

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 criteria 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

The screenshot shows a web browser window with the URL https://www.bls.gov/cex/pumd_doc.htm. The page is titled "Consumer Expenditure Surveys" and is part of the "PUMD Documentation" section. The header includes the "UNITED STATES DEPARTMENT OF LABOR" and "BUREAU OF LABOR STATISTICS" logos, along with navigation links like "A to Z Index", "FAQs", "About BLS", "Contact Us", and "Subscribe to E-mail Updates". A search bar is also present.

The main content area is divided into two columns. The left column contains a "BROWSE CE" menu with links to "CE HOME", "CE OVERVIEW", "CE NEWS RELEASES", "CE PUBLICATIONS", "CE TABLES", "CE LABSTAT DATABASE", "CE PUBLIC USE MICRODATA", "CE WORKSHOP AND SYMPOSIUM", "CE GEOGRAPHIC DATA", "CE EXPERIMENTAL RESEARCH PRODUCTS", and "CONTACT CE". Below this is a "SEARCH CE" box and a "CE TOPICS" section with links to "INFORMATION FOR CE RESPONDENTS".

The right column is titled "PUMD Documentation" and contains the following text: "This page contains documentation for the public-use microdata (PUMD) for years starting in 1996. Documentation for years prior to 1996 are available USB flash drive for [purchase \(PDF\)](#). The documentation falls into two major types: [Documentation that covers all years](#) and [documents that cover one particular year](#). If you are new to CE PUMD data, you may want to explore the [CE PUMD Getting Started Guide](#)."

Below this text are two sections: "Documents covering all years" and "Documents covering specific years".

Documents covering all years

- [Consumer Expenditure Surveys Public-use Microdata Getting Started Guide](#) provides documentation for the CE PUMD, its conventions, files, sample code, and methodology.
- [Consumer Expenditure Surveys Program Considerations When Using the Public-use Microdata](#) discusses considerations when preparing and interpreting estimates with PUMD.
- [Dictionary for Interview and Diary Surveys \(XLSX\)](#) provides variables and codes from 1996 forward.
- [Source selection file \(XLSX\)](#) identifies which survey data variable comes from when combining the two CE surveys for 1996 forward.
- [Description of income imputation](#) provides information on the methods BLS uses to estimate income since 2004.

Documents covering specific years

- [Hierarchical groupings](#) lists the relation between the summary variables and their contributing variables as they are used in the [published tables](#). Integrated stub (IntStub) lists the variables as BLS integrates them from the Interview and Diary Surveys.

Spreadsheet 1996-2017

ce_source_integrate (1) - Excel

Creech, Brett J. - BLS

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW

Clipboard Font Alignment Number Styles Cells Editing

AutoSum Fill Clear Sort & Find & Filter Select

Survey Source of Data for Consumer Expenditure Survey Integrated Tables

The detailed list of characteristics, income, and expenditure items below shows which component—the Diary Survey ("D") or the Interview Survey ("I")—was used as the source for that item in the published Consumer Expenditure Survey data tables for each year

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
3	Vehicles	VEHICLES	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C	C

READY

Type here to search

9:08 AM 7/8/2019

Interview Example

ce_source_integrate - Excel

Creech, Brett J. - BLS

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW

Clipboard Font Alignment Number Styles Cells Editing

B679 : X ✓ fx New aircraft

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC			
1	Survey Source of Data for Consumer Expenditure Survey Integrated Tables																															
2	The detailed list of characteristics, income, and expenditure items below shows which component—the Diary Survey ("D") or the Interview Survey ("I")—was used as the source for that item in the published Consumer Expenditure Survey data tables for each year																															
3																																
4	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	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	USEDCCARS	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							

AllYears_IntStub

READY

Type here to search

10:58 AM 7/16/2019

Diary Example

ce_source_integrate - Excel

Creech, Brett J. - BLS

FILE

HOME

INSERT

PAGE LAYOUT

FORMULAS

DATA

REVIEW

VIEW

Paste

Clipboard

Calibri

11

A

A

</

Source Selection Change

ce_source_integrate - Excel

Creech, Brett J. - BLS

FILE HOME INSERT PAGE LAYOUT FORMULAS DATA REVIEW VIEW

Clipboard Font Alignment Number Styles Cells Editing

B649 Infant nightwear, loungewear

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z	AA	AB	AC			
1	Survey Source of Data for Consumer Expenditure Survey Integrated Tables																															
2	The detailed list of characteristics, income, and expenditure items below shows which component—the Diary Survey ("D") or the Interview Survey ("I")—was used as the source for that item in the published Consumer Expenditure Survey data tables for each year																															
3																																
4	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							
646	4	Infant coat, jacket, snowsuit	410110	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I							
647	4	Infant dresses, outerwear	410120	D	D	D	D	D	I	I	I	I	I	I	I	I	I	I	I	I	I	D	D	D	D							
648	4	Infant underwear	410130	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D							
649	4	Infant nightwear, loungewear	410140	I	D	D	D	D	I	I	I	I	I	I	I	I	I	I	I	I	I	D	D	I	I							
650	4	Infant accessories	410901	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D							
651	3	Footwear	FOOTWEAR	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G							
652	4	Men's footwear	400110	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D							
653	4	Boys' footwear	400210	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D							
654	4	Women's footwear	400310	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D							
655	4	Girls' footwear	400220	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D							
656	3	Other apparel products and services	OTHAPPRL	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G	G							
657	4	Material for making clothes (thru Q20124)	420110						D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D							
658	4	Sewing patterns and notions (thru Q20124)	420120						D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D	D							
659	4	Material and supplies for sewing, needlework, quilting (includes household items) (new UCC Q20131)	420115	D	D	D	D	D																								
660	4	Watches	430110	D	D	D	D	D	D	D	D	D	D	I	I	I	I	I	I	I	I	I	I	D	D							
661	4	Jewelry	430120	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	D	D							
662	4	Shoe repair and other shoe service	440110	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I								
663	4	Coin-operated apparel laundry and dry cleaning	440120	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	D	D							
664	4	Alteration, repair and tailoring of apparel and	440130	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I	I								

READY AVERAGE: 410140 COUNT: 24 SUM: 410140 83%

Type here to search

10:55 AM 7/16/2019

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