Survey Methodology, December 2006
Vol. 32, No. 2, pp. 197-216
Statistics Canada, Catalogue No. 12-001

On Sample Survey Designs for Consumer Price Indexes

Alan H. Dorfman, Janice Lent, Sylvia G. Leaver and Edward Wegman [1]

Abstract

Survey sampling to estimate a Consumer Price Index (CPI) is quite complicated, generally requiring a combination of data from at least two surveys: one giving prices, one giving expenditure weights. Fundamentally different approaches to the sampling process–probability sampling and purposive sampling–have each been strongly advocated and are used by different countries in the collection of price data. By constructing a small "world" of purchases and prices from scanner data on cereal and then simulating various sampling and estimation techniques, we compare the results of two design and estimation approaches: the probability approach of the United States and the purposive approach of the United Kingdom. For the same amount of information collected, but given the use of different estimators, the United Kingdom's methods appear to offer better overall accuracy in targeting a population superlative consumer price index.

Key Words: Elementary index; Probability proportional to size sampling; Purposive sampling; Scanner data; Superlative index.

1. Background

From start to finish, survey sampling for the sake of a Consumer Price Index (CPI) must rank among the most complicated of sampling enterprises. The population target is hard to pin down, the appropriate domain of items debated, the definitions of the raw ingredients – prices, quantities, items – ambiguous and subject to question. The ultimate estimator – the estimator of the all-items CPI – relies on data from at least two surveys, one giving prices, and one giving "weights." Below the level of "composite items" (or "item strata") – groups of items supposed homogeneous in their price movements – there is typically no cost effective way to keep sampling weights up to date. Debate therefore goes on about the proper choice among various simple alternative estimators of price change for item categories, the "elementary aggregate indexes." The appropriate method of aggregating these price changes, using the weights, is subject also to debate.

There are two broad approaches to the sampling by which prices are collected: probability sampling and judgment sampling. The most commonly accepted approach to survey sampling in general requires injecting an element of randomness into the survey process and relying on this randomness to make inference about population characteristics of interest – probability or "design-based" sampling; see, *e.g.*, Särndal, Swensson and Wretman (1992). This approach was not always taken for granted. Early in the 20^{th} century, "purposive" or "representative" sampling was considered a viable, and possibly better, option. More recently, the prediction-based school of Royall has challenged design-based assumptions; see *e.g.*, Valliant, Dorfman and Royall (2000).

In the U.S., all CPI-related surveys are carried out using complex probability sampling techniques. Around the world, most CPI's are constructed from judgment samples, in which experts on the different item strata choose broader or narrower classes of items for which field representatives collect prices. The fundamental reason for this is the difficulty of getting all the data one needs on the plethora of items sold, and the places where they are sold, to make probability sampling feasible.

The interesting fact is that there has been very little assessment of the relative accuracy of the different approaches to sampling. Indeed it has not been clear that it is feasible to make such assessments. The underlying population price index, for even the smallest of countries, involves so many transactions on so many items in so many places as to be inaccessible. Moreover, the population of items on the market is in a constant state of flux, complicating the application of traditional population index formulas. How then can one judge the relative closeness to "truth" of different sample-based estimates? Furthermore, in most cases, not even sample information is available for a key ingredient of the population index-namely the *quantities* of items sold – so even artificially constructing a population for test purposes from sample data has not been feasible.

The relatively recent availability of *scanner* data, in the U.S. and elsewhere, presents an unprecedented opportunity for testing sampling approaches and estimators. These data include prices *and* quantities, typically on a weekly basis, of *all* the items sold in a given category within a large sample of outlets having scanner devices. Such data may be used to construct realistic populations of transactions for which the true price index is *known*. We can then use various methods to sample from this population, construct different index

1. Alan H. Dorfman, Office of Survey Methods Research, and Sylvia G. Leaver, Office of Prices and Living Conditions, U.S. Bureau of Labor Statistics, 2 Massachusetts Ave NE, Washington, D.C., U.S.A., 20212; Janice Lent, U.S. Bureau of Transportation Statistics, 400 7^{th} Street, SW, Washington, D.C., U.S.A., 20590; Edward Wegman, Center for Computational Statistics, George Mason University, Fairfax, VA, U.S.A., 22030.

estimates of interest, and compare the results to the known population parameters. One such study, described by de Haan, Opperdoes and Schut (1999), seems to show that "cutoff sampling," the sampling of the few largest (in terms of revenue generated) items in the population, outperforms two important design-based approaches: simple random sampling (*srs*), and probability proportional to size sampling (*pps*) (where the size measure is, again, revenue).

One difficulty in any study making such comparisons is the task of maintaining a "level playing field." If one sampling method, for example, makes use of (population) information that might not actually be available in practice, while another does not, the comparison of methods is undermined. Similarly, if one method provides only one sample or very few samples, and another provides thousands, special precautions are needed in comparing the two; indeed, such a comparison might require serious qualifications. Given the complexity of the sampling and estimation methods used in price index computation, it is not surprising that these and many other difficulties complicate experiments designed to compare various methods.

Ideally, to compare the approaches, for example, of two countries, we would mimic the whole complex sampling and estimation process of each and evaluate its costs. Both processes would be allowed the same budget, and we would be able, by some predetermined and equitable measure, to evaluate each estimate's proximity to a known target index.

This paper comprises two studies, a large primary study, and a smaller secondary, follow-up study.

