Developing a Core Classification System for U.S. Personal Consumption Data

Prepared by Dale A. Smith and Mary Lynn Schmidt
U.S. Bureau of Labor Statistics

Introduction

In the U.S. and throughout the world it is common to use standardized classification systems such as the Standard Industrial Classification (SIC), the International Standard Industrial Classification System (ISIC), or the North American Industry Classification (NAICS) for the coding and classification data relating to industrial production and/or enterprises. In the same way, the Classification of Individual Consumption by Purpose (COICOP) as part of the 1993 System of National Accounts (SNA) is widely used throughout the world for the classification and coding of personal consumption-related data associated with consumer expenditure surveys, personal consumption estimates for the national accounts and Consumer Price Indices (CPIs). However, neither COICOP, nor any other official standardized classification system, is presently used for the purpose of classifying and coding these official databases and statistical series in the United States.

After providing some background information on the current classification and coding structures in place for personal consumption-related data in the U.S., the concept of a Core Classification System (CCS) is presented. In generic terms, a CCS is used for standardizing the coding and classification of conceptually similar statistical data that are obtained from various sources, by various entities, and employed in diverse uses. Since personal consumption data in the U.S. are characterized by these conditions, it is natural to think of using a CCS to organize and standardize these data.

In order to illustrate the need for standardization, we present an example of the detailed classification and coding structures of several food components compared across the CE, the CPI and the PCE. This example serves to highlight a number of issues regarding the need for a CCS.

Diverse classification and coding systems presently used for the preparation of the CE, the CPI, and the PCE

For a number of years the Bureau of Labor Statistics (BLS) has had the goal of standardizing the classification and coding of personal consumption expenditures as used in the preparation of the CPI and CE publications. The creation of the Universal Classification Code (UCC) in the late 1970s represented a significant step toward the achievement of this goal. However, in spite of the existence and use of the UCC for the past 20 years, a number of different coding and classification systems (both explicit and implicit) continue to be used for survey form design, data collection, data processing and publication of the CPI and the CE results. In addition to the standardization of the coding and classification systems used for processing CE data within and between the BLS and the Census Bureau, a major issue also exists regarding standardization between the BLS/Census systems and the coding and classification systems used for processing and publishing personal consumption expenditures (PCE) and the implicit deflator by the Bureau of Economic Analysis (BEA).

While the UCC is the backbone of the current U.S. classification system for the CE and the CPI, there are a number of other coding systems, which are presently being used in the production of the CE and CPI at various intermediate stages. The current UCC is a 6 digit coding system that reflects the CPI sampling and publication structure for the 1977 CPI revision. The first two digits correspond to the Expenditure Class (EC) structure which reflects the classification of expenditures for the CPI market basket according to type of expenditure, i. e., bakery products, cereals, dairy products, etc. The next two digits represent the Item Strata for the CPI sampling frame/weighting structure/publication structure. These classifications by Strata are basically determined by the size of the expenditure or the weight in the index. In other words the Item Strata are roughly equivalent in weight except where a given item has a large expenditure weight in and of itself. The fifth digit represents the ELI (Entry Level Item). For the U.S. CPI the ELI's are generally rather broad categories of expenditures. For example, Men's Shirts would include all types of shirts such as dress shirts, knit shirts, sport shirts, T-shirts, etc. There are usually no more than a few ELI's in a given Item Stratum. For the CPI the sixth digit is used for the cluster classification below the ELI level when necessary. For the CE the sixth digit is also used for further disaggregation, but

it may be different form that used for ELI clusters in the CPI.

In addition to the UCC there are several other coding systems used for the production of the CE and the CPI. For the Diary portion of the CE a separate 6 digit coding system is used. The first digit designates the major group of expenditure such as food, housing, clothing, medical care, etc. The second digit is used to designate sub-groups such as meat, poultry and fish or furni-ture, etc. The third digit corresponds to the CPI expenditure classes in at least certain cases, although there is no completely consistent mapping between EC's and the first three digit groups in the Census diary codes. The fourth and fifth digits are used to identify specific items. The sixth digit is used in various ways. For food it is used to designate whether an item is fresh, frozen, canned or other. For clothing it is used to designate whether the item is for men, women, boys, girls or infants.

