2014 Users' Documentation

Interview Survey
Public-Use Microdata (PUMD)
Consumer Expenditure

September 03, 2015

Table of Contents

I.   Introduction

The Consumer Expenditure Survey (CE) program provides data on the buying habits of American consumers. These data are primarily used as weights for the Consumer Price Index. However, CE also provides the data to the public for research in two formats. The first format are tabulations of average and aggregate expenditures and income in news releases, databases, and tables. The second format are individual responses to the CE Survey in the Public-Use Microdata (PUMD). For broad analysis the former format is better suited for detailed studies the latter format may proof more useful.

This document discusses the methodology of PUMD for the Interview Survey. This survey collects 95 percent of the total expenditures and income by households. The current data set covers 2014 and the first quarter of 2015. The data are presented in major data files, processing files, and detailed expenditure files. To provide novice users additional assistance, CE prepared a "Getting started with Consumer Expenditure Public-use Microdata."

This document does not list the changes from the previous year, the topcoding and other suppressions of sensitive data, and the response rates. These items can be found in the 2014 Documentation zip file.

The microdata files are in the public domain and, with appropriate credit, may be reproduced without permission. A suggested citation is: "U.S. Department of Labor, Bureau of Labor Statistics, Consumer Expenditure Survey, Interview Survey, 2014."

II.   File Information

The Interview Survey microdata are provided as SAS, STATA, SPSS. or ASCII comma-delimited files. The 2014 Interview release contains three groups of files:

- **8 major data files** (FMLI, MEMI, MTBI, ITBI, ITII, NTAXI, FPAR, and MCHI)
- **4 types of processing files**
- **43 detailed expenditure data files** (EXPN files)

Six of the eight major data files (FMLI, MEMI, MTBI, ITBI, ITII, and NTAXI) are organized by the calendar quarter of the year in which the data were collected. (For a description of calendar and collection years, see Section Calendar Period Versus Collection Period) There are five[1] quarterly data sets for each of these files, running from the first quarter of 2014 through the first quarter of 2015.

The FMLI file contains CU characteristics, income, and summary level expenditures; the MEMI file contains member characteristics and income data; the MTBI file contains expenditures organized on a monthly basis at the UCC level; the ITBI file contains income data converted to a monthly time frame and assigned to UCCs; and the ITII file contains the five imputation variants of the income data converted to a

---

[1] NTAXI, the new tax files are only available the four quarters of 2013Q2-2014Q1.

monthly time frame and assigned to UCCs. The NTAXI file contains federal and state tax information for each tax unit within the CU.

In 2009, CE added the FPAR and MCHI files. They are grouped as 2-year datasets (2013 and 2014), plus the first quarter of the 2015 and present paradata about the interview survey. The FPAR file contains CU level data about the Interview survey, including timing and record use. The MCHI file contains data about each interview contact attempt, including reasons for refusal and times of contact. Both FPAR and MCHI files contain five quarters of data.

The processing files enhance computer processing and tabulation of data, and provide descriptive information on item codes. CE provides these processing files:

- **ISTUB** and **INTSTUB** provide the aggregation scheme used in the published consumer expenditure survey interview tables and integrated tables. These files contain UCCs and their abbreviated titles, identifying the expenditure, income, or demographic item represented by each UCC.
- **CAPIVEHI** provides the codes for the vehicle makes.
- **Sample programs** with code that approximates the tables that CE publishes in SAS and R.
- "**User's Guide to Income Imputation in the CE**" provides information on how to appropriately use the imputed income data.

The EXPN files contain expenditure and ancillary descriptive data, that are often not in the FMLI or MTBI files. EXPN files facilitate the analysis by allowing users identify distinct spending categories. EXPN files are organized by their respective sections in the Computer Assisted Personal Interview (CAPI) and cover five quarters. The table below lists the names of the EXPN files, the CAPI section, and a brief description. For more detail, see the questionnaire on the CE Survey webpage.

List of EXPN files

File name	Description
RNT Section 2	*Rented Living Quarters – CU Tenure, Rental Payments, Facilities, and Services for Sample Unit and Other Units* Section 2 collects rent and related expenses from households who rent their homes or other properties. The questions asked during the first interview vary from those asked during subsequent interviews.
OPB Section 3, Part B	*Owned Living Quarters and Other Owned Real Estate – Detailed Property Description* Section 3, Part B collects detailed information about owned properties reported in Section 3, Part A, including the date of settlement, total cost, current market value, and annual property taxes.
OPD Section 3, Part D	*Owned Living Quarters and Other Owned Real Estate – Disposed of Property* Section 3, Part D collects information on properties that have been sold, traded, given to someone outside of the household, or otherwise disposed of by the household.
MOR Section 3, Part F	*Owned Living Quarters and Other Owned Real Estate – Mortgages* Section 3, Part F deals with mortgages and home equity loans, including the type of loan, interest rate and term, and amount of payment.
HEL Section 3, Part F	*Owned Living Quarters and Other Owned Real Estate – Lump Sum Home Equity Loans* Section 3, Part F deals with mortgages and home equity loans, including the type of loan, interest rate and term, and amount of payment.

File name	Description
OPH Section 3, Part H	*Owned Living Quarters and Other Owned Real Estate – Line of Credit Home Equity Loans* Section 3, Part H covers payments made on home equity lines of credit.
OPI Section 3, Part I	*Owned Living Quarters and Other Owned Real Estate – Ownership Costs* Section 3, Part I collects ownership costs, including extra mortgage and home equity loan payments, ground rent, homeowners' association fees, condominium and cooperative fees, and special assessments. The respondent is also asked to provide an estimate of the owned property's rental value.
UTA Section 4, Part A	*Utilities and Fuels for Owned and Rented Properties – Telephone Expenses* Section 4, Part A deals with expenditures for telephone services, including residential service and cellular service.
UTP Section 4, Part B	*Utilities and Fuels for Owned and Rented Properties – Other Telephone Expenses* Section 4, Part B deals with other telephone expenses, including the purchase of pre-paid telephone and cellular cards and spending on public telephone use.
UTI Section 4, Part C	*Utilities and Fuels for Owned and Rented Properties – Internet Service Expenses* *Section 4, Part C collects expenditures on cable, satellite, and internet services for the household residence and other owned properties, including cable or satellite TV, satellite radio services, internet service provider, online games, and internet services at web cafes or internet kiosks.*
UTC Section 4, Part D	*Utilities and Fuels for Owned and Rented Properties – Detailed Questions* *Section 4, Part D collects expenditures on fuels and utilities for the household residence and other owned properties as well as rented vacation properties, including electricity, natural gas, other fuels, water service, sewer maintenance, garbage collection, and cable television or satellite service.*
CRA Section 5, Part A	*Construction, Repairs, Alterations, and Maintenance of Owned and Rented Property* *Section 5, Part A deals with expenses for supplies and services related to home construction, repair, alteration and maintenance.*
CRB Section 5, Part B	*Construction, Repairs, Alterations, and Maintenance of Owned and Rented Property – Jobs Not Yet Started* *Section 5, Part B deals with expenses for supplies and services related to home construction, repair, alteration and maintenance.*
APA Section 6, Part A	*Appliances, Household Equipment, and Other Selected Items – Purchase of Household Appliances* *Section 6, Part A covers purchases and rentals of major household appliances, such as kitchen appliances, clothes washers, and clothes dryers.*
APB Section 6, Part B	*Appliances, Household Equipment and Other Selected Items – Purchase of Household Appliances and Other Selected Items* *Section 6, Part B deals with purchases and rentals of small appliances, televisions, radios, sound equipment, sports and exercise equipment, and miscellaneous other household items.*
EQB Section 7	*Household Item, Repairs, Service Contracts*

File name	Description
	Section 7 covers expenditures for maintenance, repair, and service contracts for appliances, televisions, computers, tools, pest control service, and other household items.
FRA Section 8, Part A	*Home Furnishings and Related Household Items – Purchases* *Section 8, Part A deals with purchases of furniture, household decorative items, dishes, household linens, floor coverings, and window coverings.*
FRB Section 8, Part B	*Home Furnishings and Related Household Items – Rental, Leasing, or Repair of Furniture* *Section 8, Part B deals with expenditures for furniture rental and repair.*
CLA Section 9, Part A	*Clothing and Jewelry – Clothing, Watches, and Jewelry* *Section 9, Part A deals with purchases of clothing for persons age 2 years old and older.*
CLD Section 9, Part B	*Clothing and Jewelry – Clothing Services* *Section 9, Part D deals with expenses for clothing services, including alterations, jewelry repair, clothing rental, and clothing storage.*
RTV Section 10	*Rented and Leased Vehicles – Rented Vehicles* *Section 10 deals with vehicle rentals and leases. The questions asked during the first interview vary from those asked during subsequent interviews.*
LSD Section 10	*Rented and Leased Vehicles – Leased Vehicles* *Section 10 in a first interview asks if there are any vehicle lease payments or new leases, then collects details about those vehicles and expenses.*
OVB Section 11	*Owned Vehicles – Detailed Questions* *Section 11 collects expenditures for owned vehicles. The questions asked depend on whether it is the first interview or a subsequent interview, and whether there are any previously reported vehicles owned by the consumer unit.*
OVC Section 11	*Owned Vehicles – Disposal of Vehicles* *Section 11 collects expenditures for owned vehicles. The questions asked depend on whether it is the first interview or a subsequent interview, and whether there are any previously reported vehicles owned by the consumer unit.*
VEQ Section 12, Part A	*Vehicle Operating Expenses – Vehicle Maintenance and Repair, Parts and Equipment* *Section 12, Part A deals with expenses for vehicle services, parts and equipment.*
VLR Section 12, Part B	*Vehicle Operating Expenses – Licensing, Registration, and Inspection of Vehicles* *Section 12, Part B deals with expenses for driver's licenses, vehicle registration, and vehicle inspection.*
VOT Section 12, Part C	*Vehicle Operating Expenses – Other Vehicle Operating Expenses* *Section 12, Part C deals with other vehicle operating expenses, including a monthly average expenditure on gasoline, purchases of oil and other fluids, parking fees, towing charges, docking or landing fees, and expenses for auto repair service policies and clubs.*
INB Section 13, Part B	*Insurance Other Than Health – Detailed Questions*

File name	Description
	Section 13, Part B collects detailed information about each type of non-health insurance policy that was reported.
IHB Section 14, Part B	*Hospitalization and Health Insurance – Detailed Questions* *Section 14, Part B collects detailed information about each health insurance policy that was reported in Section 14, Part A.*
IHC Section 14, Part C	*Hospitalization and Health Insurance – Medicare, Medicaid, and Other Health Insurance Plans Not Directly Paid for by the Household* *Section 14, Part C covers participation in health insurance plans for which the household does not pay directly, such as Medicare, Medicaid, and military health care plans.*
IHD Section 14, Part C	*Hospitalization and Health Insurance – Medicare Prescription Drug Program* *Section 14, Part C covers participation in health insurance plans for which the household does not pay directly, such as Medicare, Medicaid, and military health care plans.*
MDB Section 15, Part A	*Medical and Health Expenditures – Payments* *Section 15, Part A collects out-of-pocket medical payments, including payments for medical services, prescription drug purchases, and rentals or purchases of medical supplies and equipment.*
MDC Section 15, Part B	*Medical and Health Expenditures – Reimbursements* *Section 15, Part B covers reimbursements received by the consumer unit for medical services, prescription drugs, and medical supplies or equipment.*
EDA Section 16	*Educational Expenses* *Section 16 collects educational expenses, including recreational lesson fees, tuition, room and board, purchases of school books and equipment, and other educational expenses.*
SUB Section 17	*Subscriptions, Memberships, Books, and Entertainment Expenses* *Section 17 deals with expenditures for subscriptions, mail order clubs, season tickets, reference books, recreational club memberships and shopping club memberships.*
TRD Section 18, Part A	*Trips and Vacations – Screening Questions* *Section 18, Part A is used to determine whether the household has taken any trips during the reference period, or to follow up on previously reported trips. Specific questions in this section are used to distinguish between trip expenses paid by the household and those paid by someone else. Only expenses paid by the household are included in CE Survey estimates.*
TRV Section 18, Part BC	*Trips and Vacations – Detailed Questions* *Section 18, Part BC collects detailed information about the trips identified in Part A, including the value of any package deals and expenses for transportation, lodging, food, and entertainment on trips.*
TRE Section 18, Part E	*Trips and Vacations – Trip Expenses for Non-Household Members* *Section 18, Part E deals with trip expenses paid by the household for someone outside of the household.*

