above, then use the Interview Survey
- If all three z-scores are -1.000 and lower, then use the Diary Survey
- Any remaining scenarios, the source remains the same.

Exclusions – Items stay in the Interview Survey

- Expenditures for items net of reimbursements
  - Medical Care
  - Auto Repairs
- Reimbursements are captured in the Interview survey
  - Not captured in the Diary survey
- Transportation UCCs
  - Trade-in vehicle values are deducted from purchases of new cars in out-of-pocket expense calculations

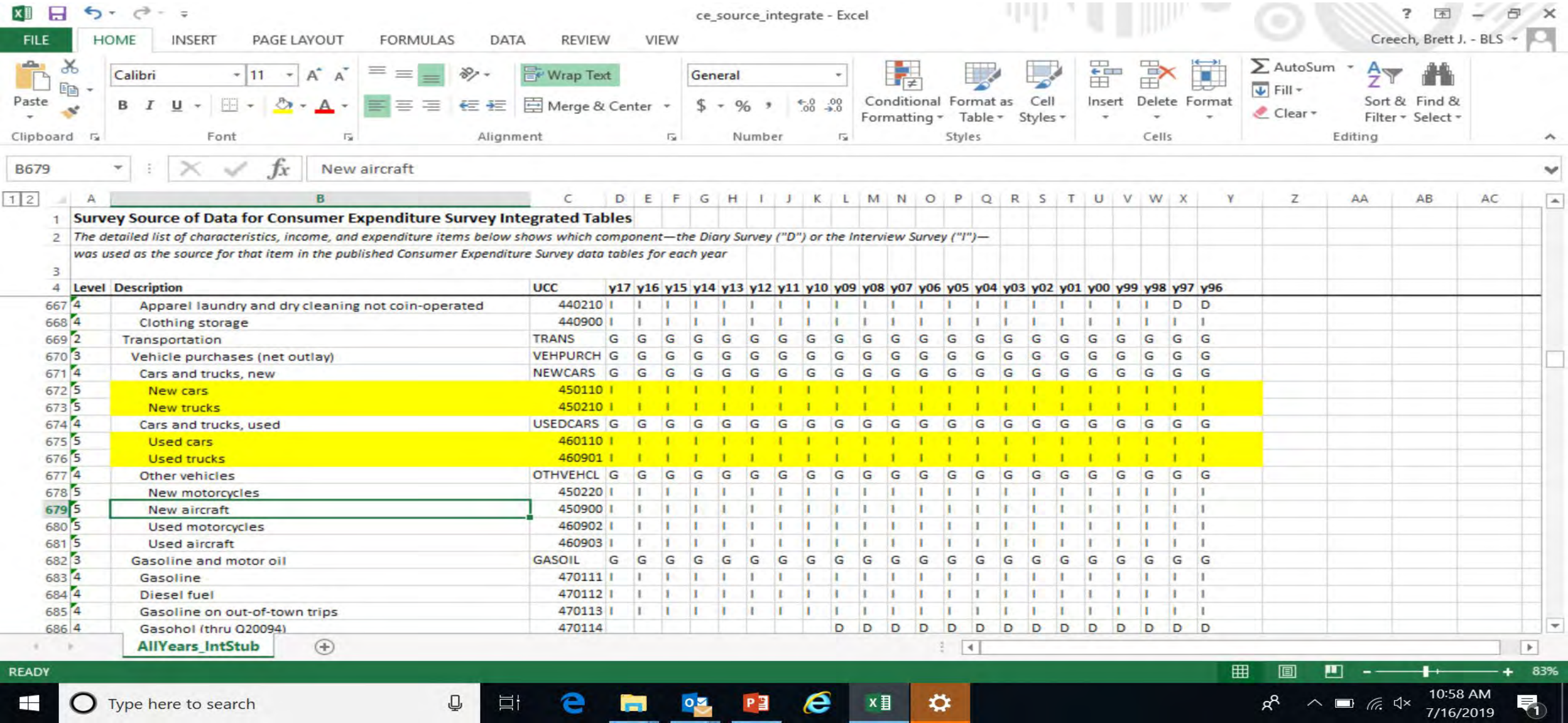
Where to find the Source Selection spreadsheet

