Volume Title: Scanner Data and Price Indexes

Volume Author/Editor: Robert C. Feenstra and Matthew D. Shapiro, editors

Volume Publisher: University of Chicago Press

Volume ISBN: 0-226-23965-9

Volume URL: http://www.nber.org/books/feen03-1

Conference Date: September 15-16, 2000

Publication Date: January 2003

Title: Using Scanner Data in Consumer Price Indexes. Some Neglected Conceptual Considerations

Author: Jack E. Triplett

URL: http://www.nber.org/chapters/c9734

6

Using Scanner Data in Consumer Price Indexes
Some Neglected Conceptual Considerations

Jack E. Triplett

6.1 Introduction

New data present not only opportunities, but also new problems. They often require new statistical techniques to explore them and new analytical tools to understand what they are telling us. This chapter explores some problems that arise in using scanner data in the consumer price index (CPI).

6.2 Cost-of-Living Index Theory and Scanner Data

The theory of the cost-of-living index (COLI) provides a way to reason about practical decision making in the CPI (Triplett 2001). However, for scanner data we need first to consider some aspects of COLI theory that are normally left unstated or are presently undeveloped.

One area in which COLI theory is undeveloped is the distinction between consumption periodicity and acquisition periodicity. Standard COLI theory rests on a theory of consumption behavior and not on a theory of acquisition behavior. Accordingly, the standard COLI relates solely to consumption periodicity. One usually assumes in the standard COLI theory that consumption periodicity and acquisition periodicity are the same, that is, that price changes have exactly the same effect on consumption and on purchases. In the long run, this may be acceptable, but empirically consumption periodicity and acquisition periodicity are not equal.

Durable goods provide a well-known example. Acquisition periodicity is very long for durable goods, although consumption periodicity for the services of durable goods may be quite short. High-frequency data (time use or

Jack Triplett is a visiting fellow in the economic studies program at the Brookings Institution.

actual consumption studies) can capture consumption of services from durables by individual households, but it is well known that high-frequency expenditure data are unlikely to capture acquisitions.

The distinction between consumption periodicity and acquisition periodicity arises in storable nondurable goods as well. Walter Oi, in discussion at the conference, gave Tabasco sauce as an example of a commodity that is purchased infrequently. For nondurable goods, the wedge between acquisition periodicity and consumption periodicity involves storage costs, as well as search and information costs. Household search and shopping behaviors matter, and not just *consumption* behavior.

Because the standard COLI theory implicitly assumes that consumption periodicity and acquisition periodicity are the same, it ignores search, information, and shopping costs. However, those costs are part of the total cost of consumption. The shorter the period during which prices and quantities are observed—high-frequency data—the smaller the linkage between purchases and consumption.

The economic behavior recorded in high-frequency price and quantity data will be dominated by acquisition periodicity, *which a theory of consumption behavior in response to changes in relative prices does not address.* Toilet paper and soft drinks are frequently on sale. Consumers know they are frequently on sale and can plan their acquisitions accordingly. Surely sale prices for storable commodities affect the timing—and location—of acquisitions far more than they influence the quantity of consumption. To confront the household behavior recorded in high-frequency data requires a theory that adequately describes search, storage, shopping, and other household activities that drive a wedge between acquisition periodicity and consumption periodicity.

Applying the standard COLI paradigm to CPI component indexes for storable nondurables like toilet paper and soft drinks, or for many of the other commodities available in scanner data—and indeed for the CPI price index for bananas—may be instructive and enlightening. I have in mind contributions such as that by Diewert (1995). However, because the standard theory is a theory of consumption, not of acquisition, it is incomplete.[1] Applying the theory to data on acquisitions may yield misleading conclusions. In this, I agree with Pollak (1998) that we need a more elaborate COLI theory for guidance, perhaps building on the neglected work of Baye (1985).

Readers who are familiar with the international literature on CPIs may have noted already that many statistical agencies have recognized pragmatically the problems posed by consumption periodicity and acquisition periodicity and have sought to define them away by saying that their CPIs are

1. Moreover, adding to the standard consumption model tastes over different retail outlets, in parallel with tastes for different commodities, does not make the problem more tractable. See section 6.4.

indexes not of consumption prices but of acquisition prices. In the same breath, they often also state that their CPIs are not COLIs. They do not, however, escape the problems described in the next two sections.