File name	Description
TRF Section 18, Part F	*Trips and Vacations – Local Overnight Stays* *Section 18, Part F collects detailed information about local overnight stays, including the value of any package deals and expenses for lodging, food, and entertainment.*
MIS Section 19, Part A	*Miscellaneous Expenses* *Section 19, Part A covers miscellaneous expenses such as funeral expenses, legal and accounting fees, various household services, babysitting and adult care, toys and games, lotteries, and pet expenses.*
CNT Section 19, Part B	*Miscellaneous Expenses – Contributions* *Section 19, Part B deals with payments and contributions to persons outside of the household, and to religious, political, educational and other charitable organizations.*
XPA Section 20, Part A	*Expense Patterns For Food, Beverages, and Other Selected Items – Food and Beverages* *Section 20, Part A asks for expenditure estimates for groceries, alcoholic beverages, and meals away from home.*
XPB Section 20, Part B	*Expense Patterns For Food, Beverages, and Other Selected Items – Selected Services and Goods* *Section 20, Part B deals with expenses for dry cleaning, laundry service, cigarettes, personal services, banking fees, taxis, limousines, and mass transportation.*

A. Dataset Names

The file naming convention is listed in the table below. The files are compressed and can be uncompressed with most unzip utilities.

\INTRVW14\FMLI141x.* (Interview FMLI file for first quarter, 2014)
\INTRVW14\MEMI141x.* (Interview MEMI file for first quarter, 2014)
\INTRVW14\MTBI141x.* (Interview MTBI file for first quarter, 2014)
\INTRVW14\ITBI141x.* (Interview ITBI file for first quarter, 2014)
\INTRVW14\ITII141x.* (Interview ITII file for first quarter, 2014)
\INTRVW14\NTAXI142.* (Interview NTAXI file for the second quarter, 2014)
\INTRVW14\FMLI142.*   (etc.)
\INTRVW14\MEMI142.*
\INTRVW14\MTBI142.*
\INTRVW14\ITBI142.*
\INTRVW14\ITII142.*
\INTRVW14\NTAXI143.*
\INTRVW14\FMLI143.*
\INTRVW14\MEMI143.*
\INTRVW14\MTBI143.*
\INTRVW14\ITBI143.*
\INTRVW14\ITII143.*
\INTRVW14\NTAXI143.*
\INTRVW14\FMLI144.*
\INTRVW14\MEMI144.*
\INTRVW14\MTBI144.*

\INTRVW14\ITBI144.*	
\INTRVW14\ITII144.*	
\INTRVW14\FMLI151.*	
\INTRVW14\MEMI151.*	
\INTRVW14\MTBI151.*	
\INTRVW14\ITBI151.*	
\INTRVW14\ITII151.*	
\INTRVW14\NTAXI151.*	
\INTRVW14\VEHI14.txt	
\PARA14\FPAR1214.*	
\PARA14\MCHI1214.*	
\EXPN14\RNT14.*	
\EXPN14\OPB14.*	
\EXPN14\OPD14.*	
\EXPN14\MOR14.*	
\EXPN14\HEL14.*	
\EXPN14\OPH14.*	
\EXPN14\OPI14.*	
\EXPN14\UTA14.*	
\EXPN14\UTP14.*	
\EXPN14\UTI14.*	
\EXPN14\UTC14.*	
\EXPN14\CRA14.*	
\EXPN14\CRB14.*	
\EXPN14\APA14.*	
\EXPN14\APB14.*	
\EXPN14\EQB14.*	
\EXPN14\FRA14.*	
\EXPN14\FRB14.*	
\EXPN14\CLA14.*	
\EXPN14\CLD14.*	
\EXPN14\RTV14.*	
\EXPN14\LSD14.*	
\EXPN14\OVB14.*	
\EXPN14\OVC14.*	
\EXPN14\VEQ14.*	
\EXPN14\VLR14.*	
\EXPN14\VOT14.*	
\EXPN14\INB14.*	
\EXPN14\IHB14.*	
\EXPN14\IHC14.*	
\EXPN14\IHD14.*	
\EXPN14\MDB14.*	
\EXPN14\MDC14.*	
\EXPN14\EDA14.*	
\EXPN14\SUB14.*	
\EXPN14\TRD14.*	
\EXPN14\TRV14.*	
\EXPN14\TRE14.*	
\EXPN14\TRF14.*	
\EXPN14\MIS14.*	
\EXPN14\CNT14.*	
\EXPN14\XPA14.*	

```
\EXPN14\XPB14.*
```

B. Record Counts

The following are the number of records in each data set. Each EXPN file contains 5 quarters of data within a single data set.

Data set	Record counts
FMLI141x	6,483
FMLI142	6,489
FMLI143	6,466
FMLI144	6,470
FMLI151	6,413
MEMI141x	15,997
MEMI142	16,107
MEMI143	15,999
MEMI144	15,968
MEMI151	15,661
MTBI141x	477,277
MTBI142	469,510
MTBI143	477,769
MTBI144	474,474
MTBI151	480,389
ITBI141x	396,714
ITBI142	397,584
ITBI143	396,039
ITBI144	395,673
ITBI151	382,866
ITII141x	553,455
ITII142	555,345
ITII143	551,985
ITII144	553,395
ITII151	540,000
NTAXI141x	7,728
NTAXI142	7,820
NTAXI143	7,796

Data set	Record counts
NTAXI144	7,797
NTAXI151	7,741
FPAR1314	107,635
MCHI1314	551,032
APA14	2,452
APB14	30,846
CLA14	130,330
CLD14	1,966
CNT14	30,482
CRA14	793
CRB14	9,548
EDA14	12,192
EQB14	4,742
FRA14	24,720
FRB14	328
HEL14	616
IHB14	33,693
IHC14	13,500
IHD14	5,660
INB14	65,551
LSD14	1,691
MDB14	51,438
MDC14	1,175
MIS14	48,897
MOR14	12,789
OPB14	23,779
OPD14	117
OPH14	1,356
OPI14	37,930
OVB14	57,053
OVC14	1,969
RNT14	13,256
RTV14	897
SUB14	58,533
TRD14	5,794
TRE14	3,453
TRF14	362
TRV14	11,316
UTA14	39,788

Data set	Record counts
UTC14	91,595
UTI14	34,991
UTP14	32,320
VEQ14	34,610
VLR14	11,519
VOT14	32,320
XPA14	32,320
XPB14	32,319

C. Data Flags

Data fields on the FMLI, MEMI, MTBI, NTAXI and EXPN files are explained by flag variables following the data field. The names of the flag variables are derived from the names of the data fields they reference.

In general, the rule for naming variable flags is to add an underscore to the last position of the data field name, for example SALARYX becomes SALARYX_. However, if the data field name is eight characters in length, then the fifth position is replaced with an underscore. If this fifth position is already an underscore, then the fifth position is changed to a zero, so that RETSURVX becomes RETS_RVX, but EDUC_REF becomes EDUC0REF.

Flag value	Description
A	Valid blank; a blank field where a response is not anticipated
B	Invalid blank due to invalid nonresponse; nonresponse that is not consistent with other data reported by the CU
C	Blank due to "Don't know," refusal, or other nonresponse
D	Valid value, unadjusted
E	Valid value, allocated
F	Valid value, imputed or adjusted in some other way
G	Valid value, allocated *and* imputed
H	Valid blank for an expenditure that is a "parent record" where the expenditure was allocated to other records and the original expenditure was overwritten with a blank
T	Valid value, topcoded or suppressed
U	Valid value, allocated *then* topcoded or suppressed
V	Valid value, imputed or adjusted in some other way *then* topcoded or suppressed
W	Valid value, allocated *and* imputed or adjusted in some other way *then* topcoded or suppressed

D. Income Imputation

Beginning in 2004, the CE implemented multiple imputation of income data. Imputation allows income values to be estimated when they are not reported. Many income variables and other income related variables are now imputed using a multiple imputation process. These imputed income values are included in the FMLI, MEMI, ITBI and ITII files. The multiple imputation process derives five imputation values, and a mean imputation value, for each selected income variable. More information on the imputation process and how to appropriately use the data are found in the document "User's guide to Income Imputation in the CE."

In the public-use microdata, not all of the imputed income variables contain the derived imputation values. For some income variables, the five derived imputations are excluded and only the mean of those imputations is available. For these variables, there are 3 associated income variables in the FMLI and

MEMI files (*INCOME*M, *INCOME*M_, and *INCOME*I). For all other imputed income variables, there are 7 associated variables in the FMLI and MEMI files:

*INCOME*1   the first imputed income value or the reported income value, if non-missing
*INCOME*2   the second imputed income value or the reported income value, if non-missing
*INCOME*3   the third imputed income value or the reported income value, if non-missing
*INCOME*4   the fourth imputed income value or the reported income value, if non-missing
*INCOME*5   the fifth imputed income value or the reported income value, if non-missing
*INCOME*M   the mean of the five imputed income values
*INCOME*M_   the flag variable for the imputed variable (see Section III.C. Data Flags)
*INCOME*I   the imputation indicator variable (see below)

Income variables that have imputed values as components (ex: FINCBTXM) will also have 5 imputed values and a mean based on each of the imputed components.

The imputation indicator variable is a 3 digit number that is coded as follows:

The first digit in the 3 digit code defines the imputation method as such:

1. No Imputation
2. Multiple Imputation due to invalid blank only
3. Multiple Imputation due to bracketing only
4. Multiple Imputation due to invalid blanks and bracketing
5. Multiple Imputation due to conversion of a valid blank to an invalid blank (this occurs only when initial values for all sources of income for the CU were valid blanks)

The meaning of the last two digits of the three digit code differs depending on whether you are looking at one of the components of overall income, like FSALARYM, or you are looking at the summary level variable FINCBTXM. For the components, the last 2 digits represent the number of family members who had their data imputed for that source. For example, if a family had a value of 302 for FSALARYI that would mean that 2 of the members in the family had their salary income imputed and that in both cases the imputation was due to bracketing only. For the summary level variable FINCBTXM which is a summation of all of the income components, the last 2 digits represent the number of income sources imputed for each member added together. For example, if a family had 3 members and 2 had salary income imputed due to invalid blank only, and 2 had self-employment income imputed due to bracketing only, and that was the only income data imputed for members of that family, then FSALARYI for the family would be 202, FSMPFRMI would be 302, and FINCBTXI would be 404.