Spreadsheet 1996-2017



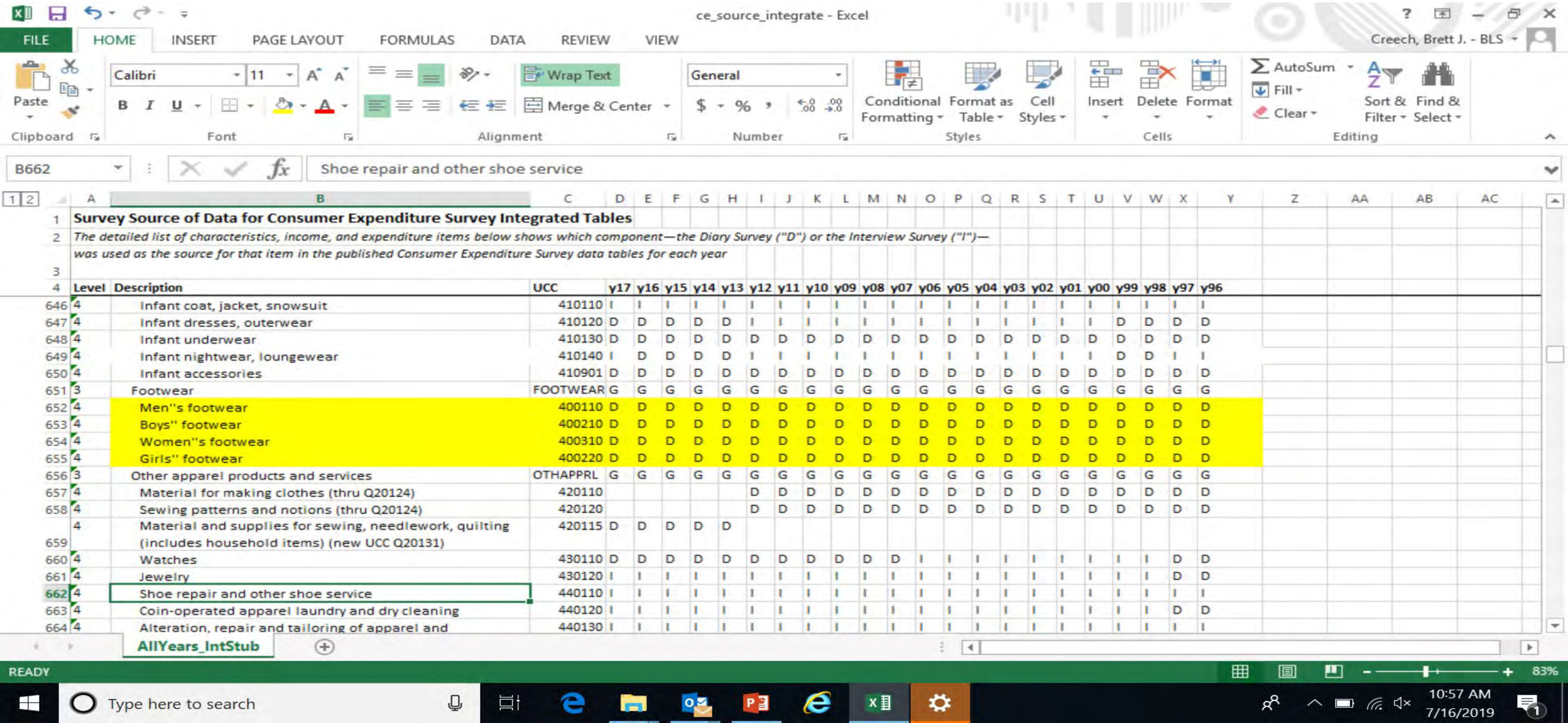
Level	Description	UCC	y17	y16	y15	y14	y13	y12	y11	y10	y09	y08	y07	y06	y05	y04	y03	y02	y01	y00	y99	y98	y97	y96
1	Integrated stub parameter file	HEADINTG	H	H	H	H	H	H	H	H	H	H	H	H	H	H								
1	Number of consumer units (in thousands)	CONSUNIT	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
1	Lower limit	QUINTLIM	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
1	Percent distribution of consumer units	CUDISTRB	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S	S
1	Number of sample diaries	SAMPDIAR								S														
1	Consumer unit characteristics:	TITLECU	T	T	T	T	T	T	T	T	T	T	T	T	T	T								
2	Income before taxes	INCBFTAX	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
3	Meals as pay	800700	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
3	Rent as pay	800710	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
3	Income before taxes	980000	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
2	Income after taxes	INCAFTAX	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
3	Meals as pay	800700	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
3	Rent as pay	800710	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
	Income after taxes (new UCC Q20132)	980071	I	I	I	I	I																	
3	Income after taxes (thru Q20131)	980070					I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
3	2008 Tax stimulus (thru Q20091)	950031									I	I												
2	Age of reference person	980020	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
2	Average number in consumer unit:	TITLEACU	T	T	T	T	T	T	T	T	T	T	T	T	T	T								
3	People	980010	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
3	Children under 18	980050	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
3	People 65 and older	980060	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
3	Earners	980030	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I