The main study is described in Sections 2 through 4. Section 2 describes the construction of the target population. Section 3 describes the "US" and "UK" methodologies and outlines the simulation details. No attempt is made to assess relative costs (thus falling short of the ideal), but the competing approaches are made as equal as possible in terms of the information they use. Results, which favor the UK approach, are given in Section 4.

The follow-up study, in Section 5, attempts to disentangle the effects of different components of the two approaches, in particular sampling method and elementary index formula. Section 6 gives a final summary and discussion.

Note on the target indexes. The price index literature contains myriad formulas for calculating price change between one period and another. Different indexes are compatible with different assumptions regarding the "average" consumer's buying behavior in response to price change. The "fixed market basket" indexes, the commonly employed Laspeyres and less used Paasche formulas, are compatible with the assumption that consumers continue to purchase the same items in the same quantities regardless of changes in relative prices. The Laspeyres index projects the period 1 ("base period") quantities forward to period 2 ("current period"), while the Paasche applies the period 2 quantities to period 1. The geometric index (or "Jevons" or "*geomean*"), usually weighted with base period expenditure shares, assumes that consumers adjust the quantities they purchase in such a way that the expenditure share for each item remains constant across time. The "superlative" Fisher, Törnqvist, and Walsh index formulas, which rely on quantity (or expenditure share) information for both periods, do not require these assumptions. Formulas for these indexes, with a superscript y representing the base period, $y+1$, the current period, and i the item purchased, for the indexes are given in Appendix A.

The debate on the all items target index usually comes down to choosing between the Laspeyres and one of the superlative indexes. Most countries select a Laspeyres target index, but a strong case (Diewert 1997) can be made that the proper target is a superlative index (usually the Fisher or Törnqvist), even if the formula for the estimator does not resemble one of the superlative population index formulas. Because of the form of the US elementary aggregates – geometric mean – and the fact that previous research (Dorfman, Leaver and Lent 1999) indicated that the lowest level of estimation can have a major impact, the weighted *geomean* will be included among the potential targets. Targets for a given domain are calculated based on prices and quantities of all items in the domain following the formulas in Appendix A (a single-stage aggregation of prices and quantities).

Note: These formulas are deceptively simple, requiring the notation of section 3 for their full development. Thus, in a formula such as that for the Fisher index F (which we will take as our target in the main study of sections 2 – 4) "i" represents an item i belonging to a small class c (an "ELI" or "representative item" – see section 3), where c is itself a subset of wider classes; further, the item i is sold in a particular outlet j, classified as part of a particular chain k, and located in a particular sampled geographic area, the primary sampling unit (*psu*) l. Thus, the expression for a sum $\sum_i$, in the case of the overall population index, is really shorthand for $\sum_{l=1}^{3}\sum_{k=1}^{8}\sum_{j\in(k,l)}\sum_{C=1}^{4}\sum_{h\in C}\sum_{c\in h}\sum_{i\in(j,c)}$; a similar remark holds for $\prod_i$. In short, these are sums and products over *all* items in the population. Contractions of this full expansion will give the population indexes for the different classes C, *etc*.

2. The Population for the Primary Study

The data source for the present study is a scanner data set for breakfast cereal for the years 1995 through 2000 in three separate but contiguous sections of a single large

metropolitan area. The data set was purchased from the A.C. Nielsen Corporation by the U.S. Bureau of Labor Statistics for the purpose of determining the feasibility of incorporating scanner data into the U.S. CPI; see Richardson (2000).

From these data, artificial "populations" were drawn by the method described below. Thus the study encompasses an apparently narrow world, that of cereal, within a fairly restricted geographic domain. Even this restricted world, however, allows for rather discrepant price trends over the six years. Thus, although we will not be able to generalize, in any simple fashion, to global price indexes encompassing a wide heterogeneity of products, we may be able to derive important clues on the effects of different sampling methods and the behavior of particular estimators.

The six years' worth of data available provided the opportunity for establishing fairly long-term price trends. In order to keep the data manageable and avoid the complications of seasonality, we limited ourselves to February data. For February of year y, for each item (*i.e.*, each particular combination k of brand, type, size) in a particular outlet, four weeks t of price and quantity data were combined into a single month's price and quantity, by using the sum of quantities $q_k^{\text{Feb},\,y} = \sum_{t\in \text{Feb},\,y} q_k^t$ sold during the month as the quantity, and the *unit value* $p_k^{\text{Feb},\,y} = \sum_{t\in \text{Feb},\,y} q_k^t\, p_k^t / \sum_{t\in \text{Feb},\,y} q_k^t$ as the price. Unit values computed over short periods of time (*e.g.*, a month) give perhaps the most meaningful sense of the "average" price for a particular item. The use of unit values smoothes the data and reduces it to more manageable proportions; for a discussion of circumstances under which use of unit values is appropriate or not appropriate see Balk (1999).

For the purposes of the study, the population of breakfast cereals was divided into four groups:

1. Hot Cereals (H)
2. "Sugary" cereals (S)
3. "Fruity" cereals (F)
4. "Plain" cereals (P), *i.e.*, cold cereals not falling into categories (2) and (3).

For each group, for each successive pair of years, superlative and non-superlative indexes were calculated, using item-outlet combinations available in both years. In practice, there is generally a problem with getting perfect match-ups from period to period, and finding means to deal with this by finding substitutes for original products or by other means is important; this study bypasses this particular problem.