6.3 Cost-of-Living Index Theory and Aggregation Questions

Cost-of-living index theory is, partly, a theory of commodity aggregation. The theory tells us how to aggregate commodities (i.e., don't use fixed consumption weights to form intertemporal comparisons).

Other aggregations exist in price and quantity data that are assumed away in the theory's simplifying assumptions. One aggregates, inevitably, over time. High-frequency price collections may show price variations that are invisible in lower-frequency collections. Some studies have endorsed unit values to reduce high-frequency price variation, but this implicitly assumes that the high-frequency variation represents simply noise in the data and is not meaningful in the context of a COLI. That is debatable. We need to develop a theory that confronts the data, not truncate the data to fit the theory.

Conventional price and quantity data are also aggregations over people. Cost-of-living index theory rests on a theory of an individual consumer's behavior.

Scanner data are, in general, less aggregated. However, scanner data are typically disaggregated by store, not by individual consumers, so they are still aggregations over individuals. Any aggregation over consumers poses knotty questions (Pollak 1989), which scanner data do not circumvent. Indeed, scanner data on prices charged and quantities sold *by stores* may exacerbate the consumer aggregation problem, as suggested below.

Statistical agencies now collect the prices *offered by retailers,* and they aggregate them, or their price relatives, across stores. Cost-of-living index theory requires the *prices paid by consumers* and assumes that one price prevails (perhaps because search theory was not a prominent part of economics before Stigler [1961]). Accordingly, there is no store aggregation problem. The average price across retail outlets—strictly speaking, the average price change—is the price that is deemed relevant to the theory.

It is well established, however, that prices vary across stores at any instant of time, even for precisely defined commodities, and the variation does not seem fully accounted for by differences in retailing services. Variability in store prices presents both opportunities and problems. In some of the existing empirical work using scanner data, researchers have reaggregated the store data into unit values, partly to reduce the size of the database (for example, Reinsdorf's [1999] study of coffee prices). However, if households shift their purchases across stores in response to sale prices, special promotions, and so forth, one does not want to aggregate across stores; household search and shopping behavior is a serious topic for investigation. In any

Table 6.1 **Numerical Example of Semihigh Frequency Data**

	Observation Period (Month)			
	1	2	3	4
	A. The Data			
Actual price observations (¢)				
Store 1	49	99	33	49
Store 2	99	49	99	99
Hypothetical quantities				
Store 1	300	100	350	300
Store 2	50	250	50	50
Hypothetical revenue ($; price × quantity)				
Store 1	147.00	99.00	115.50	147.00
Store 2	49.50	122.50	49.50	49.50
Hypothetical revenue shares (w_t)				
Store 1	0.75	0.45	0.70	0.75
Store 2	0.25	0.55	0.30	0.25
	B. Weighted CPI Basic Component Calculations			
Chained Laspeyres	1.000	1.636	2.071	2.774
Chained Paasche	1.000	0.747	0.332	0.440
Chained Fisher	1.000	1.105	0.830	1.104
Chained geometric mean	1.000	1.418	1.280	1.688
Chained unit value	1.000	1.127	0.735	1.000

Source: Actual price observations from extract from unpublished Canadian CPI data from Schultz (1994)

case, unit values across stores are not the prices actually faced by households and do not represent the per-period price in the COLI, even if the unit values are grouped by type of retail outlet.[2]

These several aggregations—across time, across people, and across retail outlets—cause analytical difficulties that the standard theory does not address. High-frequency collection of price and quantity data from retailers, feasible with scanner information, may result in statistics that describe the behavior of no consumer. The following section demonstrates.

6.4 Some Interpretive Examples with High-Frequency Price and Quantity Data

The price information in table 6.1 is an extract of data from Schultz (1994), which were used in Triplett (1998). They are actual monthly prices in one city for a particular size and brand of soft drink, collected for the Canadian CPI. Since they are monthly, they are only medium-frequency, not high-frequency, data. Scanner data in principle yield even higher fre-

2. For a contrary position, see Balk (1999) and Diewert (1995).

quency price and quantity data, which might magnify substantially the effects in this example.