Interview Example

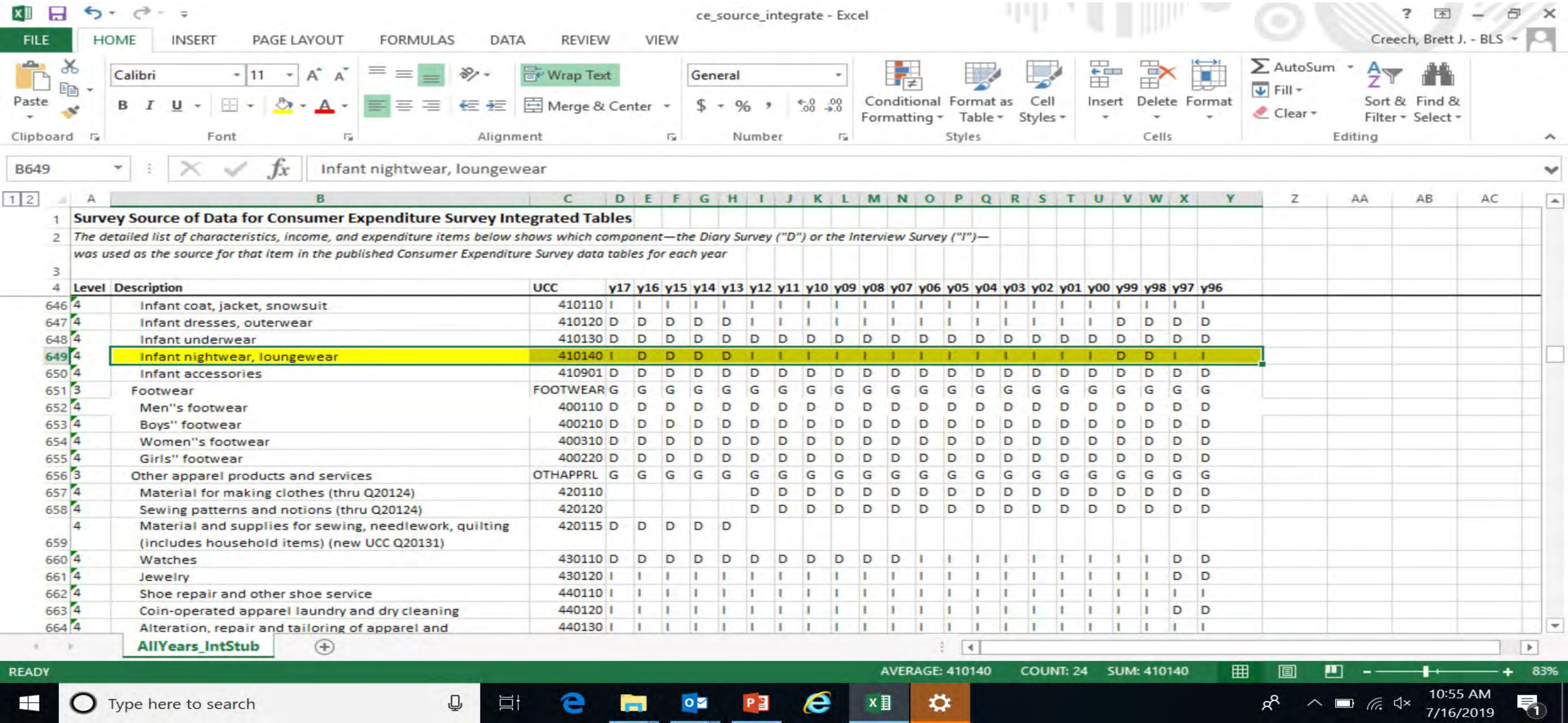


	Level	Description	UCC	y17	y16	y15	y14	y13	y12	y11	y10	y09	y08	y07	y06	y05	y04	y03	y02	y01	y00	y99	y98	y97	y96
667	4	Apparel laundry and dry cleaning not coin-operated	440210	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	D	D
668	4	Clothing storage	440900	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
669	2	Transportation	TRANS	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
670	3	Vehicle purchases (net outlay)	VEHPURCH	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
671	4	Cars and trucks, new	NEWCARS	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
672	5	New cars	450110	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
673	5	New trucks	450210	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
674	4	Cars and trucks, used	USEDCARS	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
675	5	Used cars	460110	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
676	5	Used trucks	460901	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
677	4	Other vehicles	OTHVEHCL	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
678	5	New motorcycles	450220	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
679	5	New aircraft	450900	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
680	5	Used motorcycles	460902	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
681	5	Used aircraft	460903	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
682	3	Gasoline and motor oil	GASOIL	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G
683	4	Gasoline	470111	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
684	4	Diesel fuel	470112	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
685	4	Gasoline on out-of-town trips	470113	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I
686	4	Gasohol (thru Q20094)	470114									D	D	D	D	D	D	D	D	D	D	D	D	D	D

Diary Example



Source Selection Change



Reference

Brett Creech and Barry Steinberg: CE Source Selection for Publication Tables

https://www.bls.gov/cex/anthology11/csxanth3.pdf

What's next?

- Team is being formed to revisit the current Source Selection methodology
- 2019 data: Use current methodology while testing new approach
- 2021 data: Potential new methodology implemented

Contact Information

Brett J. Creech
(202) 691-5120
Creech.Brett@bls.gov