Long range indexes ('95 to '00) were calculated both directly and by chaining annual indexes. Additionally, indexes were calculated on the "core" items, meaning those available in all six years. On a year-to-year basis, the core items represented between 53% and 61% of a given year's items available for year-to-year comparison; core expenditure was from 83% to 91% of the total expenditure on all (core and non-core) items. There were 326 core items, and a total of 848 distinct items over the course of '95 to '00.

Values of year-to-year population indexes are represented in Figures 1 through 5. Figure 1 gives the index $\hat{I}^{y,\,y+1}$ values for Sugary cereals based on all items sold in stores in both y and $y+1$, for (February of) $y =$ 1995, ..., 1999 (the "all items"). Values for five indexes are shown, including the Paasche P and, as being of academic interest, a unit value index U, the ratio of quantity weighted mean prices, averaged over all item types and outlets. Figure 2 gives results of the same calculations, but limited to "core" items. Figures 1 and 2 are almost identical, and the resemblance between indexes calculated using all items and those using just the core items held for the other cereal categories as well. Figures 3 through 5 give the results for the core-based indexes for Hot, Fruity, and Plain cereals. For any given index, the figures indicate serious differences across cereal categories. The price trends of the four major groups are quite different: H increases, S decreases sharply, F decreases modestly, and P increases modestly.

Table 1 gives long range ('95 – '00) direct indexes and chained indexes for "all items" and "core items." ("All items" for constructing an index between two given years, are all those item/outlets with positive quantities sold, both years). Again, there is very little difference between the values for "core items" and "all items," and sharp differences from one cereal category to another. The chained and direct results are close for the superlative indexes but tend to be discrepant for the *geomean*, Laspeyres, and Paasche. The chained and direct unit value indexes are close and in fact the latter would be identical to the chained based on the core items, except that, for convenience, the year to year indexes were based only on item-outlet combinations available for both years.

Except for some oscillation of position of the unit value index, we observe a clear ordering of index values by formula, the same across categories, which may be summarized as follows: (1) The superlative indexes differ relatively little from each other, a noteworthy result given the amount of variability in the item-outlet price relatives and quantities, due to "sales." (2) The traditional non-superlative indexes differ wildly from each other and the superlatives, with the *geomean*, weighted by first period expenditures, running much *higher* than the superlatives, the Laspeyres still higher, and the Paasche (not surprisingly) much lower. These results suggest that, in Cereal World, not only quantity, but expenditure share, tends to drop in period 2 on

an item having a sharp increase in price in that period. (3) The unit value index is low as well, but, except in the case of Hot cereals, is better than the tradional non-superlative indexes in approximating the superlatives. (4) In the light of later developments in this paper, and at the suggestion of a referee, two non-quantity based indexes are included in this table (although not in the figures): the *dutot* index, which is a simple ratio of average prices (RA) – see Appendix A, and an *unweighted* (that is, constant weighted) *geomean*; both are usually reserved for computing indexes at the elementary level. The results are surprising: in approximating the superlatives, they do as well as or better than the traditional, quantity based non-superlatives, about on a par with the unit value index.

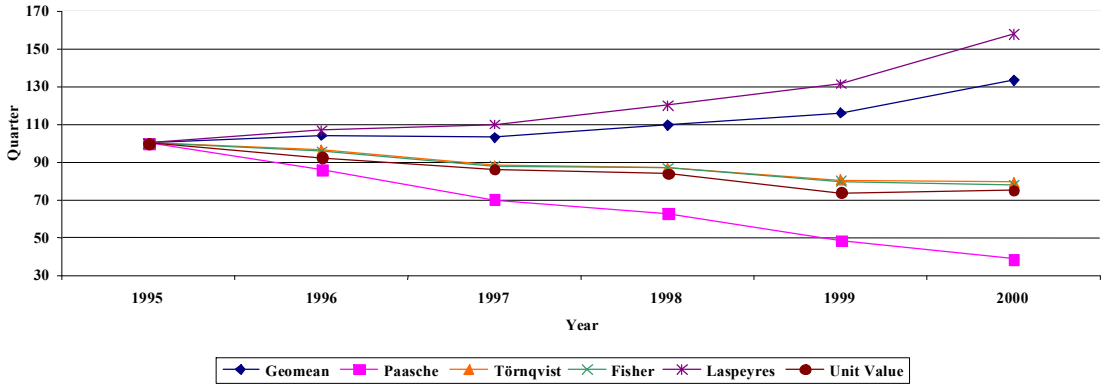

Figure 1. Annually Chained Population Target Indexes for All Sugary Cereals February-to-February Indexes, 1995 = 100.

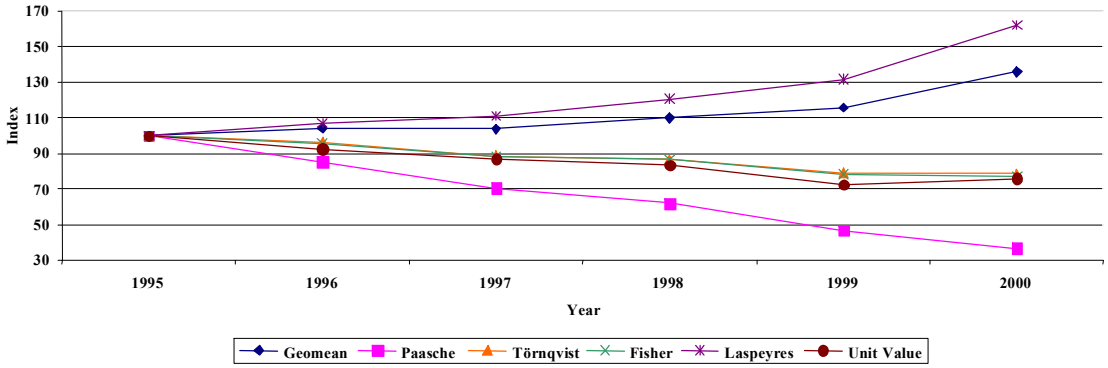

Figure 2. Annually Chained Population Target Indexes for Core Sugary Cereals February-to-February Indexes, 1995 = 100.