The ITBI file includes income UCCs mapped from the associated *INCOME*M variable in the FMLI files. The ITII file includes UCCs mapped from income variables subject to income imputation, including the variable IMPNUM to indicate the imputation number 1 - 5.

E. File Notation

Every record from each data file includes the variable NEWID, the CU's unique identification number, which is used to link records of one CU from several files across all quarters in which they participate.

Data fields for variables on the microdata files have either numeric or character values. The format column in the data dictionary distinguishes whether a variable is numeric (NUM) or character (CHAR) and shows the number of field positions the variable occupies. Variables that include decimal points are formatted as NUM(t.r) where t is the total number of positions occupied, and r is the number of places to the right of the decimal.

In addition to format, this data dictionary give an item description, questionnaire source, and identification of codes where applicable for each variable. The questionnaire source format will now indicate the CAPI section where the question can be found.

An asterisk (*) is shown in front of new variables, those which have changed in format or definition, and those which have been deleted. Variables whose format has expanded are moved to the end of the files, and their original positions are left blank. New variables are added to the end of the files after variables whose format has changed. The positions of deleted variables are left blank.

Some variables require special notation. The following notation is used throughout the data dictionary for all files:

*D(Yxxq) identifies a variable that is deleted as of the quarterly file indicated. The year and quarter are identified by the 'xx' and 'q' respectively. For example, the notation *D(Y142) indicates the variable is deleted starting with the data file of the second quarter of 2014.

*N(Yxxq) identifies a variable that is added as of the quarterly file indicated. The year and quarter are identified by the 'xx' and 'q' for new variables in the same way as for deleted variables.

*C(Yxxq) identifies a variable's content or description has changed beginning in the quarterly file indicated. The year and quarter are identified by the 'xx' and 'q' for new variables in the same way as for deleted variables.

*L indicates that the variable can contain negative values.

A. State Identifier

The variable STATE identifies the state of residence of respondents. Since the CE survey is not designed to produce state-level estimates, summing the CU weights by state will *not* yield representative state population totals because of three reasons:

- CU's basic weight reflects its *national* probability of selection among a group of primary sampling units of similar characteristics. For example, sample units in an urban nonmetropolitan area in California may represent similar areas in Wyoming and Nevada.
- CUs are post-stratified nationally by sex-age-race. For example, the weights of CUs containing a black male, age 16-24 in Alabama, Colorado, or New York, are all adjusted equivalently.
- Some CUs are located in PSU that span over two states or are suppressed due to nondisclosure requirements by Census. For information, see 2014 Topcoding and Suppression in the 2014 Documentation zip file.

Nevertheless state-level estimates that are unbiased in a repeated sampling sense can be calculated for various statistical measures, such as means and aggregates. However, the estimates will generally be subject to large variances and may be far from the true state population.

List of state identifiers

01	Alabama	29	Missouri
02	Alaska	30	Montana
04	Arizona	31	Nebraska
05	Arkansas	32	Nevada
06	California	33	New Hampshire
08	Colorado	34	New Jersey
09	Connecticut	36	New York
10	Delaware	37	North Carolina

11	District of Columbia	39	Ohio
12	Florida	40	Oklahoma
13	Georgia	41	Oregon
15	Hawaii	42	Pennsylvania
16	Idaho	44	Rhode Island
17	Illinois	45	South Carolina
18	Indiana	46	South Dakota
20	Kansas	47	Tennessee
21	Kentucky	48	Texas
22	Louisiana	49	Utah
23	Maine	51	Virginia
24	Maryland	53	Washington
25	Massachusetts	54	West Virginia
26	Michigan	55	Wisconsin
27	Minnesota		
28	Mississippi		

F. Allocation and Record Origin (EXPN)

Expenditures on the EXPN files that have been allocated can be identified through their flag variable, which will have a value, set to 'H' (see Section III.C. Data Flags). These expenditures can be recreated using the fields SEQNO and ALCNO. SEQNO is a counter assigned to make records unique. ALCNO is zero for all original expenditure records. If ALCNO is greater than zero, the corresponding expenditure record is the result of allocation of an original record whose expenditure field has been replaced with a blank for that CU. By summing expenditures for records with ALCNO greater than zero and the same SEQNO as the original record, one can arrive at the value which was allocated.

For every EXPN record, the codes for the variable REC_ORIG are the following codes:

REC_ORIG Code	Description
1	Data reported in the current quarter's interview. Ex: expenditures reported in the current reference period for the current reference period
2	Data reported in the previous quarter's interview that are encompassed by the current reference period. These data are brought forward through the reference period adjustment process. Ex: expenditures reported in the previous reference period for the current reference period. This can happen when a respondent reports making routine payments (rent, phone bill, etc.) that will continue throughout the year. Instead of the FRs asking all the questions again, they simply verify that these payments still exist and whether or not there are any new ones.
3	Data reported in the previous quarter's interview that are encompassed by the current reference period, and this logical record duplicates a logical record from the current interview month. These data are brought forward through the reference period adjustment process; the data duplication is also identified during this process. Ex: similar to REC_ORIG code 2, except through some oversight the expenditure that was carried over from the previous reference period is also being collected in the current reference period. Processing identifies and removes these duplicates.

REC_ORIG Code	Description
4	Inventory data reported in previous quarters' interviews brought forward through the inventory update process. No updates are applied to this logical record as none are indicated in the current inventory chart. Ex: inventory items that are reported in a previous reference period that still apply to the current reference period. This includes items such as automobiles, houses, and insurance policies. These items are stored ("inventoried") and carried forward throughout all interviews, being updated as necessary along the way. These data are used to ease the flow of the interview, for example, instead of asking whether or not the CU owns a car each interview, the FR can reference the previously reported car and ask relevant expenditure questions relating to said car. If there is no update to the inventoried item it is given code 4. However, if there is an update applied (insurance policy remains intact but changes, for instance) it receives a code 5.
5	Inventory data reported in previous quarters' interviews brought forward through the inventory update process. Updates are applied based upon data contained in the current inventory chart. Ex: similar to REC_ORIG code 4, except that an update has been applied to the inventoried item
6	Data created by the processing system. Ex: rare though they are, these data have been created by CE. There are very few instances where this may occur, but it is most often used to correct inconsistencies

G. Notes on Files

There are some specifics that are unique to particular files to be aware of when working with the datasets. Important notes that were previously listed with the variable descriptions can now be found in this section of the documentation. Each note is broken into file and category.

1. Consumer Unit (CU) Characteristics and Income File (FMLI)

The "FMLI" file, also referred to as the "Consumer Unit Characteristics and Income" file, contains CU characteristics, CU income, and characteristics and earnings of the reference person and of the spouse. The file includes weights needed to calculate population estimates and variances. (See Sections V. Estimation Procedures and VI. Reliability Statement.)

Summary expenditure variables in this file can be combined to derive quarterly estimates for broad consumption categories. Demographic characteristics, such as family size, refer to the CU status on the date of the interview. Demographic characteristic information may change between interviews if, for example, a member enters or leaves the CU. Income variables contain annual values. Income data are collected in the second and fifth interviews only and cover the 12 months prior to the date of interview. Income data collected in the second interview are copied to the third and fourth interviews. Income data are updated only if a CU member over 13 is new to the CU or has not worked in previous interviews and has now started working. When there is a valid nonresponse, or where nonresponse occurs and there is no imputation, there will be missing values. The type of nonresponse is explained by associated data flag variables described in Section III.C. Data Flags.

Summary Expenditure Data

Main Summary Level Expenditure Variables

For each summary expenditure category listed below there are two variables. They apportion expenditures reported for the three-month reference period of the interview to the calendar quarters, relative to the month of interview, in which the expenditures occurred. The first variable contains expenditures made by the CU in the calendar quarter previous to the month of interview. These "previous quarter" expenditure variables are identified by "PQ" placed as the last two letters of the variable name. The second variable contains expenditures made in the calendar quarter of the month of interview (last two letters of the variable name "CQ"). So if CUs were interviewed in May (when they reported their February, March, and April expenditures), the "PQ" variable would contain their February and March expenditures since the previous calendar quarter to a May interview is from January to March. The "CQ" variable for these CUs would contain only their April expenditures. The variables are set up this way to facilitate analysis by calendar time period. For example, to calculate an expenditure category mean for a given calendar quarter, expenditures from the "CQ" variable for interviews conducted during the quarter of interest are added to amounts from the "PQ" variable for interviews conducted during the subsequent quarter prior to dividing by the number of observations. To derive expenditure statistics by collection period, i.e., for interviews conducted during a specific period, it is necessary to obtain all expenditures reported during each interview by summing the "PQ" and "CQ" variables of the desired expenditure category. See Section V.A.1.b. Calendar Period Versus Collection Period for a detailed explanation of calendar and collection periods.

Note:

MISC2PQ(CQ) contains UCCs that are a subset of those included in MISCPQ(CQ) – miscellaneous expenditures. Component UCCs in MISCPQ(CQ) have been separated according to collection method. UCCs for which the values are obtained from questions asked in interviews 2 through 5 are now in MISC1PQ(CQ), while MISC2PQ(CQ) contains those UCCs from questions asked only in the fifth interview.

To obtain population or sample estimates, the summary variable MISCX4PQ(CQ) has been created. It is comprised of MISC1PQ(CQ) expenditures and MISC2PQ(CQ) expenditures that have been multiplied by four, in order to account for families not in their fifth interviews. Similarly, TOTEX4PQ(CQ) reflects the adjustments for "non-fifth interview" families in MISC2PQ(CQ) and CASHCOPQ(CQ).

For 2014Q1 MISCX4CQ(PQ) and TOTEX4PQ(CQ) overestimate the values of CASHCOPQ(CQ) and a portion of MISC2PQ(CQ) for "fifth interview" CUs and should only be used for population estimates.

Travel related summary expenditure variables

The summary level "travel" expenditure variables (T-variables) describe expenditures by consumer units on out-of-town trips. These variables have been constructed to facilitate research on travel related spending. Because the UCCs describing these items are scattered across several categories, they are collected in one format for the convenience of the user. As is the convention with the main summary level expenditure variables, each of the T-variable categories are sorted by expenditures that took place during the previous calendar quarter and current calendar quarter. However for the T-variables, the previous quarter expenditure variables are appended with "P," and the current quarter expenditure variables are appended with "C."

Expenditure Outlays Summary Variables

Expenditure outlay summary level variables (EVARS) are used to provide a measurement of all expenditure outlays. These variables are constructed similarly to the main summary level expenditure

variables in that they contain interest payments for home mortgage and vehicles when financed. The difference with the EVARS are that they also include payments on principle for home mortgages and vehicles. Note: main summary level expenditure variables are components of the higher aggregated EVARS. The EVARS follow the same naming convention as the main summary level expenditure variables. Expenditures within the collection quarter are sorted by whether they occurred in the previous calendar quarter or in the current calendar quarter. As in the Travel related summary variables, the EVARS are appended with a "P" for previous or "C" for current.

2. Member Characteristics and Income File (MEMI)

The "MEMI" file, also referred to as the "Member Characteristics and Income" file, contains selected characteristics for each CU member, including identification of their relationship to reference person. Characteristics for the reference person and spouse appear on both the MEMI file and FMLI file. Demographic characteristic data, such as age of CU member, refer to the member status on the date of the interview. Characteristic information may change between interviews. Income data are collected in the second and fifth interviews for all CU members over 13 years of age and in the third and fourth interviews for members over 13 who are new to the CU or who previously reported not working and are now working. Member income data from the second interview are carried over to the third and fourth interviews subject to the above conditions. Income variables contain annual values for the 12 months prior to the interview month. Income taxes withheld and pension and retirement contributions are shown both annually and as deductions from the member's last paycheck. When there is a valid nonresponse, or where nonresponse occurs and there is no imputation, there will be missing values. The type of nonresponse is explained by associated data flag variables described in Section III.C. Data Flags.