The quantity data in table 6.1 are hypothetical, put together only on the hypothesis that periodic sales of soft drinks "work," in the sense that they result in larger quantities of soft drinks sold in any store that is offering a temporary sale price than otherwise. Although the quantity data are hypothetical, I interpret the paper by Feenstra and Shapiro (chap. 5 in this volume) as showing that my quantity data are indeed realistic as a description of what actual scanner data will show. The hypothetical quantity data in table 6.1 are designed to encompass consumer preferences for store 1, which always sells more for any particular price than store 2, along the lines suggested by Diewert (1995).

Note that in period 4 the prices return to exactly their values in period 1. For heuristic reasons, the hypothetical quantity data also return to their exact period 1 values when the prices return to their initial values.

Section B of table 6.1 shows that chained versions of standard price index formulas behave perversely, in the sense that none recovers the initial period's level when the prices and quantities return (in period 4) to their initial period levels. As this example suggests, chaining is part of the "problem" with high-frequency data, as is the common presumption that the price indexes should necessarily be time reversible. These topics are not explored here.

Applying the conventional theory of consumption to the quantity changes for store 1 shown in table 6.1 implies that this store's customers gorge themselves on soft drinks during the sale month and go thirsty in nonsale months, and similarly for store 2's customers. However, the quantity changes in these data are unlikely to represent changes in *consumption* of soft drinks in response to sale prices. Instead, at least two things are driving the data: Households switch stores in response to sales, and they stock up on soft drinks when they are on sale, consuming the sale-price soft drinks in other periods when they do not buy them.

Thus, I speculate that some households exhibit shopping and inventory behavior, although others may not. In these circumstances, what are households' acquisition prices for soft drinks? Their consumption prices? Consumption prices are relevant for the COLI, the former for CPIs of those countries that refer to their indexes as non–cost-of-living acquisition price indexes (such as Australia and the European Union's Harmonized Indexes, or HICP). Both acquisition prices and consumption prices depend on household shopping and inventory behaviors. Section B of table 6.2 lists some possibilities.

One type of household doesn't shop and doesn't inventory: call it the habit purchaser. The prices this household faces (and the period-to-period price relatives) are given by lines (1a) and (1b).

For the habit purchaser household, and only for the habit purchaser, the price changes Statistics Canada (or the Bureau of Labor Statistics [BLS])

Table 6.2 Price Changes Faced by Different Consumer Types

	Observation Period (Month)			
	1	2	3	4
	A. The Data			
Actual price observations, soft drink, Canadian city				
Store 1	49¢	99¢	33¢	49¢
Store 2	99¢	49¢	99¢	99¢
	B. Prices and Price Index Relatives for Different Consumer Types			
1a. The habit purchaser (store 1)	49¢	2.02 (99/49)	0.33 (33/99)	1.48 (49/33)
1b. The habit purchaser (store 2)	99¢	0.49 (49/99)	2.02 (99/49)	1.00 (99/99)
2a. The shopper	49¢	1.00 (49/49)	0.67 (33/49)	1.48 (49/33)
2b. The shopper, with assumed 15¢ search costs	49¢	1.31 (64/49)	0.98 (48/49)	1.94 (64/33)
3. The inventory/shopper (store 1)	49¢	(no purchase)	0.67 (33/49)	(no purchase)

Source: Actual price observations from extract from unpublished data from Schultz (1994).

collects from the stores match exactly both acquisition prices and consumption prices. The habit-purchaser households may vary their consumption of soft drinks in response to price changes (they drink more fruit juice in months when soft drinks are not on sale), matching the behavior that is embodied in COLI theory, but they do nothing more. If store quantities actually measure purchases (and therefore consumption) by habit-purchaser households, then conventional price index calculations on store price and quantity data measure lower-level COLIs, along the lines developed in Diewert (1995). I have no data, but I presume that these households do not account for much of the variation in store quantities that typical scanner data show result from soft drink sales.

Next is the "shopper" household. This household switches stores, only buying at the sale price, and it consumes all that it purchases in each period. If the household ignores switching and shopping costs, then the acquisition prices it faces are given by line (2a). As with the habit purchaser, the acquisition price for the shopper is also the household's consumption price. This household never pays the nonsale price, so only the sales prices are relevant. This household's price index is an index of the minimum prices prevailing in each period. In the second period, for example, when prices in store 1 and store 2 just reverse themselves, this shopper household faces no change in price.[3]

3. Note that the unit value index in table 6.1 does not represent the acquisition price index for this shopper, essentially because this shopper has no preferences between the two stores, and I have built store preferences into the quantity data used in table 6.1. This is an artifact of the hypothetical data, but not an unreasonable one. Unit values across stores do not in general correspond to the prices that are relevant for COLI theory, nor do they represent acquisitions prices.