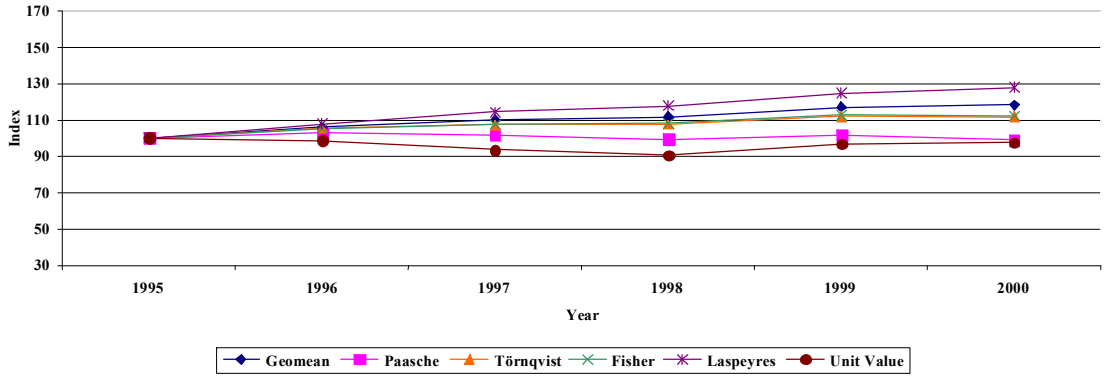

Figure 3. Annually Chained Population Target Indexes for Core Hot Cereals February-to-February Indexes, 1995 = 100.

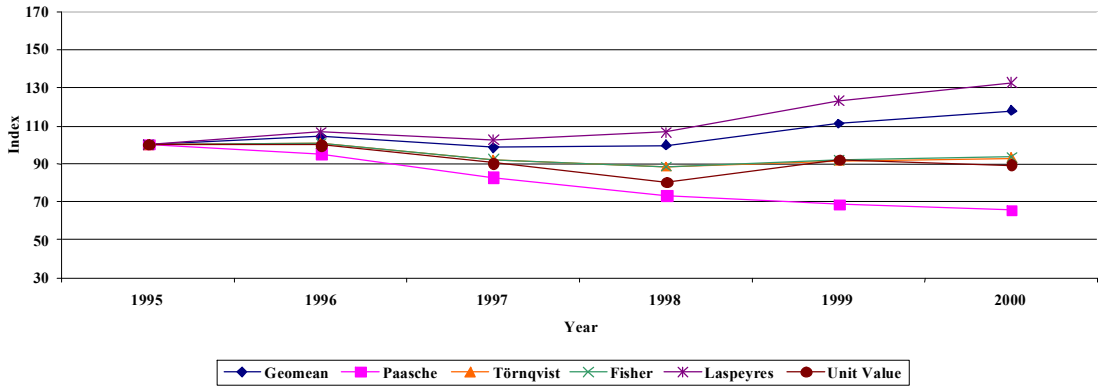

Figure 4. Annually Chained Population Target Indexes for Core Fruity Cereals February-to-February Indexes, 1995 = 100.

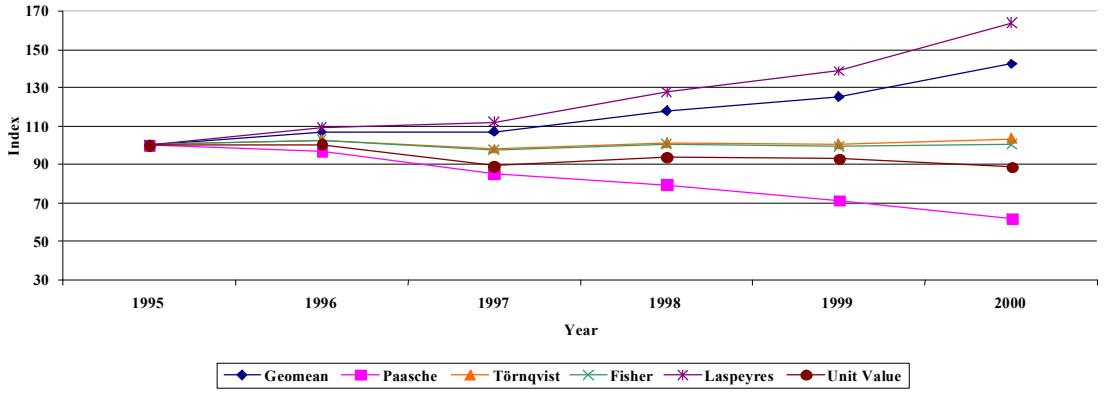

Figure 5. Annually Chained Population Target Indexes for Core Plain Cereals February-to-February Indexes, 1995 = 100.