3. Monthly Expenditure File (MTBI)

In the MTBI file, each expenditure reported by a CU is identified by UCC, gift/nongift status, and month in which the expenditure occurred. UCCs are six digit codes that identify items or groups of items. The expenditure data record purchases that were made during the three month period prior to the month of the interview. There may be more than one record for a UCC in a single month if that is what was reported to the interviewer. There are no missing values in this file. If no expenditure was reported for the item(s) represented by a UCC, then there is no record for the UCC on the file.

4. Income File (ITBI)

The "ITBI" file, also referred to as the "Income" file, contains CU characteristics and income data. This file is created directly from the FMLI file and contains the same annual and point-of-interview data in a monthly format. It was created to facilitate linking CU income and characteristics data with MTBI expenditure data. As such, the file structure is similar to MTBI. Each characteristic and income item is identified by UCC (For a list of the UCCs, see Istub), gift/nongift status, and month. There are no records with missing values in ITBI. If the corresponding FMLI file variable contained a missing value, there is no record for the UCC.

5. Imputed Income File (ITII)

As a result of the introduction of multiply imputed income data in the Consumer Expenditure Survey, the ITII file is now included in the Public Use Microdata. It is very similar to the ITBI file, except that the variable IMPNUM. will indicate the number (1-5) of the imputation variant of the income variable and it only contains UCCs from variables subject to income imputation.

6. Income Tax File (NTAXI)

The NTAXI or "Income Tax File" contains income data collected by the CE, estimated income tax data (federal, state, and combined), NEWID to merge the file with other PUMD files, and flags to identify processing on individual variables. For a complete list of the variables available, see the data dictionary. The data are presented by Tax Unit, which are individuals or groups filing a taxes.

CE does not estimate taxes when our methods have determined that there is a valid nonresponse. In these cases, the data will have missing values. The type of nonresponses are explained by the flags, which are described in Section III.C. Data Flags.

More information about, see the section on Income Tax Estimation.

7. Paradata Files

With the development of computer-assisted modes of data collection, data on the survey process automatically generated by the new electronic modes became known as "paradata." [2] The scope of paradata now includes computer-generated as well as other types of interviewer-reported data about the process of collecting survey data.

Starting in 2005, the CE began recording data about attempts to contact the sample unit through the Contact History Instrument (CHI), developed by the U.S. Census Bureau. CHI provides interviewer-observations for each contact attempt with a sample unit, regardless of whether contact is made.

Additional paradata is collected about the interview within the interview collection instrument (CAPI). This data includes information on the amount of time required to collect each interview and interview section, as well as other interviewer-entered information about the resulting survey.

The paradata files include all eligible interviews for both completed interviews and eligible but non-responding sample units (Type A non-interviews), in Interviews 1 through 5. The case's final disposition for a sample unit can be found in the variable "OUTCOME" in the FPAR file. All other (non-paradata) files in the microdata include only completed interviews (OUTCOME = '201' and '203') and interviews 2 through 5.

The paradata files FPAR1314 and MCHI1314 each contain 9 quarters of data. This allows users to have a possible complete set of interviews (1-5) for respondents in 2013. These files include the variable CUID, which allows users to link the same CU across quarters (and interviews). It also includes the variable NEWID, which allows users to link the paradata for a particular quarter (interview) with other data from that quarter.

The paradata are in two files:

- **CU Level Paradata File (FPAR):** The CU level paradata contains one record per CU per interview. Most of the data included in the file are only relevant to completed or partially completed interviews and will have missing data for non-interviews. The non-interviews in these cases will still have an ID and OUTCOME code.

---

[2] Couper, M. (1998). Measuring survey quality in a CASIC environment. Pp. 41-46 in Proceedings of the Section on Survey Research Methods. Alexandria, VA: American Statistical Association.

This file is derived from information captured automatically in the CAPI instrument in addition to responses entered directly by the interviewer in the CAPI instrument.

This file includes information on the total amount of time needed to complete each section. For a description of the sections and questions, see the interview form of the CAPI.

- **Contact History Attempt File (MCHI):** The contact history attempt file consists of data collected through the CHI instrument. There can be multiple records per CU per quarter.

  Examples of CHI information include whether contact was made, the mode of contact (e.g., by telephone or in person), reasons for non-interview, the strategies the interviewer used when attempting to contact the sample unit, as well the interviewer's observations about interactions with a sample unit that was contacted.

  Interviewers can make a CHI entry immediately after a contact attempt or at a later time (for example, at home). Every time the survey questionnaire is accessed on the laptop, CHI launches automatically upon exiting the questionnaire, at which point, interviewers can complete a CHI entry. Alternatively, a contact attempt entry can also be recorded by selecting a case from the Case Management System and bringing up CHI without opening the survey itself. Interviewers are instructed to complete a CHI record each time a contact attempt is made.[3]

8. Detailed Expenditures Files (EXPN)

The variables QYEAR, NEWID, SEQNO, ALCNO and REC_ORIG are common to all sections of EXPN. Descriptions of these variables can be found in the data dictionary.

9. Processing Files

a. Istub File

Stub files show the hierarchy or aggregation scheme used in the published consumer expenditure tables. The IStub provides the hierarchy for the Interview Survey and the IntStub to integrate both surveys. Each stub file has 7 columns. The stub files are in the documentation zip file and are formatted as follows:

X:\Programs\Istub2014.txt

Name	Content	Code	Format
Type	If information in this line contains aggregation data or not	**1**: Row contains content **2**: Row contains overflow space for descriptions *****: Row contains title	CHAR(1)
Level	Aggregation level	Lowest number (1) corresponds to the highest level of aggregation and the highest number to the lowest level (9)	CHAR(1)
Title	Title of the line item	NA	CHAR(60)
Variable or UCC	Variable or UCC number	NA	CHAR(8)
Source	Source of the information	**I**: UCCs from Interview file **D**: UCCs from Diary file **G**: BLS derived variables **T**: Section title (no data)	CHAR(1)

---

[3] In theory, interviewers are expected to record a CHI entry whenever CHI automatically launches. However, the first CHI screen does have an "out" by allowing interviewers to select the category "Looking at a case – exit CHI". Therefore it is possible for interviewers to complete an interview without ever having recorded a single CHI entry.

Name	Content	Code	Format
		S: Administrative variables	
Factor	Factor applied to data for annual estimation.	NA	CHAR(2)
Group	If the item is an expenditure, income, or asset	**CUCHARS**: CU characteristics **EXPEND**: Expenditures **Food**: Food expenditures **Income**: Income data **Assets**: Data on assets **Addenda**: Data describing variables	CHAR(7)

Each title row represents an aggregate category of its more detailed subcategories which can be an aggregate of more detailed data or the most detailed level of data for that particular category, which would be represented by a row with a UCC rather than a variable name in column 4.

b. Vehicle File

New vehicle codes were introduced with the CAPI instrument and should be used for vehicle information collected from the 2003q2 survey on. These codes can be found in the variable MKMDEL (the first 3 characters) in EXPN Section 10, Part B (Rented and Leased Vehicles – Detailed Questions for Leased Vehicles) and MAKE in EXPN Section 11, Part B (Owned Vehicles - Detailed Questions).

X:\EXPN14\CAPIVEHI14.TXT

CAPIVEHI14.TXT is formatted as follows

DESCRIPTION	FORMAT
Make code	CHAR(3)
Make of vehicle	CHAR(32)

III.   Estimation Procedure

A. Description of Procedures

The following section describes procedures for using microdata for the estimation of descriptive statistics such as aggregates and mean. Sample programs are available online with the downloadable data files illustrate this methodology.

1. General Concepts

a. NEWID and CUID, connecting data across files

NEWID and CUID provide the identification number of each consumer unit (CU) across different PUMD files and interview waves. To connect data for one CU across different files use either variable. However they are differently constructed.

The CUID only identifies the CU. It consists of 7 digits. The first digit is a leading "blank" and is not visible in the data. The next 6 digits identify the CU.

The NEWID identifies the CU and the interview wave. It consists of 8 digits. The first seven digits are identical to the CUID and the last digit identifies the interview wave.

It is not possible to connect CUs in the Interview Survey to CUs in the Diary Survey, because the two surveys survey different CUs.

b. Sample Versus Population Estimates

As described in Section X.C. Weighting, each CU in the CE sample represents a given number of CUs in the U.S. population. The translation of sample CUs into a population estimate is accomplished by weighting. FINLWT21, one of the 45 weight variables associated with each CU, is used to estimate the population. Procedures for estimating sample (unweighted) and population (weighted) statistics are described in Sections V.A.2. Estimation of Unweigted Statistics and V.A.3.Estimation of Weighted Statistics below.

b. Calendar Period Versus Collection Period

Because the rotating panel design of the Interview survey affects the structure of the data files, one must be aware of the distinction between calendar period and collection period in producing estimates. (See Section X.A. Survey Sample Design for a description of the panel rotation scheme.)

Respondents are asked to report expenditures made since the first of the month three months prior to the interview month. For example, if a CU is interviewed in February of 2014, they are reporting expenditures for November and December of 2013, and January of 2014. This is illustrated in the rotation chart below. The period between November 1 and January 31 is referred to as the reference period for the interview.

Month of Expenditure	Month of Interview					
	January Panel A	February Panel B	March Panel C	April Panel A	May Panel B	June Panel C
October	X					
November	X	X				
December	X	X	X			
January		X	X	X		
February			X	X	X	
March				X	X	X
April					X	X
May						X

The microdata files are organized and identified by collection period, i.e., the month of the interview. Thus, the MTBI file for the second quarter of 2014 contains expenditure data collected in interviews that took place in April, May, and June of 2014. Referring to the rotation chart, one can see that this MTBI file contains expenditures made between January 2014 and May 2014. Similarly, the MTBI file for the third quarter of 2014 (interviews conducted between July and September) contains expenditures made between April and August 2014. To obtain all expenditures made in January 2014, one should access the MTBI files for both the first and second quarters of 2014. The MTBI file for the first quarter of 2014 would contain January expenditures made by CUs interviewed in February and March 2014, while the MTBI file for the second quarter of 2014 would contain January expenditures made by CUs interviewed in April 2014.

As a consequence, users should be clear as to whether they desire estimates based on when expenditures were reported (collection period) or when expenditures were made (calendar period).

To produce an annual estimate for 2014 based on collection period, that is, from all interviews conducted in 2014, data users need data only from Q141 through Q144 files. However, to produce a 2014 annual estimate based on expenditures made in 2014 (calendar period), one needs to access five collection-quarter files, the first quarter of 2014 through the first quarter of 2015. (The estimates published by BLS are based on calendar periods that require the subsequent year's first quarter data).

The ITBI files are derived in a slightly different manner than MTBI. As was mentioned in the description of the ITBI file, the data on the file represents the conversion of annual and point-of-interview data into a monthly format compatible with MTBI. Looking at a CU interviewed in January 2014, as an example, nonfarm business income earned over the previous 12 months would be collected and recorded as such on the FMLI file. For the ITBI file, this annual amount would be divided by 12, and separate records would be created for October, November, and December each containing that amount.