However, the price changes collected by Statistics Canada or the BLS do not measure changes in the prices that the shopper household faces, nor do store-level scanner data. As table 6.1 shows, weighted Laspeyres, Fisher, geometric mean, and unit value indexes of store prices all rise between periods 1 and 2, but the shopper household's price index is unchanged (as is, coincidentally, the ratio of unweighted average prices [RA]). For the third period, the shopper household's price index falls by one-third, which is more than any of the weighted indexes in table 6.1 (ignoring the Paasche formula), because for the shopper household the price rise in store 2 is irrelevant.

A variant is the shopper household that considers switching and shopping costs before changing stores. Acquisition costs for this household are given by line (2b). In this case, none of the store prices or their changes measure acquisition costs directly, nor do they measure consumption prices. For example, in the second period, this household experiences a price increase because obtaining the 49¢ soft drink from store 2 entailed a 15¢ switching cost. Collecting scanner data by retail outlet does not provide the relevant measure of price change faced by this household, nor does the collection methodology of the BLS or Statistics Canada. Shopping and switching costs are outside the domain of the CPI.

The final case is the inventory shopper. This shopper knows that soft drinks are frequently on sale and follows the rule: Stock up when they are on sale and consume from inventory when they are not. Although I have no data, I presume that this household type accounts for a large amount of the quantity variation when soft drinks go on sale. Line (3) shows acquisition prices faced by the inventory shopper. A similar inventory shopper exists for store 2, but the data are omitted from the table.

The inventory shopper makes no purchases during months 2 and 4, in which soft drinks are not on sale in this household's favorite store. The household's acquisition price is not defined in those months.

What about this household's consumption price? One could elaborate on inventory, storage, and capital costs and calculate a user cost equation for consumption of soft drinks from inventory. Or one could assert that the household should charge itself the opportunity cost (the nonsale price?) for consumption out of inventory. However, the point is that it is not obvious how the inventory purchaser should be treated in scanner data for a COLI or in scanner data for a non–cost-of-living "acquisitions" price index.

The COLI, or the CPI, should be viewed as the average of the indexes across households. If these five household types were equally distributed across the population, one could average the price relatives from table 6.2, making some rule to allow for the inventory shoppers' consumption. I have not presented that "democratic" CPI in table 6.2. Such a democratic CPI, calculated on actual data, is unlikely to resemble any of the commonly used index number formulas in table 6.1.

What is to be concluded from these examples? First, explorations of scanner data, and indeed of methods for calculating component indexes of the CPI, have mostly employed standard index number formulas from the existing price index literature, applied to store data. Theoretical analyses of price index basic components also follow the standard index number commodity substitution paradigm. This is understandable, since it is a relatively new topic. This empirical and theoretical work assumes, implicitly, that scanner data are measuring the commodity substitution behavior that is incorporated in the usual COLI theory.

The examples in table 6.2 suggest that conventional index number approaches only capture acquisition and consumption prices for the household that doesn't shop and doesn't inventory (the habit purchaser). No index number formula, *applied to period-to-period store prices,* can solve the problem that such prices are not the period-to-period transactions prices for the shopper households or for households that inventory storable commodities. Moreover, it is hard to see how an index number formula, no matter how ingenious, can deal with the zeros in the inventory purchasers' transactions for periods when the soft drink is not on sale. An index number formula cannot solve the problem that we are collecting, in the CPI and in scanner data, prices from sellers. To understand household behavior with respect to periodic sale prices, we need prices paid by buyers.

Second, price indexes calculated using scanner data seem always to differ from the CPI. It is not clear why. However, it is also not clear that we have been addressing the problems posed by high-frequency data with the right theoretical tools, and the right tools are always necessary to an understanding of any economic phenomenon. Much more work needs to be done on the theoretical and practical frameworks for using scanner data in the CPI.