Table 1
Direct and Chained Indexes for '95 – '00

		Geometric Mean	Paasche	Törnqvist	Fisher	Laspeyres	Unit Value
Hot	Direct	1.1176	1.0253	1.0847	1.0891	1.1569	0.9576
	Chained, All Items	1.1801	0.9874	1.1159	1.1216	1.2742	0.9453
	Chained, Core Items	1.1804	0.9865	1.1160	1.1221	1.2763	0.9759
Sugary	Direct	0.8855	0.6739	0.7913	0.7898	0.9257	0.7417
	Chained All Items	1.3341	0.3825	0.7925	0.7771	1.5786	0.7506
	Chained, Core Items	1.3591	0.3661	0.7849	0.7704	1.6212	0.7585
Fruity	Direct	0.9716	0.8676	0.9319	0.9296	0.9960	0.8932
	Chained All Items	1.2202	0.6849	0.9661	0.9696	1.3728	0.9308
	Chained, Core Items	1.1808	0.6557	0.9320	0.9328	1.3269	0.8950
Plain	Direct	1.0811	0.8641	1.0045	0.9816	1.1150	0.8554
	Chained All Items	1.3969	0.6330	1.0333	1.0053	1.5965	0.8935
	Chained, Core Items	1.4234	0.6175	1.0353	1.0054	1.6370	0.8879

Table 1
Direct and Chained Indexes for '95 – '00

		Geo-metric Mean*	Paasche	Törnqvist	Fisher	Las-peyres	Unit Value	RA	Geo-metric Mean+
Hot	Direct	1.1176	1.0253	1.0847	1.0891	1.1569	0.9576	1.1192	1.0949
	Chained, All Items	1.1801	0.9874	1.1159	1.1216	1.2742	0.9453	1.1395	1.1128
	Chained, Core Items	1.1804	0.9865	1.1160	1.1221	1.2763	0.9759	1.1374	1.1151
Sugary	Direct	0.8855	0.6739	0.7913	0.7898	0.9257	0.7417	0.8817	0.8702
	Chained All Items	1.3341	0.3825	0.7925	0.7771	1.5786	0.7506	0.9124	0.9010
	Chained, Core Items	1.3591	0.3661	0.7849	0.7704	1.6212	0.7585	0.8984	0.8894
Fruity	Direct	0.9716	0.8676	0.9319	0.9296	0.9960	0.8932	0.9815	0.9726
	Chained All Items	1.2202	0.6849	0.9661	0.9696	1.3728	0.9308	1.0263	1.0165
	Chained, Core Items	1.1808	0.6557	0.9320	0.9328	1.3269	0.8950	0.9935	0.9820
Plain	Direct	1.0811	0.8641	1.0045	0.9816	1.1150	0.8554	1.0620	1.0511
	Chained All Items	1.3969	0.6330	1.0333	1.0053	1.5965	0.8935	1.0642	1.0572
	Chained, Core Items	1.4234	0.6175	1.0353	1.0054	1.6370	0.8879	1.0653	1.0571

* Weighted by base period expenditure.
\+ Unweighted.

Based on this preliminary investigation, and for relative simplicity, we restricted our further investigations to the core data. To investigate the relative accuracy of probability and purposive sampling, as applied in practice to construct CPI's, we endeavored to approximate the sample designs used by the United States and the United Kingdom-representing probability based and judgment sampling respectively. In both cases we were fortunate to have detailed information on the complex survey processes, in the form of manuals, and contacts within the respective agencies. The basic idea was to repeatedly sample from a given population, for example the core transactions in the years '95 and '96. Each "run" was a composite of sampling and estimation activities carried out according to the methods of one country or the other. It should be borne in mind that our interest was in comparing the merits of *methodologies*, not in measuring the success of the US and UK in estimating their target population parameters.

The four "natural" population groups described above, referred to as "Major Groups" in the UK and "Expenditure Classes" in the US, were divided into finer sub-groups. In practice, sub-groups would be defined in terms of types of commodity. One justification for this, besides any intrinsic interest there might be in those commodities themselves, is that sub-groups so formed will tend to be homogeneous in their price trends. For purposes of this simulation study we therefore defined sub-groups as follows:

1) Long range price change for each of the 326 items in the core data were calculated, using unit value indexes for the items (across outlets) for '00 versus '95.
2) Noise was added to these indexes, items within a major group were sorted by their values of the perturbed index, and adjacent items were grouped together. The grouping of items with close long term indexes was to make the subgroups homogeneous, and the addition of noise was done so that the homogeneity would be realistically imperfect.

Table 2 gives the population item structure that was constructed, including the nomenclature in use in each of the two countries, the number of groups at each level of refinement, and the corresponding symbol for each class level used in this paper. The "Representative Item" is the lowest level at which an index is produced in the UK. This corresponds to the US's Entry Level Item (ELI), actually a collection of similar or related items. In the US, indexes are produced for categories one level up, *i.e.*, at the "Item Stratum" level, but these categories are further divided by the geographic areas in which the items are sold. Note that there are 2 or 3 Item Strata/Sections h in a Class/Major Group C, 3 ELI/Representative Items c per Item Stratum/Section h (except in one instance 2), and 10 or 11 items/varieties i in each ELI/Representative Item c. (*Note*: an actual UK class might be larger or smaller than the

corresponding US class; for example as a rule the ELI probably takes in more sorts of specific items than does the Representative Item. We had to force equivalence to ensure that the same amount of information was used in each approach. This adjustment will not affect our conclusions regarding the relative merits of the basic methods used in the two countries).