The variables REF_MO, REF_YR, QINTRVMO, and QINTRVYR indicate reference month of expenditure, reference year of expenditure, interview month, and interview year, respectively. REF_MO and REF_YR, in the MTBI and ITBI files, can be used to select all data for the desired period in which expenditures were made. Because of the interview rotation pattern, there is a one-month to three-month lag between the time an expenditure occurs and the time it is reported. QINTRVMO and QINTRVYR can be used to identify the collection reference period.

In addition to its effect on the selection of data prior to estimation, this distinction between collection period and calendar period also directly affects the estimation procedure for producing means. In computing means based on data collected from all CUs interviewed in a given time frame (e.g., year, quarter, 8 months), the potential contribution of each CU to the mean is the same. That is each CU can contribute data from the entire reference period to the estimate. On the other hand, in computing means based on expenditures made in a given time frame, the potential contribution of each CU to the mean varies depending on how closely the reference period for an interview coincides with the time frame desired. To see this more clearly, refer once again to the rotation chart. To compute a mean for expenditures made during the first quarter of the year, one would obtain data from CUs interviewed between February and June. However, their potential contributions to the mean are not equal. CUs interviewed in February only contribute 'one-third' of the expenditures they made during the reference period to the estimate (their January expenditures), while CUs interviewed in April contribute all their expenditures to the estimate.

As a result, the population (the denominator in the equation for a mean) has to be adjusted to account for the difference in contribution among CUs. At BLS we create a variable, MO_SCOPE, that shows the number of months a CU's interview can contribute to the mean or is "in scope" for the time period the estimate will cover. All CUs interviewed in the same month will have identical values for MO_SCOPE, as their potential contribution to the mean is the same. Thus, MO_SCOPE will be conditioned on the value of QINTRVMO (and possibly QINTRVYR).

Continuing with our example of estimating a mean for expenditures made during the first quarter of the year, we would access data from files for the first and second quarter of the year. MO_SCOPE would be derived as explained below.

    If QINTRVMO is 1 then MO_SCOPE is 0
    if QINTRVMO is 2 then MO_SCOPE is 1
    if QINTRVMO is 3 then MO_SCOPE is 2
    if QINTRVMO is 4 then MO_SCOPE is 3
    if QINTRVMO is 5 then MO_SCOPE is 2
    if QINTRVMO is 6 then MO_SCOPE is 1

Note that MO_SCOPE has a value of 0 for CUs interviewed in January, as they report expenditures for October through December, totally outside the period of interest. One could extract a data set of only CUs interviewed between February and June to eliminate that condition. How MO_SCOPE is used in estimation will be discussed later.

c. Time Period Differences

It has been mentioned previously that these files contain data that can cover a variety of time periods. Values for MTBI and ITBI variables are monthly. Values for variables on the FMLI and MEMI files can vary. For example income variables are for annual time periods and demographic variables are as of the time of interview.

This is particularly important where the user may have a choice between variables on two files that contain the same data adjusted to reflect different time periods. For instance, FMLI income data are annual covering the 12-month period prior to the collection month, whereas in ITBI these income data have been converted into monthly values. Selected demographic characteristic variables in the FMLI files contain values as of the date of interview. In the ITBI files, these values are treated as if they were "annual" amounts, and are converted to monthly records by dividing the values by 12. To illustrate each of these cases, the following example looks at a CU interviewed in April whose reference person is 60 years old at the time of interview and where CU income from wages and salaries over the previous 12 months is $48,000.

| FMLI | | | ITBI | |
VARIABLE	AMOUNT	UCC	AMOUNT	MONTH
FSALARYM	$48,000	900000	$4,000	JAN
		900000	$4,000	FEB
		900000	$4,000	MAR
AGE_REF	60	980020	5	JAN
			5	FEB
			5	MAR

Users should be aware of these time period differences when using the data.

d. Comparisons with Published CE Data

The mean values for some income and expenditure items which appear in CE publications are different than those derived from the Interview public-use microdata because some variables are topcoded or suppressed on the public-use files, but are not so treated on BLS's own data base in producing published data. For detailed topcoding information, see 2014 Topcoding and Suppression in the 2014 Documentation zip file.

2. Estimation of Unweighted Statistics

a. Aggregate Statistics

To compute unweighted aggregate expenditures from data on the MTBI files, one would sum the value of the COST field for MTBI records of interest. These records could be selected on the basis of factors such as item category, month or year of occurrence, or characteristics of the CU or its members. While MTBI is a monthly file, there is no summation done at the monthly level for each CU for expenditures with similar UCC and gift characteristics. Thus one may find multiple MTBI records with identical characteristics including COST, if the CU reported the expenditures as discrete purchases. A similar approach can be applied to estimate aggregate income from data on the ITBI files, summing the VALUE field on the appropriate records.

Certain MTBI and ITBI item categories are collected only in the 5th interview. Therefore, the data are reported by only one-fourth of the sample at any time. For some categories, the reported values have been multiplied by 4 to expand them to represent the total sample, while in other categories, this has not

been done. When estimating for these UCCs, values should be multiplied by 4 for total sample representation. (See sections on Monthly Expenditure File (MTBI) and Income File (ITBI)).

The estimation of aggregates for FMLI and MEMI file variables is similar to that for MTBI and ITBI variables. To estimate aggregates from data on the FMLI file, one would sum the value of the desired variable field for FMLI records selected on the basis of, for example, other CU characteristic variables on the FMLI file, characteristics of CU members, expenditures made, and month or year of interview. Aggregates for MEMI file variables would be developed in a similar fashion.

The user must be careful in interpreting what the aggregate represents because of the time period differences between variables on different files. For example, summing the COST field of MTBI records representing purchases for a UCC that occurred in a specific month will yield an aggregate monthly expenditure for that UCC. However, summing the value of a FMLI file variable such as FSALARYM for all CUs interviewed in a specific month will yield an aggregate annual value for that variable.

In general, one can use an aggregate derived for a certain time period to extrapolate an aggregate estimate for a longer time period. A typical case is the estimation of annual aggregates based on an aggregate using less than 12 months of data. To do this, divide the number of months for which the estimate is desired (12) by the number of months of expenditure data being used and multiply the aggregate by that quotient.

b. Means

There are two types of means that are customarily derived from CE data. The most common is the sample mean computed over all CUs. The other is the mean of those reporting computed over only those CUs actually reporting the item. The following sections look at each type of mean.

1   Sample_Means

Unweighted sample means are derived by computing an aggregate estimate for the desired item and dividing it by the sample size over the time period being estimated. Deriving an aggregate estimate has already been discussed; ascertaining the correct sample size is the next task.

The Interview survey is designed such that the CUs interviewed in each quarter represent one independent sample. Since there is one FMLI record for each sample CU, the national sample for the first quarter of 2014 is 6769 (see Section on Record Counts). The appropriate sample size for any time period will reflect the number of interviewed CUs eligible to report data over the period adjusted by the number of independent samples represented. As explained earlier, the major consideration is whether the desired estimate is a collection period estimate or a calendar period estimate.

To calculate the sample size for a collection period estimate, divide the total number of CUs interviewed by the quotient of the number of months in which these interviews occurred divided by 3. For example, one might wish to estimate the annual sample mean expenditure for men's shirts for all CUs interviewed in 2014. If one were to divide the aggregate expenditure on men's shirts from these interviews by the total number of CUs interviewed, one would get an annual sample mean about 1/4 as large as it should be, since the number of CUs interviewed represented four independent samples (one sample for each quarter of 2014). In fact, one would have derived the average quarterly sample mean rather than the annual sample mean. To get the annual sample mean, one would have to divide the total number of CUs interviewed by 4 (12 months divided by 3), thereby computing the average sample size over the year, and divide the aggregate by that amount.

As mentioned earlier, when one computes a calendar period estimate, the variable MO_SCOPE is required to adjust the sample size for the difference in potential contribution among CUs. Since one independent sample of CUs is represented in each quarter, the sum of MO_SCOPE for one quarter can be up to 3 times the independent sample (if MO_SCOPE = 3 for every CU interviewed in the quarter, the sum of MO_SCOPE would be equal 3 times the independent sample). To calculate the sample size for a

calendar period estimate, sum MO_SCOPE for the appropriate CUs and divide by 3. Note that this makes sense in those instances where MO_SCOPE does not equal 3. Referring to the example where MO_SCOPE was introduced, we can see that summing MO_SCOPE for CUs interviewed in the second quarter of the year (QINTRVMO = 4-6) would yield approximately one independent sample as CUs interviewed in June would be counted twice while CUs interviewed in April would not be counted. Dividing this amount by 3 would yield a sample size of 1/3 the independent sample. Keep in mind that only 1/3 of the expenditures reported in those interviews occurred within the time period of the aggregate being estimated. Only April data from May interviews and April-May data from June interviews would be included in the aggregate.

One can see how the computation of sample size is affected when one calculates the commonly-used annual calendar period estimate. A 2014 estimate would be based on data from interviews over five quarters. MO_SCOPE would take on the following values:

	2014				**Interview Month and Year** **2014**				
	Jan	*Feb*	*Mar*	*Apr*	*May*	*Jun*	*Jul*	*Aug*	*Sep*
MO_SCOPE	0	1	2	3	3	3	3	3	3

	2014			**2015**		**2015**
	Oct	*Nov*	*Dec*	*Jan*	*Feb*	*Mar*
MO_SCOPE	3	3	3	3	2	1

Summing MO_SCOPE for each of the five quarters and dividing by 3 would yield a value of 1/3 the independent sample for the first quarter of 2014, 2/3 the independent sample for the first quarter of 2014, and one independent sample for the second, third, and fourth quarters of 2014. Summed over the five quarters, this represents 4 independent samples, so the result should be divided by 4 to get the correct sample size of one average independent sample. Thus, the general rule in computing sample size for deriving an annual calendar period estimate is to sum MO_SCOPE over the five quarters and divide by 12.

> 2   *Means of Those Reporting*

The only difference between estimating a mean-of-those-reporting and estimating a sample mean is in selecting the appropriate CUs to use in the computation. The CUs to be used depend on the objective of the analysis. In deriving a sample mean, all sample units interviewed over the time period covered are included in the computation of sample size whether or not they reported the item being estimated. In computing a mean of those reporting, only those CUs reporting the desired item would be included. The aggregate estimate used in the numerator is the same in either case. The adjustments made for MO_SCOPE and the fact that each quarter represents one independent sample would apply in this case as well. It should be noted that means of those reporting cannot be used in all analyses in the same ways that means estimated for the U.S. population can. For example, means of those reporting specific items, such as rented dwellings, owned dwellings and other lodging, cannot be aggregated to compute means of those reporting larger categories, such as shelter. Similarly, the ratio of the mean for those reporting a specific item (e.g., rented dwellings) to the mean of those reporting an expenditure for at least one element of the larger category (e.g., shelter), cannot be interpreted as the expenditure share for those reporting either the specific item or the larger category. Proper care should be used when interpreting results computed only from those reporting an expenditure.

3. Estimation of Weighted Statistics

By applying weights when computing aggregates or means, one transforms the results from sample estimates to population estimates. There are 45 weight variables on the FMLI file, WTREP01-WTREP44

and FINLWT21. All the WTREP variables are half-sample replicate weights that should be used in variance computation. Use FINLWT21 to estimate weighted statistics for the population of CUs.