A completely separate system is used for mapping the CE interview data into the UCC codes. The CE Interview Survey is conducted using a necessarily complicated form, which also serves as the instrument for collecting a great deal of collateral non-expenditure data. In particular the CE provides data on household demographics, labor force participation, income, housing characteristics, inventories of vehicles and household equipment, credit, insurance, gifts, savings, etc., in addition to the simple expend-iture data. For certain types of expenditure data the format for collecting the data is also complex. For example, for clothing there is a list of items with corresponding 3 digit codes. Expenditures for these items are recorded in a generic field along with the code for the item. In addition there is also a field for entering the name of the individual in the consumer unit for which the item was purchased and the line number on which this individual is listed during the initial phase of the survey. This means that in order to compute tabulation for men's shirts it is necessary to crosstabulate clothing expenditures for the 3 digit code for shirts and all the individual numbers which are classified as men.

In order to handle the complex organization of the U.S. CE data collection form it is necessary to have an elaborate system of codes and mapping documents. Since Census collects and processes the data, they assign processing codes and develop a data base dictionary. This allows them to do the edits and preliminary tabulations. The data base dictionary consists of a series of variable names and a mapping from these variables to the specific line item responses on the data collection form At BLS a new data base dictionary is defined that has some differences with the Census database dictionary. Also, an elaborate mapping document is prepared for documenting how the expenditure data are derived from the collection instrument and mapped into the specific UCC categories.

The UCC-classified data are then used for the construction of the CPI market basket weights and the preparation of the line item totals for the CE publication structure.

At the BEA a separate structure is used for classifying, coding and aggregating personal consumption data for the national accounts. These items are more oriented toward sources of production rather than purposes of consumption. Aggregations of the data also differ significantly for certain types of consumption and are quite similar for others such as food.

Specification of a Core Classification System

Since personal consumption data in the U.S. is collected from various sources and is used in diverse ways by different government agencies, there is a critical need for standardization at some level for these data. The first step in accomplish-
ing this goal is the development of what will be termed the Core Classification System (CCS) that would serve as a benchmark for all classifi-cation and coding (both explicit and implicit) of all CE, CPI and PCE data from the data collection stage to publication of results and official series.

The CCS should have the following desirable characteristics:

1. Natural interpretation for the end use of the data—publication and analysis—for the CPI, CE and PCE. 2. Natural interpretation for relevant survey forms and data collection efforts—CE and Point of Purchase Surveys (TPOPS). 3. Suffi-cient detail so that the CCS-coded expenditure data can be mapped unambiguously into other relevant classification systems such as: UCC (EC's, strata, ELI's, etc.), CPI market basket and publication format, EC publication format, COICOP, NAICS, ISIC, CPC, PCE

classification structure, etc. The CCS will probably be more detailed than anything it is mapped into. It will also probably be more detailed/complete than parts of the data collection forms since it must anticipate new "variables" or expenditure aggregates that will be created in the database.
4. Branching that is intuitive, balanced and consistent across major groups, subgroups, classes, strata, etc. a. Intuitive—the BCS should seem natural. If the classification system is not intuitive, it will not survive over time. People will always be tempted to use a system that is more convenient. b. Balanced—the n (th) digit levels of aggregation across groups, subgroups, etc. should have similar interpretations and reasonable relative budget shares. E.g., using an extra digit to divide total expenditures into one group having a budget share of 2 percent and another group having a budget share of 98 percent is awkward at best. c. Consistent: i. The classification hierarchy should allow the same number of digits for similar levels of aggregation across groups, subgroups, etc. If subgroups for Food are represented by 2 digits, subgroups for clothing should be represented by 2 digits as well. ii. All codes ending in "9" could designate "other" for example. iii. Codes ending in "0" refer to a level of aggregation. E.g., "34000" is some level of aggregation; "34001 is an item within the classification "34000"; etc.

The codes should be easy to work with and easy to remember. In general, alphabetic codes are confusing. They are perhaps OK for designating major groups, but they should be avoided for subgroups, etc.