Table 2
Population Structure of "Cereal World": Items

UK	US	Number of Groups	Symbol
Major Group	Expenditure Class	4	C
Section	Item Strata	10	h
Representative Item	Entry Level Item (ELI)	29	c
Variety	Item	326	i

In addition to the item structure, each population of transactions has a "spatial" structure, characterizing where an item was sold. This structure is summarized in Table 3. Outlets belong to chains (*e.g.*, Safeway, Kroger), which cut across the three US geographic primary sampling units from which the cereal data were collected. (In the UK terminology, chains are called "multiples.) Outlets in a given chain share common ownership, with the exception of "Chain 8," which was a "catch-all" group consisting of outlets *not* belonging to a major chain (there may have been some "mini-chains"). In matching this "chain structure" to the classification of shops used in UK sampling, Chain 8 was considered a set of "independents" (the term used for independently-owned shops in the UK). Chain 4, which appeared to have the greatest homogeneity of pricing across outlets, was regarded as a "centrally collected multiple," the term used in the UK for groups of outlets with centrally controlled pricing. Each remaining chain was a non-centrally collected multiple. The manner of collection and estimation for each of these three types is given in the description of UK methodology below.

Thus the population consists of $N^{95} \approx 20{,}000$ records for '95 – '96 indexes, each record representing the purchase of an item i within an outlet j. Attached to each item/outlet are its PSU/Region l, its chain/shop-type k, the outlet/shop j, the item/variety i, the ELI/representative item c, the item stratum/section h, the expenditure class/major group C, and p^y, q^y, p^{y+1}, and q^{y+1}, the prices and quantities (in ounces) of the items sold in (February of) the two years in question. We used this population file (henceforth referred to simply as "the file") to simulate all phases of the US and UK operations.

3. Sampling Methodologies Simulated

The complicated sampling procedures we used to simulate the US and UK approaches are patterned on the respective practices of these two countries. These practices change over time, and have variants even at a given point in time. Our goal was not to determine which country does better, nor to encompass all variants. Rather it was to compare two distinct modes of sampling, with the range of complexity those modes entail. The interested reader can find a description of the US construction of the CPI in the *BLS Handbook of Methods* (2005), Chapter 17. For the UK's Retail Price Index (RPI), we relied on *The Retails Prices Index Technical Manual* (1998). A description of more current UK practice can be found in the *Consumer Price Indexes Technical Manual* (2005).

3.1 US Sampling Methodology

We first describe the US sampling methodology, which requires three surveys employing probability sampling: (1) a household survey, the Consumer Expenditure Survey (CEX), to estimate household allocation of expenditure to different categories of goods, (2) a second household survey, the Point of Purchase Survey (POPS) to estimate, within item groups, the relative amounts spent in different outlets, and (3) an outlets survey, through which individual items are selected and priced. In all three cases, sampling for the simulation is random with replacement (though the sampling employed in practice is considerably more complicated). The first two surveys are based on simple random samples, and the last is based on a probability proportional to size (*pps*) sample, where the size measures are a function of expenditures as estimated from the CEX and POPS. The sample for the third survey is a collection of items within outlet/ELI combinations.

Table 3
Population Structure of "Cereal World": Outlets

UK		US	#	Symbol
Region		Primary Sampling Unit	3	l
Shop type:	Independents	Chain 8		k
	Multiples: Central	Chain 4		
	Multiples: Non – central	Chains 1 – 3; 5 – 7		
Shop		Outlet	~300	j

3.1.1 CEX (Household Survey)

The goal is to estimate E_{lc}, the gross household expenditure on ELI c within PSU l. We sampled using simple random sampling with replacement (*srswr*) from the file described above, within PSU, in such a manner as to get unbiased expansion estimates

$$\hat{E}_{lc}^{95} = \frac{N_l^{95}}{n_{xl}} \sum_{j \in l \cap s(xl)} \sum_{i \in c \cap s(xl)} E_{ljci}^{95},$$

where $E_{lji}^{y} = q_{lji}^{y} p_{lji}^{y}$, N_l^{95} was the population size (number of records for *psu* l in '95 – '96), and n_{xl} was the sample size of the CEX sample $s(xl)$ in PSU l, chosen to match actual US CEX sample sizes and to achieve coefficients of variation of the estimates that approximated those achieved through the actual US CEX; the x in $s(xl)$ and n_{xl} is meant merely to differentiate the CEX from the POPS survey (which has a corresponding "p"; see below) or the prices survey. This "imitation CEX" was a simplified version of the actual survey. Our methodology tacitly assumed that all customers in a given outlet bought items in the same proportions; it did not allow for the inevitable measurement error that accompanies any actual expenditure survey, and (for '95 – '96) it was too current: real CEX data often predate by several years the outlet surveys for which they are used. Since, however, the "household data" collected were used in the corresponding UK methodology (see below) as well, the simplified version sufficed for the intended comparison of methodologies.

Higher level expenditures were estimated by simple addition. Thus, for example, the total expended across PSU's in a given ELI c is estimated by $\hat{E}_{.c}^{95} = \sum_l \hat{E}_{lc}^{95}$, *etc*. There were 500 CEX samples taken, each producing a corresponding set of expenditure estimates.