6.5 Classifications

In their chapter in this volume, Hawkes and Piotrowski point out that ACNielsen and BLS classifications diverge. Their table 1.4 points out that to ACNielsen ice cream is a frozen food, whereas to BLS it is a dairy product. From a conceptual point of view, what commodity classification is appropriate for the CPI?

Classifications are not just definitions. Classifications group data. It is not generally recognized that economic theory exists that can guide thinking about how economic classifications should be devised. A review of the theory of economic classifications—including classifications for the CPI—is Triplett (1990).

The conceptual approach to economic classifications, as developed in my 1990 paper, has transformed thinking about industry classification systems (see "Preface" to U.S. Office of Management and Budget 1998; Kort 2001).

However, I have to say that the use of economic theory has had no impact whatever on classification systems used for consumer price indexes, despite the fact that the CPI is the economic statistic that is probably most closely aligned with a concept from economic theory, by which I mean the theory of the COLI.

One reason, no doubt, is the inevitable lag between ideas and implementations (although the implementation lag was short for industrial classifications). Additionally, empirical implementation of an economic concept for CPI classification is harder than for industry classification systems, partly because there is more than one way to proceed and partly because the empirical knowledge is scant for implementation of the theoretical principle.

For CPI classifications, economic theory tells us that we should look for separable "branches" of the utility or the consumer cost function. The basic references are Pollak's (1989) paper on the subindex of the COLI and Blackorby and Russell (1978). This classification concept is very hard to implement empirically, partly because separability is a mathematical condition that is not very intuitive. Triplett (1990) also discusses chains of close substitutes.[4]

In any case, the appropriate COLI classification concept is drawn from consumer demand theory. In the language used in Triplett (1990), CPI classifications are "product" grouping systems. Theoretical classifications for a COLI do not depend on supply conditions; production conditions determine industry classification systems, not CPI classifications.

Nielsen and BLS seem both to have used some sort of supply-side reasoning about classifications. Nielsen classifies ice cream according to the way it is handled, transported, and stored; for the CPI, what is important is how the product is used, not how it is produced. The BLS classifies ice cream according to one of its major ingredients, dairy products (even, I gather, when the ice cream contains no dairy product!). Classification by materials inappropriately applies a supply-side criterion; a demand-side criterion is appropriate for CPI classifications. Thus, the Neilsen and the BLS classifications are both inappropriate for a COLI.

I do not have any empirical data on which to rely, but my intuition suggests that ice cream belongs in a "dessert" branch. Of course, practicality considerations ought to enter as well. For efficient collection of prices, one ought to be able to find products that are grouped together in CPI food index components in the same part of the grocery store, which might force modification of the theoretical principle (perhaps in Neilsen's direction).

As a final point on this important question of classifications for CPIs, users of the CPI should know of a proposal that is currently being consid-

4. Hicksian aggregation also figures in the theoretical literature, and (implictly) in deflation practices for national accounts.

ered by the international group that is writing an international manual for consumer price indexes, because it seems a wholly inappropriate way to proceed. They propose to use the classification of consumer expenditures that was published in the 1993 system of national accounts (SNA93; Inter-Secretariat Working Group on National Accounts 1993). This classification system is called COICOP (Classification of Individual Consumption by Purpose). COICOP was plucked out of the air by a small group of people who did not even consult with consumer price index experts to see if their classification system reflected CPI practice or accumulated wisdom. Constructing a classification with neither theory nor practice is the very worst methodology.

The right way to proceed is, surely, to use economic theory and empirical analysis to determine the appropriate classifications in the CPI. Because these will also be classifications for grouping consumption expenditures, the same classifications should be used in the consumption portions of national accounts.

As developed in SNA93, COICOP was not very detailed (it did not distinguish anything more detailed than "food," for example), so little harm was done in practice. More recently, however, COICOP was elaborated with more detail, but without any use of economic principles for classifications, so far as I know. The classification system for international CPIs should not be developed in the same way the COICOP system was developed for national accounts. Economic principles should be employed for CPI classifications.

It is also worth noting that classifications of consumption in the U.S. CPI and in the U.S. national accounts do not agree. This is a source of serious problems for economists who wish to analyze consumer behavior, and it should receive attention from both agencies. However, a new and improved classification should also incorporate economic theory, to the extent possible with available knowledge. The United States should not just adopt COICOP, without evidence that it meets the conceptual principles of economic classifications for consumption and the CPI.