Following are suggestions for implementation of the production of the CPI, CE and PCE using the CCS:

1. To the extent possible the structure of the data collection instruments (survey forms for the diary and recall portions of the CE and the TPOPS, etc.) should reflect the structure of the CCS. Where possible, the numbering of the sections and subsections should be the same as the corresponding numbers in the CCS codes. 2. All coding for processing and tabulation should correspond to the CCS. The database dictionary should reflect the CCS coding system. The number codes and the names assigned to each line item of the survey tabulation should be consistent with the number codes and names assigned to the codes of the CCS. The use of abbreviated names should be avoided. It is understood that certain databases require abbreviated names. 3. All explicit coding and implicit coding in publication formats should reflect and be reflected in the CCS. If new aggregations are developed, new CCS codes should be created. 4. Consideration should be given to making the CCS consistent with, and/or easily mapable to, COICOP.

Developing a national classification system based on an international classification system

The use of an international classification system by any country requires the development of an extended version of that system that is consistent with the international system and meets the needs of the specific country. A classification system is extended by the use of additional digits to account for the specific needs of the country. For example, COICOP-HBS, a version of COICOP that was designed for use with household budget surveys, has five digits. If tabulations for the CE, CPI and PCE were available at the five-digit level of COICOP-HBS, the requirement for consistency with COICOP is met. However, more digits would be needed to account for the detail associated with the processing of primitive data and to provide flexibility for tabulation and publication needs.

To illustrate the process of extending an international classification system for use in a given country we reference work that has been done by BLS[1] to map detailed CE diary data into COICOP-HBS. Table 2 presents a sample of the results of this work. It will be noted that the six-digit CE Diary item codes were recoded using a nine-digit extended COICOP coding system. The additional four digits of this nine-digit extended COICOP coding system provide sufficient detail for the preservation of the current mapping of these diary codes to UCC codes for use in publishing the CE results and for use in estimating weights for the CPI. At the same time, the first five digits are used to present the structure for aggregating the detailed CE diary data to the five-digit level of COICOP-HBS. It should be noted that COICOP-HBS is used extensively throughout the world for consumer expenditure surveys and consumer price indices.

[1] Smith, Dale A., "CE Diary Item Codes Mapped to Nine-digit Extended COICOP-HBS," preliminary version, May 2000.

For personal consumption data for the national accounts most countries use the traditional four-digit version of COICOP recommended in the 1993 SNA.

. In general, many-to-one mappings of primary data to different classification system do not provide a link between the two target classification systems. It is not enough to simply have a number of concordances available. The mappings generally do not have inverses and are generally not transitive.

Consistency of classification systems for production and consumption

With the introduction of NAICS there is an interest in developing a product classification system that is consistent with the economic activities specified by NAICS. This new product classification system would be similar to the Classification of Products by Economic Activity (CPA) that is related to the economic activities specified by the General Industrial Classification of Economic Activities within the European Communities (NACE).

Interest has also been expressed in using this new system (being developed for classifying production goods) for personal consumption data as well. However, personal consumption data is more naturally classified by purpose of consumption than by considerations relating to production. For example, if production data for clothing were classified by the type of inputs and/or process of production, knit goods such as sweaters and underwear might well be combined with knit shirts. On the other hand, if purpose of consumption were the standard of classification, knit shirts would be combined with shirts made of woven fabric. The solution, of course, is to develop both a production-based and a consumption-based classification system with a CCS for each that takes into account the other.

Comparison of the Dairy Products and the Fats and Oils components of the CE, CPI and PCE

It is reasonable to assume that it would be straightforward to compare major food components across the CE, the CPI and the PCE. However, Table 1 illustrates that this is not the case for the Dairy Products and the Fats and Oils components of these series. The classification and coding structure of these components for the CE, the CPI and the PCE along with the corresponding structures for the UCC and COICOP-HBS are included. In order to demonstrate the differences in these components across the different series it is necessary to use a very detailed item structure. Since two of the series being considered are based on BLS data, it is convenient to carry out the analysis at the level of the detailed CE Diary items rather than the detailed PCE items.

In Table 2 the detailed CE Diary items found in the last three columns of Table 1 are classified into five-digit COICOP-HBS Categories. It will be noted that for certain items the classification structure of COICOP is quite different from those of the PCE, the CE and the CPI, even at the four-digit Class level. For example, salad dressings and even mayonnaise (not separated out for the CE at any level) are classified with the five-digit Category Sauces, Condiments under the four-digit Class Food Products n.e.c. rather than with Fats and Oils.

Conclusions

1. The internationally recommended COICOP classification system is not presently employed for any official U.S. statistics relating to personal consumption.