3.1.2 POPS (Household Survey)

Here the goal is to estimate the distribution of expenditures at different outlets for particular classes of goods. These classes could be ELI's or groups of ELI's; in the present study we assume they are the ELI's. The actual US TPOPS (Telephone Point of Purchase Survey) is, as its name suggests, conducted by phone, using a sample rotation scheme with a four-year cycle. We endeavored, as we did with the CEX, to match statistical properties of our procedure to the actual TPOPS, but it turned out that to match sample sizes on our file of 20,000 would have given larger than desirable sampling fractions within PSU's. We therefore cut the sample sizes in half – our "imitation POPS" should have precision about $1/\sqrt{2}$ of the actual TPOPS. Again, this modification will not affect the conclusions of this study, because we used the identical data in the UK construction. Samples $s(pl)$ of size n_{pl} were drawn by *srswr*, and estimation was by the expansion estimator:

$$\tilde{E}_{lcj}^{y} = \frac{N_l^{y}}{n_{pl}} \sum_{i \in c \cap s(pl)} E_{ljci}^{y}.$$

Since the POPS survey tends to be more up-to-date than the CEX, we allow y to be the base year of the index, '95 in '95 – '96, but '96 in '96 – '97, *etc*. There were 500 runs and sets of estimates, each to be matched with a CEX run.

3.1.3 Outlet Sampling

For each year y, selection of items from which to collect prices involves the following steps:

(a) For each PSU l, and each of the 10 item strata h, we select 2 ELIs c by probability proportional to size with replacement sampling (*ppswr*), with size measure $\hat{E}_{lc}^{95}$ derived from the CEX.

(b) For each ELI c selected, we select 8 outlets j by *ppswr*, using as size measure POPS expenditure estimates $\tilde{E}_{ljc}^{y}$. Thus altogether there are 160 ELI-outlet pairs per PSU, and 480 total, with a certain amount of repetition possible.

(c) Within outlet/ELI (j, c) we "go" (as the field representative would literally go) to the outlet and "list" all items belonging to the ELI and their corresponding first period expenditures E_{ljci}^{y}, and, with this within-outlet frame, sample 1 item by *pps*.

For each item so selected, we record the prices p_{ljhci}^{y}, $y = 1, 2$. Thus we note that all aspects of the outlet sampling are *pps* with replacement, based on estimates of expenditure from one or other of the 2 household surveys or from within the selected store. Again, we performed 500 runs, each run corresponding to a single CEX/POPS run.

3.2 US Estimation

"Elementary aggregates" $\hat{I}_{lh}^{y,\,y+1}$, index estimates at the PSU × Item stratum level, are the building blocks from which the CPI is constructed. In most CPI's around the world, the lowest level indexes are unweighted averages of one sort or another, as is the UK's RA estimator discussed below, and expenditure data are only used to aggregate these to higher levels. In the US, the elementary indexes are basically Horvitz-Thomson estimators relying explicitly or implicitly on expenditure estimates from both the CEX and POPS. In recent years, the US has for most item strata adopted a *geomean* formula (see Appendix A), so that estimates at this level take the form

$$\hat{I}_{lh}^{y,\,y+1} = \prod_{\substack{j\in l,\\ i\in c\in h\\ (i,\,j)\in s}} \left(\frac{p_{ljhci}^{y+1}}{p_{ljhci}^{y}} \right)^{s_{ljhci}},$$

where

$$s_{ljhci} = \frac{w_{ljhci}}{\sum_{\substack{j\in l,\, c\in h,\, i\in c\\ (i,\,j)\in s}} w_{ljhci}},$$

with

$$w_{ljhci} = \frac{\tilde{E}_{lc}\,\hat{E}_{lh}}{\hat{E}_{lc}}\, w_{ljhci},$$

$j \in l$, $i \in c \in h$ and $(i,\ j) \in s$. Note that the weights are not particular to the i^{th} item; we omit the time superscripts for brevity. They are not simply equal to the reciprocal of the number n_{lh} of sample items in lh, as sample unbiasedness considerations might lead one to expect (Balk 2003), because the sampling probabilities do not reflect exact base period expenditures on items; see the *BLS Handbook of Methods* (2005).

Then the elementary indexes are aggregated using estimated expenditures from the CEX according to a Laspeyres formula, for example

$$\hat{I}_{.h}^{y,\,y+1} = \frac{\sum_{l} \hat{E}_{lh}\,\hat{I}_{lh}^{y,\,y+1}}{\sum_{l} \hat{E}_{lh}}$$

to get the index for a given item stratum h, across PSU's.

3.3 UK Sampling Methodology

The UK, like the US, combines three components in its estimation methodology: (1) a household survey, the Family Expenditure Survey (FES), to get estimates of amounts spent on different item groups, (2) a shops survey, the Annual Retailing Inquiry (ARI) to get expenditure information by section and shop type, and (3) an outlet survey of shops, to select items for pricing.

3.3.1 FES (Household Survey)

The goal is to estimate expenditures $E_{..c}$ for representative items c, and $E_{l..h}$ expenditures for region/section combinations. For purposes of this study we will assume that the data for the US's CEX and the UK's FES coincide run by run, so there are, again, 500 FES data sets. Note that the UK does not aim at the more detailed estimates $E_{l..c}$ which the US targets.