Users should follow the procedures for estimating unweighted statistics described above. When estimating weighted aggregates, the desired cost or value field should be multiplied by FINLWT21 at the CU level before summing across all appropriate records. In determining the proper sample size when computing collection period means, divide the sum of FINLWT21 for the CUs interviewed by the quotient of the number of months in which these interviews occurred divided by 3. Where calendar period means are to be estimated, multiply MO_SCOPE by FINLWT21 for each CU prior to summing and dividing by 3.

B. Description of Formulas

Expenditure items will be referred to in these descriptions, but income items can be handled similarly except where otherwise stated.

Definition of Terms:

Let

S = all CUs in the subpopulation of interest
k = item(s) of interest
q = number of months for which estimate is desired
m = number of months of interviews whose expenditures are to be used in calculating the estimate (collection period estimate)
r = number of months in which expenditures were made to be used in calculating the estimate (calendar period estimate)
j = individual CU in subpopulation S
t = month of expenditure
i = month of interview
MSC = MO_SCOPE value

Then

$E_{j,k,i}$ = 3-month expenditure by CU$_j$ on item k reported at month i interview
$E_{j,k,t}$ = monthly expenditure by CU$_j$ on item k made during month t
$W_{j,i,F21}$ = weight assigned to CU$_j$ for interview at month i
$W_{j,t,F21}$ = weight assigned to CU$_j$ for interview where CU$_j$ makes expenditure during month t

The F21 denotes FINLWT21, which is used for population estimates.

1. Aggregate Expenditures Estimates (Unweighted)

An estimate of unweighted aggregate expenditures for a collection period can be expressed as:

$_{UK} X _{(S,k)(q,m)}$ = an unweighted collection (*UK*) period estimate of aggregate expenditures (*X*) by CUs in subpopulation *S*, indexed from *j* = 1 through *n*, on item *k* over *q* months of interviews, where data collected over *m* months of interviews are used.

or

$$_{UK} X _{(S,k)(q,m)} = \left(\frac{q}{m}\right)\sum_{i=1}^{m}\left(\sum_{j=1}^{n} E_{k,j}\right)_i$$

An estimate of unweighted aggregate expenditures for a calendar period can be expressed as:

$_{UC}X_{(S,k)(q,r)}$ = an unweighted calendar (UC) period estimate of aggregate expenditures (X) by CUs in subpopulation S, indexed from j = 1 through n, on item k over q months, where expenditures made over r months are used.

or

$$_{UC}X_{(S,k)(q,r)} = \left(\frac{q}{r}\right)\sum_{t=1}^{r}\left(\sum_{j=1}^{n}E_{k,j}\right)_{t}$$

2. Sample Mean Expenditures Estimates (Unweighted)

An estimate of an unweighted mean expenditure for a collection period can be expressed as:

$_{UK}\overline{X}_{(S,k)(q,m)}$ = an unweighted collection period estimate of the mean expenditure by CUs in subpopulation S on item k over a period of q months, where data collected over m months of interviews are used.

or

$$_{UK}\overline{X}_{(S,k)(q,m)} = \left(\frac{_{UK}X_{(S,k)(q,m)}}{\dfrac{\sum_{i=1}^{m}\left(\sum_{j=1}^{n}S_{j}\right)_{i}}{\left(\dfrac{m}{3}\right)}} \right)$$

An estimate of an unweighted mean expenditure for a calendar period can be expressed as:

$_{UC}\overline{X}_{(S,k)(q,r)}$ = an unweighted calendar period estimate of the mean expenditure by CUs in subpopulation S on item k over a period of q months, where expenditures made over r months are used.

or

$$_{UC}\overline{X}_{(S,k)(q,r)} = \left(\frac{_{UC}X_{(S,k)(q,r)}}{\dfrac{\sum_{t=1}^{r+3}\left(MSC\sum_{j=1}^{n}S_{j}\right)_{t}}{r}} \right)$$

Note: For t =1, MO_SCOPE (MSC) = 0, since CUs interviewed in the first month for which the estimate is to be generated report expenditures outside the estimate period, i.e., in the previous

quarter, month, etc. For *t* = (*r+3*), MO_SCOPE = 1 since only 1 month's worth of expenditures have a chance to contribute to the calendar period of *r* months.

3. Aggregate Expenditures Estimates (Weighted)

An estimate of weighted aggregate expenditures for a collection period can be expressed as:

$_{WK}X_{(S,k)(q,m)}$ = a weighted collection (*WK*) period estimate of aggregate expenditures by CUs in subpopulation *S* on item *k* over a period of *q* months, where data collected over *m* months of interviews are used.

    or

$$_{WK}X_{(S,k)(q,m)} = \left(\frac{q}{m}\right)\sum_{i=1}^{m}\left(\sum_{j=1}^{n}(W_{j,F21}E_{k,j})\right)_i$$

An estimate of weighted aggregate expenditures for a calendar period can be expressed as:

$_{WC}X_{(S,k)(q,r)}$ = a weighted calendar (*WC*) period estimate of aggregate expenditures by CUs in subpopulation *S* on item *k* over *q* months, where expenditures made over *r* months are used.

    or

$$_{WC}X_{(S,k)(q,r)} = \left(\frac{q}{r}\right)\sum_{t=1}^{r}\left(\sum_{j=1}^{n}(W_{j,F21}E_{k,j})\right)_t$$

4. Sample Mean Expenditures Estimates (Weighted)

An estimate of a weighted mean expenditure for a collection period can be expressed as:

$_{WK}\overline{X}_{(S,k)(q,m)}$ = a weighted collection (*WK*) period estimate of the mean expenditure by CUs in subpopulation *S* on item *k* over a period of *q* months, where data collected over *m* months of interviews are used.

    or

$$_{WK}\overline{X}_{(S,k)(q,m)} = \left(\frac{_{WK}X_{(S,k)(q,m)}}{\dfrac{\sum\limits_{i=1}^{m}\left(\sum\limits_{j=1}^{n}W_{j,F21}\right)_i}{\left(\dfrac{m}{3}\right)}} \right)$$

An estimate of a weighted mean expenditure for a calendar period can be expressed as:

$_{WC}\overline{X}_{(S,k)(q,r)}$ = a weighted calendar (*WC*) period estimate of the mean expenditure by CUs in subpopulation *S* on item *k* over a period of *q* months, where expenditures made over *r* months are used.

or

$$_{WC}\overline{X}_{(S,k)(q,r)} = \left(\frac{_{WC}X_{(S,k)(q,r)}}{\dfrac{\sum_{t=1}^{r+3}\left[(MSC)\left(\sum_{j=1}^{n}W_{j,F21}\right)\right]_t}{r}} \right)$$

Note: For *t* =1, MO_SCOPE (*MSC*) = 0, since CUs interviewed in the first month for which the estimate is to be generated report expenditures outside the estimate period, i.e., in the previous quarter, month, etc. For *t* = (*r+3*), MO_SCOPE = 1 since only 1 month's worth of expenditures have a chance to contribute to the calendar period of *r* months.

IV.   Reliability Statement

A. Description of Sampling and Non-Sampling Errors

Sample surveys are subject to two types of errors, sampling and non-sampling. Sampling errors occur because observations are not taken from the entire population. The standard error, which is the accepted measure for sampling error, is an estimate of the difference between the sample data and the data that would have been obtained from a complete census. The sample estimate and its estimated standard error enable one to construct confidence intervals.

Assuming the normal distribution applies to the means of expenditures, the following statements can be made:

(1)   The chances that an estimate from a given sample would differ from a complete census figure by less than one standard error are approximately 68 out of 100.

(2)   The chances that the difference would be less than 1.6 times the standard error are approximately 90 out of 100.

(3)   The chances that the difference would be less than two times the standard error are approximately 95 out of 100.

Non-sampling errors can be attributed to many sources, such as definitional difficulties, differences in the interpretation of questions, inability or unwillingness of the respondent to provide correct information, mistakes in recording or coding the data obtained, and other errors of collection, response, processing, coverage, and estimation of missing data. The full extent of the non-sampling error is unknown. Estimates using a small number of observations are less reliable. A small amount of non-sampling error can cause a small difference to appear significant even when it is not. It is probable that the levels of estimated expenditures obtained in the Interview survey are generally lower than the "true" level due to the above factors.

B. Estimating Sampling Error

1. <u>Variance Estimation</u>

Variances can be estimated in many ways. The method illustrated below (a pseudo replication technique) is chosen because it is accurate and simple to understand. The basic idea is to construct several artificial "subsamples" from the original sample data such that the variance information of the original data is preserved in the subsamples. The subsamples (or pseudo replicates) can then be used to approximate variances for the estimates. Forty-four separate subsamples can be extracted from the data base using the replicate weight variables, WTREP01-WTREP44, associated with each CU. Note that only half of the CUs are assigned to each of the 44 replicates. The replicate weight variable contains a value greater than 0 for CUs assigned to that replicate. A value of missing is assigned to the weight variable for those CUs not included in a particular replicate.

The notation for the weighted collection period and calendar period estimates of aggregate expenditures in the section on <u>Aggregate Expenditure Estimates (Weighted)</u> does not explicitly identify the replicate as a variable because to calculate an aggregate (or mean) only FINLWT21 is used.

An estimate for the variance of an aggregate or mean estimate can be computed by generating 44 separate estimates using the 44 replicate weights and employing the standard formula for computing sample variance. To illustrate the estimation of variance, the notation must first be expanded to include the replicates explicitly.

Expenditure items will be referred to in these descriptions, but income items can be handled similarly except where otherwise stated.

Let the subscript "*a*" represent one of the 44 sets of replicate weights on the FMLI files. Following the earlier notation in the section on the <u>Description of Formulas</u>, we have:

$_{AK}X_{(S,k)(q,m),a}$ = a collection period estimate of aggregate expenditures by CUs in subpopulation S on item k over a period of q months, using data collected over m months of interviews, calculated using the weights of the ath replicate

and,

$_{AK}\overline{X}_{(S,k)(q,m),a}$ = a collection period estimate of the mean expenditure by CUs in subpopulation S on item k over a period of q months, using data collected over m months of interviews, calculated using the weights of the ath replicate

Note that an estimate using any one of the first 44 replicate weights uses only part of the expenditure data; in general: $_{AK}X_{(S,k)(q,m),1}, \cdots, _{AK}X_{(S,k)(q,m),44} \neq _{WK}X_{(S,k)(q,m)}$

Using standard variance formula, the variance of aggregate expenditures can be estimated as follows:

$$V\left(_{WK}X_{(S,k)(q,m)}\right) = \frac{1}{44}\sum_{a=1}^{44}\left(_{AK}X_{(S,k)(q,m),a} - _{WK}X_{(S,k)(q,m)}\right)^2$$

Similarly, estimates for the variances of $_{WK}\overline{X}_{(S,k)(q,m)}$ can be given as:

$$V\left(_{WK}\overline{X}_{(S,k)(q,m)}\right) = \frac{1}{44}\sum_{a=1}^{44}\left(_{AK}\overline{X}_{(S,k)(q,m),a} - _{WK}\overline{X}_{(S,k)(q,m)}\right)^2$$

2. Standard Error of the Mean

The standard error of the mean, $S.E.(\bar{X})$, is used to obtain confidence intervals that evaluate how close the estimate may be to the true population mean. $S.E.(\bar{X})$ is defined as the square root of the variance of the mean. For example, if the estimated mean expenditure for alcoholic beverages is $445 and its standard error $17.34, then we can construct a 95 percent confidence interval. This interval would be bounded by 1.96 times the standard error less than and greater than the estimate, $410.32 to $479.68 respectively. We could conclude with 95 percent confidence that the true mean expenditure for alcoholic beverages lies within the interval $410.32 to $479.68**.**

3. Standard Error of the Difference Between Two Means

Standard errors may also be used to perform hypothesis testing, a procedure that evaluates population parameters using sample estimates. The most common types of hypotheses are: 1) the population parameters are identical, and 2) they are different.