2. Not only is it difficult to compare U.S. personal consumption-related data and statistical series with corresponding data from other countries, it is also difficult to compare personal consumption-related data across the PCE, the CE and the CPI.

3. While a limited amount of work has been done to map various classification systems, relating to personal consumption, to one another and to COICOP, this work is far from exhaustive, and will not serve as functional substitute for a CCS.

4. The CCS would also facilitate the development of consistent classification systems for products relating to production and consumption.

TABLE 1: Comparison of the Structure of the Dairy Products and the Fats and Oils Components of the PCE, the CE, and the CPI

PCE Line Number: PCE Category	CE Titles	CPI Index Code	CPI Titles	CPI Item Strata Code	CPI ELI Code	UCC code: UCC Title	CE Diary Item Title	Extended COICOP Code (9 digits)
110: Eggs	Meats, poultry, fish and eggs Eggs	SEFH SEFH01	Meats, Poultry Fish, and Eggs Eggs	FH01	FH011	080110: Eggs	110120: Eggs	011471010
111: Fresh milk & cream	Dairy products Fresh milk and cream	SEFJ	Dairy and related products	FJ				
	Fresh milk, all types	SEFJ01	Milk	FJ01	FJ011	090110: Fresh milk, all types	110060: Milk, fresh, whole, all grades	011411010
							110070: Milk, fresh, exc. whole, spec.	011421010
	Cream		(Cream incl. in Other Dairy Prod. for CPI)				110030: Fresh cream, exc. Non-dairy substitutes	011461030
112: Processed dairy products	Other dairy products							
	Butter		(Butter incl. in Fats and Oils for CPI)				110010: Butter	011511010
	Cheese	SEFJ02	Cheese and related products	FJ02	FJ021	100210: Cheese	110020: Cheese	011451010
	Ice cream and related products	SEFJ03	Ice cream and related products	FJ03	FJ031	100410: Ice cream and related products	110040: Ice cream and related products	011851020
							110052: Frozen yogurt	011851012
		SEFJ04	Other dairy and related products	FJ04	FJ041	090210: Cream	110030: Fresh cream, exc. Non-dairy substitutes	011461030
	Misc. dairy products				FJ041	100510: Other dairy products	1100051: Fresh yogurt	011441011
							110053: Canned yogurt	011441023
							110054: Other yogurt	011441034
							110100: Powdered milk	011431010
							110110: Other dairy products	011461020
	(Non-dairy cream and imitation milk inc. in Fats and Oils for CE)				FJ041	160310: Non-dairy cream substitutes	163010: Non-dairy substitutes	011461010
118: Fats & oils	Fats and oils	SEFS SEFS01	Fats and oils Butter and margarine	FS FS01				
	(Butter inc. in Other Dairy Prod. For CE)				FS011	100110: Butter	110010: Butter	011511010
	Margerine				FS011	160110: Margerine	163020: Margerine	011521010
	Non-dairy cream and imitation milk						163010: Non_dairy substitutes	011461010
	Salad dressing	SEFS02	Salad dressings	FS02	FS021	160212: Salad dressing	163040: Cooked dressing	011911020
							163050: Dressing for salad, ready made	011911030
							163090: Salad dressing N/ Spec.	011911040
		SEFS03	Other fats and oils including peanut butter	FS03				
	Peanut butter				FS031	160320: Peanut butter	163110: Peanut butter	011521020
	Fats and oils				FS032	160211: Fats and oils	163030: Lard and Veg. Shortening	011591990
							163100: Salad and cooking oil	011541010
							163990: Fats and oils, N/ Spec.	011591980