Classifications should get more attention from economists. The classifications that are chosen by some public or private statistical agency provide the indivisible units of economic analysis.

6.6 Conclusions

Using scanner data in the CPI is a more complicated matter than it may appear. For one thing, our theoretical tools (mainly the existing corpus of COLI theory) are not fully adequate for the economic behaviors—search, shopping, and inventory behaviors—that are incorporated into high-frequency data. For another, aggregations over time, over households, and over stores—present in existing CPI data—are not lessened with scanner

data, and their effects may be more severe with high-frequency data than with the lower-frequency data with which we have long worked. Additionally, the quantity changes that are apparent in high-frequency store data are likely to reflect inventory and shopping behavior in response to sale prices more than changes in consumption. As a result, acquisitions and consumption periodicities differ, and the period-to-period store prices (the output of scanner data) diverge from households' acquisitions and consumption prices in ways that depend on their inventory and shopping behaviors, as shown by the lines in table 6.2.

Classifications also matter. We need better classifications in the CPI in order to have CPI component indexes that are suitable for economic analysis. Collecting and processing scanner data may reduce collection costs, but scanner data will greatly increase index editing and analysis costs. As these latter problems loom larger, there is more need to think hard about the grouping of the data, because inappropriate groupings may increase the editing and processing costs unduly, as well as create groupings that are inappropriate for economic analysis.

References

Balk, Bert. 1999. On the use of unit value indices as consumer price subindices. In *Proceedings of the fourth meeting of the International Working Group on Price Indices,* ed. Walter Lane, 112–20. Washington, D.C.: U.S. Department of Labor.

Baye, Michael R. 1985. Price dispersion and functional price indices. *Econometrica* 53 (1): 213–23.

Blackorby, Charles, and Robert R. Russell. 1978. Indices and subindices of the cost-of-living and the standard of living. *International Economic Review* 19 (1) : 229–40.

Diewert, W. Erwin. 1995. Axiomatic and economic approaches to elementary price indexes. NBER Working Paper no. 5104. Cambridge, Mass.: National Bureau of Economic Research, May.

Inter-Secretariat Working Group on National Accounts. 1993. *System of national accounts 1993.* Brussels/Luxembourg, New York, Paris, Washington, D.C.: Commission of the European Communities, International Monetary Fund, Organisation for Economic Cooperation and Development, United Nations, and World Bank. Available from United Nations Publications.

Kort, John R. 2001. The North American industry classification system in BEA's economic accounts. *Survey of Current Business* 81 (5): 7–13.

Pollak, Robert A. 1989. *The theory of the Cost of Living Index.* New York: Oxford University Press.

———. 1998. The Consumer Price Index: A research agenda and three proposals. *Journal of Economic Perspectives* 12 (1): 69–78.

Reinsdorf, Marshall. 1999. Using scanner data to construct CPI basic component indexes. *Journal of Business and Economic Statistics* 17 (2): 152–60.

Schultz, Bohdan J. 1994. Choice of price index formula at the microaggregation level: The Canadian empirical evidence. In *International Conference on Price In-*

dices, Papers and Final Report: First Meeting of the International Working Group on Price Indices. Ottawa, Canada: Statistics Canada, November, 93–127. Available at http://www4.statcan/ca/secure/english/ottawagroup/toc1.htm

Stigler, George J. 1961. The economics of information. *The Journal of Political Economy* 69 (3): 213–25.

Triplett, Jack E. 1990. The theory of industrial and occupational classifications and related phenomena. *Bureau of the Census 1990 Annual Research Conference Proceedings.* Washington, D.C.: U.S. Department of Commerce, August, 9–25.

———. 1998. Elementary indexes for a consumer price index. In *Proceedings of the fourth meeting of the International Working Group on Price Indices,* ed. Walter Lane, 176–97. Washington, D.C.: Department of Labor.

———. 2001. Should the Cost-of-Living Index provide the conceptual framework for a consumer price index? *The Economic Journal* 111 (June): 312–35.

U.S. Office of Management and Budget. 1998. *North American industry classification system: United States, 1997.* Washington, D.C.: Executive Office of the President, Office of Management and Budget.