For example, the 2014 mean expenditure estimate for alcoholic beverages for CUs in the *25 to 34 years* age range is $529 and the estimate for CUs in the *35-44 years* age range is $505. The apparent difference between the two mean expenditures is $24. The standard error on the estimate of $529 is $49.01 and the estimated standard error for $505 is $33.30.

The standard error of a difference is approximately equal to

$$S.E.\left(_{WC}\bar{X}_1, _{WC}\bar{X}_2\right) = \sqrt{\left(V\left(_{WC}\bar{X}_1\right) + V\left(_{WC}\bar{X}_2\right)\right)} \tag{1}$$

where

$$V(\bar{X}_i) = \left(S.E.(\bar{X}_i)\right)^2$$

This assumes the two sample means, $_{WC}\bar{X}_1$ and $_{WC}\bar{X}_2$, are disjoint subsets of the population. Hence the standard error of the difference in rent expenditures between these two age groups is about

$$\sqrt{\left((49.01)^2 + (33.30)^2\right)} = 59.25 \tag{2}$$

This means that the 95 percent confidence interval around the difference is from -$94.05 to $142.5. Since this interval does include zero, we cannot conclude with 95 percent confidence that the mean alcoholic beverages expenditure for CUs in the *25-34 years* age range is different than the mean alcoholic beverages expenditure for CUs in the *35-44 years* age range.

Analyses of the difference between two estimates can also be performed on non-disjoint sets of population, where one is a subset of the other. The formula for computing the standard error of the difference between two non-disjoint estimates is

$$SE\left(_w\bar{X}_1 - _w\bar{X}_2\right) = \sqrt{V\left(_w\bar{X}_1\right) + V\left(_w\bar{X}_2\right) - 2\rho \cdot SE\left(_w\bar{X}_1\right) \cdot SE\left(_w\bar{X}_2\right)} \tag{3}$$

where

$$V(\bar{X}_i) = \left(S.E.(\bar{X}_i)\right)^2$$

and where r is the correlation coefficient between $_W\overline{X}_1$ and $_W\overline{X}_2$. The correlation coefficient is generally no greater than 0.2 for CE estimates.

V.   Sample programs

CE provides sample code to approximate the published tables presented by income groups. The code is available in SAS and R. The variables and ranges referred to in the program are described in the interview data dictionary. The dictionary and the code are available on the PUMD home page in the section Most Recent Data Release.

The results of the sample code may differ from the published tables due to topcoding of data and CE publication programming methodology. CE provides the programs to illustrate the estimation methodology.

VI.   Description of the Survey

The CE program consists of two separate components, each with its own questionnaire and independent sample:

1) An Interview panel survey in which each CU in the sample is interviewed once every 3 months over five consecutive quarters to obtain a year's worth of data. New panels are initiated every month of the year.

2) A Diary or recordkeeping survey completed by the sample CUs for two consecutive 1-week periods; the sample is surveyed across a 12-month period.

Data are collected by the Bureau of the Census under contract with BLS. All data collected in both surveys are subject to Bureau of the Census confidentiality requirements, which prevent the disclosure of any CU member's identity.

The quarterly Interview survey is designed to collect data on major items of expense which respondents can be expected to recall for 3 months or longer. In practice, the Interview survey collects detailed data on an estimated 60 to 70 percent of total household expenditures. In addition, global estimates are obtained for food and other selected items. These global estimates account for an additional 20 to 25 percent of total expenditures. The Interview survey does not collect expenses for housekeeping supplies, personal care products, and nonprescription drugs, which contribute about 5 to 15 percent of total expenditures. Thus, up to 95 percent of total expenditures are covered in the Interview survey. Household characteristics, income, and financial data are also collected. At BLS, each quarter of data is processed independently from other quarters. Thus the annual estimates published by BLS are not dependent on the participation of a CU for the full five interviews.

The initial interview collects demographic and family characteristics data. These pertain to age, sex, race, marital status, education, and CU relationship for each CU member. This information is updated at each subsequent interview. Expenditures are for the month prior to the interview. They are used along with the inventory information solely for bounding purposes, that is, to prevent the reporting of expenditures from an indefinite past period. Expenditure data from the first interview are not on these files since they are not included in expenditure estimation.

The second through fifth interviews use uniform questionnaires to collect expenditure information from the previous three months. Income information, such as wage, salary, unemployment compensation, child

support, and alimony, as well as information on the employment of each CU member age 14 and over, are collected in the second and fifth interviews only.

Income data and employment information collected in the second interview are carried over to the third and fourth interviews. For new CU members and CU members who started work since the previous interview, wage, salary, and other information on employment are collected in the third and fourth interviews. In the fifth interview, a supplement is used to collect information on asset values and changes in balances of assets and liabilities. These data, along with other household characteristics information, permit users to classify sample units for research purposes and allow BLS to adjust population weights for CUs who do not cooperate in the survey.

Each quarter, 20 percent of the sample are new households introduced for the first time. They replace one-fifth of the sample that completed its final interview in the previous quarter. This rotating procedure with overlap is designed to provide more efficient data collection. CUs that move away from their sample address between interviews are dropped from the survey. New CUs that move into the sample address are screened for eligibility and included in the survey. Students living in college- or university-regulated housing report their own expenditures directly, while at school, rather than being considered part of their parents' household.

VII.   Data Collection and Processing

In addition to its data collection duties, the Bureau of the Census is responsible for field editing and coding, consistency checking, quality control, and data transmittal to BLS. BLS performs additional review and editing procedures in preparing the data for publication and release.

A. The US Census Bureau Activities

Data collection activities have been conducted by the U.S. Census Bureau on a continuing basis since October 1979. Due to differences in format and design, the Interview survey and the Diary survey data are collected and processed separately.

All interviews are sent electronically to the U.S. Census Bureau headquarters in Suitland, MD, where they pass through basic quality checks of control counts, missing values, etc. Also, missing sections of questionnaires, and certain inconsistencies and errors are identified and corrected. The data are then electronically transmitted to BLS in Washington, DC.

An input file is created by the U.S. Census Bureau when the data are electronically sent to BLS. The input file is used in the next quarter's interview to prevent the recording of duplicate reports by respondents. The input file also contains data collected in the first interview about owned property, vehicles, and insurance policies. Because the input file contains this data, only updates and new records are collected about owned property, vehicles, and insurance policies in the second through fifth interviews.

B. Bureau of Labor Statistics Activities

Upon receipt from the Bureau of the Census, the data undergo a series of computer edits that identify and correct irregularities and inconsistencies. Other adjustments eliminate business and reimbursed expenses, apply appropriate sales taxes, adjust mortgages and derive CU weights based on BLS specifications. In addition, demographic and work experience items are imputed when missing or invalid. All data changes and imputations are identified with flags on the Interview data base.

Next, BLS conducts an extensive review to ensure that severe data aberrations are corrected. The review takes place in several stages: a review of counts, weighted means, and unweighted means by

region; a review of family relationship coding inconsistencies; a review of selected extreme values for expenditure and income categories; and a verification of the various data transformations.

Cases of extreme data values are investigated. Any errors discovered are corrected prior to release of the data.

Two major types of data adjustment routines--imputation and allocation--are carried out to classify expenditures and improve estimates. Data imputation routines correct for missing or invalid entries. All fields except assets are subject to imputation. Allocation routines are applied when respondents provide insufficient expenditure detail to meet tabulation requirements. For example, reports of combined expenditures for fuels and utilities are allocated among gas, electricity, and other items in this group. While not strictly an allocation routine, another adjustment separates mortgage and vehicle loan payments into principal and interest components using associated data on the interest rate and term of the loan. Another adjustment is done to prepare the data for the production of calendar year estimates. Time adjustment routines are used to classify expenditures by month. Aggregation can then be done at a monthly level, permitting the production of monthly, quarterly, annual, and other interval estimates. To analyze the effects of these adjustments, tabulations are made before and after the data adjustments. At this point, processing activities are completed and the database is ready for use.

1. Sales Tax

All expenditures are presented with state, county, and local sales taxes that were levied in the state in which the respondent resides. Generally, CE estimates the sales taxes for each expenditure based on current tax rates. CE obtains the tax rates annually from the respective "Department of Taxation and Revenue" websites of each State and the District of Columbia.

Most expenditures are reported without sales taxes and CE estimates sales taxes in two steps. First, CE matches expenditures with the tax rate of the item in the state in which the respondent resides. CE also applies the tax rate of the state of residence when an item was purchased from an out-of-state vendor. Second, CE calculates the sales taxes by multiplying the item cost by the specific state's sales tax rate. CE calculates the total cost of each expenditures by adding the calculated sales tax to the reported item cost. However, for some expenditures, respondents include sales taxes. For example for food away from home, CE specifically asks respondents to include sales taxes.

2. Mortgage Edits

CE estimates monthly values for mortgage related variables, which are primarily the current principal and interest payment variables, because these values are often inaccurately reported. CE calculates these estimates using the reported current mortgage payment, the original mortgage value, the interest rate, the mortgage term, the time of origination, and the loan type.

Generally, CE estimates the values related to original mortgages in two steps: First, CE determines the current principal and interest payment. Second, CE checks their validity. If they do not pass validity tests, CE adjusts the amounts of the escrow, the original or current principal, or the interest payment to estimate reasonable results. Escrow may consist of property taxes and insurance payments. For refinanced mortgages, monthly payments are calculated separately before and after the refinance month. The mortgage payments for properties that have been sold or gifted during the reference period are only calculated up to the month of disposal.

3. Income Tax Estimation

The CE began to estimate federal, state, and local income tax liabilities with the 2013 data (2013Q2). The CE only estimates the income taxes if our methods have determined that the tax unit (Groups of individuals, who file together) should have paid taxes. CE prepares the income tax data with a program called TAXSIM in three steps:

1. Assembles the inputs for the program, which are income and demographic data, tax codes, and tax units.
2. Feeds the inputs into the TAXSIM program and estimate the income tax liabilities. TAXSIM calculates taxes as if the respondents where filing their tax returns using the information they report to the CE with the 1040 Individual Income Tax Return Form and supplementary forms.
3. Creates quarterly tax files (NTAXI) with all inputs and outputs. The main outputs are total income taxes due and total payments[4] by Tax Units. In addition, NTAXI contains NEWIDs to allow users link NTAXI data with other PUMD files.

To calculate income tax data prior to 2013, see the TAXSIM Related Files at the NBER.

a. Background of the NBER Tax Estimation Tool

CE estimates the income tax data with a Fortan program called TAXSIM, which was developed by Amy Taylor in 1976. The National Bureau of Economic Research updates it annually to reflect the changes in the U.S. tax code. For more information, see the TAXSIM Related Files at the NBER.

CE began to use TAXSIM to estimate income tax data because the reported income tax data were not accurate enough for economic analysis. The accuracy issues were due to three concerns:

- Respondents fail to recall their tax liability or refuse to share it.
- Respondents tend to have a better idea of their total taxes paid when they file their tax return rather than when they respond to the CE survey.
- The CE collects income tax data with a changing time frame, which leads to confusion among respondents and inaccuracies within the data.

b. What is a Tax Unit?