TABLE 2: Mapping of Detailed CE Diary Item Codes to Five Digit COICOP Categories			
COICOP CODE	**COICOP TITLE/ DIARY CODE AND TITLE**	**COICOP CODE**	**COICOP TITLE/ UCC CODE AND TITLE**
01.1.4.	**Milk, cheese and eggs (ND)**	**01.1.4.**	**Milk, cheese and eggs (ND)**
01.1.4.1.	**Whole milk**	**01.1.4.1.**	**Whole milk**
011411010	110060*MILK FRESH WHOLE ALL GRADES	011411010	090110 FRESH MILK ALL TYPES
011411020	110090*MILK FRESH N/SPEC. (part)	011411020	090110 FRESH MILK ALL TYPES
01.1.4.2.	**Low fat milk**	**01.1.4.2.**	**Low fat milk**
011421010	110070*MILK FRESH EXC. WHOLE SPEC.	011421010	090110 FRESH MILK ALL TYPES
011411020	110090*MILK FRESH N/SPEC.(part)	011411020	090110 FRESH MILK ALL TYPES
01.1.4.3.	**Preserved milk**	**01.1.4.3.**	**Preserved milk**
011431010	110100*POWDERED MILK	011431010	100510 OTHER DAIRY PRODUCTS
01.1.4.4.	**Yoghurt**	**01.1.4.4.**	**Yoghurt**
011441011	110051*FRESH YOGURT	011441011	100510 OTHER DAIRY PRODUCTS
011441023	110053*CANNED YOGURT	011441023	100510 OTHER DAIRY PRODUCTS
011441034	110054*OTHER YOGURT	011441034	100510 OTHER DAIRY PRODUCTS
01.1.4.5.	**Cheese and curd**	**01.1.4.5.**	**Cheese and curd**
011451010	110020*CHEESE	011451010	100210 CHEESE
01.1.4.6.	**Other milk products**	**01.1.4.6.**	**Other milk products**
011461010	163010*NONDAIRY SUBSTITUTES	011461010	160310 NON-DIARY CREAM SUBSTITUTES
011461020	110110*OTHER DAIRY PRODUCTS	011461020	100510 OTHER DAIRY PRODUCTS
011461030	110030*FRESH CREAM EXC. NONDAIRY SUBSTIT	011461030	090210 CREAM
01.1.4.7.	**Eggs**	**01.1.4.7.**	**Eggs**
011471010	110120*EGGS	011471010	080110 EGGS
01.1.5.	**Oils and fats (ND)**	**01.1.5.**	**Oils and fats (ND)**
01.1.5.1.	**Butter**	**01.1.5.1.**	**Butter**
011511010	110010*BUTTER	011511010	100110 BUTTER
01.1.5.2.	**Margarine and other vegetable fats**	**01.1.5.2.**	**Margarine and other vegetable fats**
011521010	163020*MARGARINE	011521010	160110 MARGARINE
011521020	163110*PEANUT BUTTER	011521020	160320 PEANUT BUTTER
01.1.5.3.	**Olive oil**	**01.1.5.3.**	**Olive oil**
01.1.5.4.	**Edible oils**	**01.1.5.4.**	**Edible oils**
011541010	163100*SALAD & COOKING OIL	011541010	160211 FATS & OILS
01.1.5.5.	**Other edible animal fats**	**01.1.5.5.**	**Other edible animal fats**
01.1.5.9.	**Oils and fats not specified**		
011591980	163990*FATS & OILS N/SPEC.	011591980	160211 FATS & OILS
011591990	163030*LARD & VEG. SHORTENING	011591990	160211 FATS & OILS
01.1.8.	**Sugar, jam, honey, syrups, chocolate and confectionery (ND)**	**01.1.8.**	**Sugar, jam, honey, syrups, chocolate and confectionery (ND)**
01.1.8.5.	**Edible ices and ice cream**	**01.1.8.5.**	**Edible ices and ice cream**
011851012	110052*FROZEN YOGURT	011851012	100410 ICE CREAM AND RELATED PRODUCTS
011851020	110040*ICE CREAM & RELATED PRODUCTS	011851020	100410 ICE CREAM AND RELATED PRODUCTS
01.1.9.	**Food products n.e.c. (ND)**	**01.1.9.**	**Food products n.e.c. (ND)**
01.1.9.1.	**Sauces, condiments**	**01.1.9.1.**	**Sauces, condiments**
011911010	164040*SAUCES & GRAVIES	011911010	180510 SAUCES AND GRAVIES
011911020	163040*COOKED DRESSING	011911020	160212 SALAD DRESSINGS
011911030	163050*DRESSINGS FOR SALAD READY MADE	011911030	160212 SALAD DRESSINGS
011911040	163090*SALAD DRESSING N/SPEC.	011911040	160212 SALAD DRESSINGS
011911050	164020*OLIVES/PICKLES/RELISHES	011911050	180420 OLIVES, PICKLES, RELISHES
011911060	164050*CONDIMENTS MISC.	011911060	180520 OTHER CONDIMENTS