Tax Units are groups of individuals, who file one tax returns, rather than Consumer Units, groups of individuals who live together and share in certain expenses. Each member of a CU is placed in a Tax Unit as the tax head, spouse, or dependent.

c. Input data

Data are entered into TAXSIM by tax unit. Each tax unit is recorded on one data row. Data collected at the CU level is used to calculate the TAXSIM input variables only for the primary tax unit except where specified. Data collected at the member level in CE is used to calculate the TAXSIM input variables for the tax unit containing that member. This member-specific information is used for both the primary and secondary units.

The income taxes are based upon a full 12 months (1 year) from the month of the interview. As such, the span of the 12 months includes months from both the current and the previous year. To accurately account for both years, the tax liability created for both current and previous years will be weighted based upon the months split.

The TAXSIM program requires 22 input variables to calculate a value. "Blanks" and tax information that the CE survey does not capture (i.e. Unemployment Compensation, Short term capital gains or losses, Long term capital gains or losses, Dividend income) will be set to zero.

d. State identifiers in NTAXI

The value of the variable SOI_ST identifies the state of residence of CUs in NTAXI. Some state information must be suppressed to meet the non-disclosure requirements by CENSUS. As a result, no

---

[4] Total payments refer only to the Earned Income Credit and the Additional Child Tax Credit. These are the only tax credits that are relevant to calculate after-tax income.

state tax estimates will be generated in the PUMD for those observations with suppressed state code. For more information on the nondisclosure requirements related to the state in which the respondent resides, see Topcoding and Suppression 2014 in the 2014 Documentation zip file. For a list of the state codes, see the Interview Dictionary.

e. Federal and State Variable Calculation

Weight the prior year and current year estimates according to the interview month:

Find weighting factors using QINTRVMO:
Current year weight (CYWT):   (QINTRVMO – 1)/12
Previous year weight (PYWT): (12 – [QINTRVMO-1])/12

Calculate weighted estimate for tax liabilities for each Tax Unit:
$FTAXOWE_n$ = CYWT*FTAXO_CY$_n$ + PYWT*FTAXO_PY$_n$, where n = nth tax unit in the CU
$STAXOWE_n$ = CYWT*STAXO_CY$_n$ + PYWT*STAXO_PY$_n$ where n = nth tax unit in the CU

For example for 2013, these were the figures:
CU interviewed in May 2013, finding Federal taxes owed for Tax Unit$_1$:

They have paid:
8 months of 2013 taxes (May13 – Dec13)
4 months of 2014 taxes (Jan14 – April14)

Using calculated taxes:
Annual estimated tax liability in 2013: FTAXOWE_PY$_1$ = \$3000
Annual estimated tax liability in 2014: FTAXOWE_CY$_1$ = \$3400

Calculate weighted tax amount:
(8/12) * \$3,000 + (4/12) * \$3,400 = \$3,133

VIII.   Sampling Statement

B. Survey Sample Design

Samples for the CE are national probability samples of households designed to be representative of the total U.S. civilian population. Eligible population includes all civilian non-institutional persons.

The first step in sampling is the selection of primary sampling units (PSUs), which consist of counties (or parts thereof) or groups of counties. The set of sample PSUs used for the 2014 and 2015 samples is composed of 91 areas. The design classifies the PSUs into four categories:

- 21 "A" certainty PSUs are Metropolitan Statistical Areas (MSA's) with a population greater than 1.5 million.
- 38 "X" PSUs, are medium-sized MSA's.
- 16 "Y" PSUs are nonmetropolitan areas that are included in the CPI.
- 16 "Z" PSUs are nonmetropolitan areas where only the urban population data will be included in the CPI.

The sampling frame (that is, the list from which housing units were chosen) for the 2014 survey is generated from the 2000 Census of Population 100-percent-detail file. The sampling frame is augmented by new construction permits and by techniques used to eliminate recognized deficiencies in census coverage. All Enumeration Districts (EDs) from the Census that fail to meet the criterion for good

addresses for new construction, and all EDs in nonpermit-issuing areas are grouped into the area segment frame. Interviewers are then assigned to list these areas before a sample is drawn.

To the extent possible, an unclustered sample of units is selected within each PSU. This lack of clustering is desirable because the sample size of the Diary Survey is small relative to other surveys, while the intraclass correlations for expenditure characteristics are relatively large. This suggests that any clustering of the sample units could result in an unacceptable increase in the within-PSU variance and, as a result, the total variance.

The Interview Survey is a panel rotation survey. Each panel is interviewed for five consecutive quarters and then dropped from the survey. As one panel leaves the survey, a new panel is introduced. Approximately 20 percent of the addresses are new to the survey each month.

C. Weighting

Each CU included in the CE represents a given number of CUs in the U.S. population, which is considered to be the universe. The translation of sample families into the universe of families is known as weighting. However, since the unit of analysis for the CE is a CU, the weighting is performed at the CU level. Several factors are involved in determining the weight for each CU for which an interview is obtained. There are four steps in the weighting procedure:

1) The basic weight is assigned to an address and is the inverse of the probability of selection of the housing unit.

2) A weight control factor is applied to each interview if sub-sampling is performed in the field.

3) A non-interview adjustment is made for units where data could not be collected from occupied housing units. The adjustment is performed as a function of region, housing tenure, family size and race.

4) A final adjustment is performed to adjust the sample estimates to national population controls derived from the Current Population Survey. The adjustments are made based on both the CU's member composition and the CU as a whole. The weight for the CU is adjusted for individuals within the CU to meet the controls for 14 age/race categories, 4 regions, and 4 region/urban categories. The CU weight is also adjusted to meet the control for total number of CUs and total number of CUs who own their living quarters. The weighting procedure uses an iterative process to ensure that the sample estimates meet all the population controls.

NOTE: The weight for a consumer unit (CU) can be different for each quarter in which the CU participates in the survey, as the CU may represent a different number of CUs with similar characteristics.

IX. Interpreting the Data

Several factors should be considered when interpreting the expenditure data. The average expenditure for an item may be considerably lower than the expenditure by those CUs that purchased the item. The less frequently an item is purchased, the greater the difference between the average for all CUs and the average of those purchasing. (See Section V.A.2.b.ii. Means of Those Reporting.) Also, an individual CU may spend more or less than the average, depending on its particular characteristics. Factors such as income, age of family members, geographic location, taste and personal preference influence expenditures. Furthermore, even within groups with similar characteristics, the distribution of expenditures varies substantially.

Expenditures reported are the direct out-of-pocket expenditures. Indirect expenditures, which may be significant, may be reflected elsewhere. For example, rental contracts often include utilities. Renters with

such contracts would record no direct expense for utilities, and therefore, appear to have lower utility expenses. Employers or insurance companies frequently pay other costs. CU with members whose employers pay for all or part of their health insurance or life insurance would have lower direct expenses for these items than those who pay the entire amount themselves. These points should be considered when relating reported averages to individual circumstances.

X.   Appendix 1—Glossary

Population
> The civilian non-institutional population of the United States as well as that portion of the institutional population living in the following group quarters: Boarding houses, housing facilities for students and workers, staff units in hospitals and homes for the aged, infirm, or needy, permanent living quarters in hotels and motels, and mobile home parks. Urban population is defined as all persons living in a Metropolitan Statistical Area (MSA's) and in urbanized areas and urban places of 2,500 or more persons outside of MSA's. Urban, defined in this survey, includes the rural populations within MSA. The general concept of an MSA is one of a large population nucleus together with adjacent communities that have a high degree of economic and social integration with that nucleus. Rural population is defined as all persons living outside of an MSA and within an area with less than 2,500 persons.

Consumer unit (CU)
> A consumer unit comprises either: (1) all members of a particular household who are related by blood, marriage, adoption, or other legal arrangements; (2) a person living alone or sharing a household with others or living as a roomer in a private home or lodging house or in permanent living quarters in a hotel or motel, but who is financially independent; or (3) two or more persons living together who use their income to make joint expenditures. Financial independence is determined by the three major expense categories: housing, food, and other living expenses. To be considered financially independent, at least two of the three major expense categories have to be provided entirely or in part by the respondent.

Reference person
> The first member mentioned by the respondent when asked to "Start with the name of the person or one of the persons who owns or rents the home." It is with respect to this person that the relationship of other CU members is determined.

Income before taxes
> The combined income earned by all CU members 14 years old or over during the 12 months preceding the interview. The components of income are: Wage and salary income, business income, farm income, Social Security income and Supplemental Security income, unemployment compensation, workmen's compensation, public assistance, welfare, interest, dividends, pension income, income from roomers or boarders, other rental income, income from regular contributions, other income, and food stamps.

Income after taxes
> Income before taxes minus personal taxes, which includes Federal income taxes, state and local taxes, and other taxes.

Geographic regions
> CUs are classified by region according to the address at which they reside during the time of participation in the survey. The regions comprise the following States:

> *Northeast* - Connecticut, Maine, Massachusetts, New Hampshire, New Jersey, New York, Pennsylvania, Rhode Island, and Vermont

Midwest - Illinois, Indiana, Iowa, Kansas, Michigan, Minnesota, Missouri, Nebraska, North Dakota, Ohio, South Dakota, and Wisconsin

South - Alabama, Arkansas, Delaware, District of Columbia, Florida, Georgia, Kentucky, Louisiana, Maryland, Mississippi, North Carolina, Oklahoma, South Carolina, Tennessee, Texas, Virginia, and West Virginia

West - Alaska, Arizona, California, Colorado, Hawaii, Idaho, Montana, Nevada, New Mexico, Oregon, Utah, Washington, and Wyoming

Tax Unit
> A tax unit is a collection of members within a CU who file their tax return together, and thus the taxes incurred apply to the tax unit as a whole. There may be multiple tax units within a single consumer unit.

XI.   Appendix 3—Publications and Data Releases from the Consumer Expenditure Survey

CDs and Free Online Data
PUMD are available for 1996 forward at the [PUMD website](#) and for pre-1996 data on CD for purchase at [PUMD on CD](#).

For information and downloading of past PUMD releases, please visit the links below. Multiple zip files can also be downloaded at one time. Please see [Instructions for Downloading Consumer Expenditure Survey (CE) Microdata and Documentation](#) for information on downloading the files.

PUMD on CD are available from the Bureau of Labor Statistics for selected years: 1972-73, 1980-81, 1990-91, 1992-93, and 1993-1995. However, their content differs between years. After 1980-81, the CDs contain Interview and Diary data, while the 1972-73 CD includes Interview data only. The 1980-81 and the 1990 data in the 1990-91 CD include selected EXPN data, while the 1991 files in the 1990-91 CD and the 1992-93 CD do not. The CDs from 1994 and 1995 include Interview and Diary data as well as all EXPN files. A 1984-94 "multi-year" CD that presents Interview FMLI file data is also available. In addition to the microdata, the CDs also contain the same integrated Diary and Interview tabulated data (1984-2009) that are found on the Consumer Expenditure Survey web site (http://www.bls.gov/cex).

XII.   Inquiries, Suggestions and Comments

If you have any questions, suggestions, or comments about the survey, the microdata, or its documentation, please call (202) 691-6900 or email cexinfo@bls.gov.

Written suggestions and comments should be forwarded to:

> Division of Consumer Expenditure Survey
> Branch of Information and Analysis
> Bureau of Labor Statistics, Room 3985
> 2 Massachusetts Ave. N.E.
> Washington, DC. 20212-0001

The Bureau of Labor Statistics will use these responses in planning future releases of the microdata.