3.3.2 Annual Retailing Inquiry (Shops Survey)

The goal is to get estimates of expenditures $\tilde{E}_{kh}$, by section and shop type. This is considerably broader than the outlet (shop) by ELI (representative item) that the US's POPS seeks. We use the same data, for each of 500 runs, to construct the ARI estimates that we used to construct the POPS estimates for the simulated US CPI.

3.3.3 Outlet Sampling

Selecting items from which to collect prices involves the following steps:

(a) A "judgment sample" of representative items c is selected within each section h. In the present study (only to allow for simulation), within each section, we select the two representative items with largest values of $\hat{E}_{hc}$. Note two differences from the corresponding step (a) of the US method: (i) selection is uniform across all regions l; (ii) selection is not random, and, in particular, it does not allow for duplication of representative items. (Duplication can occur in the simulated US method, due to with replacement sampling of ELI's within item strata.)

(b) The field economists select the shops in a particular locale in which to price a given representative item. Traditionally, this was *srswor,* after the field economist had constructed a frame of appropriate shops. More recently, selection has been by *pps*, where the size measure is floor space dedicated to the type of goods the representative item represents. Field economists do not draw samples of "centrally collected" items: in the case of a very large multiple, the price of an item is collected from the multiple's central office, and taken to represent the price of that item in all shops in the multiple. In the present study we proceeded as follows: for each region l and representative item c, we selected 8 shops as follows:

- 4 from non-central multiples (Chains 1, 2, 3, 5, 6, 7)
- 1 from a central multiple (Chain 4)
- 3 from independents (Chain 8)

In each case, for simplicity, we used *srs* without replacement from shops having positive expenditure for the representative item. The number of shops in the UK (8 per representative item in each region) matches the number of "outlets" in the US; there are 160 shop/representative item pairs per region, or 480 in total. Note the following differences from the U.S. methodology:

1. Information on shop type is being used for stratification (and will play a role in estimation below). This information is available in the US sample but is bypassed in favor of the *pps* methodology.
2. We are allowing the UK to have information about the presence or absence of the specific representative item *c* (equivalent to the ELI) in the list of shops before sampling, whereas the US only in effect knows of the existence of *some* ELI in the given item stratum. (This assumes a multiple ELI-to-POPS category mapping, which was typically the case until recently in US operations; the current version of ELI-to-TPOPS (telephone point of purchase survey) category mappings is 1 to 1; that is, an outlet frame is constructed for each individual ELI.)

(c) Traditionally, for each representative item *c*, within a given shop, the field economist selects that variety *i* which he/she regards as dominating its sales – a judgment sample of the most consistently purchased variety. We formalize this as follows:

1. For a given shop/representative item pair (j, c), we list all varieties *i*.
2. For each variety, we find the minimum quantity $q_i^* = \text{Min}(q_i^y, q_i^{y+1})$ over two years.
3. We sample the variety *i* with $\text{Max}\{q_i^*\}$.

This process, of course, requires more information than a field economist would have at the earlier time period (and again is not used in the US sampling described above) but may be regarded as providing a surrogate for the field economist's appraisal of the relative continuity of goods sold.

Note: It is convenient to refer to the combination of selecting an outlet by *srswor* as in (b), and an item within the shop as described in (c), as *maxminq sampling*.

3.4 UK Estimation

Elementary aggregates for the UK were calculated by a Ratio-of-Averages (RA) formula within each cross-classification cell defined by region, shop type, and representative item This is basically an unweighted estimate, given for independent shops by

$$\hat{I}_{lkhc}^{y,\,y+1} = \frac{\frac{1}{n}\sum_{i\in c,\ j\in k} p_{ljhci}^{y+1}}{\frac{1}{n}\sum_{i\in c,\ j\in k} p_{ljhci}^{y}}.$$

In the case of multiples, a weighted version of the above formula is used with expenditures by shop type, estimated from the ARI, providing relative weights of central versus non-central multiples.

A countrywide index for representative items c in the sample (aggregated over shop types k and regions l) is then calculated by a Laspeyres type estimator:

$$\hat{I}_c^{y,\,y+1} = \sum_l \sum_k \tilde{w}_{lkhc}\, \hat{I}_{lkhc}^{y,\,y+1},$$

where $c \in h$, and $\tilde{w}_{lkhc}$ is based on $\tilde{E}_{.kh}^{y}$ from the ARI and $\hat{E}_{lh}^{95}$ from the FES (using these time periods keeps information used the same between US and UK). Further aggregation (over representative items *c*) is done using $\hat{E}_{hc}^{y}$, *etc*. from the FES.

3.5 Comparison

Table 4 gives a summary comparison of the two methodologies, US and UK, that we have been considering. The predominant feature of the US method is strict probability sampling and estimation, typically *ppswr*; that of the UK is selective sampling, taking the most important item or category as judged by expenditure or quantity sold. The methods of forming elementary aggregates are different, and the weights for aggregation in the UK are estimated at a slightly coarser level at the lower stages.

Table 5 gives a summary of what might be considered the strengths and weaknesses of the US and UK methodologies. By the advantage of "brute strength," which we attribute to the UK approach, we mean the capitalizing on a combination of two factors that often play a role in pricing and price index construction. In the first place, market leaders tend to dominate the price scene; for example, if they sharply lower or raise prices, their lesser competitors selling similar goods may think it necessary or warranted to follow suit. Secondly, even if there is variation in the price trends among similar goods, the leading sellers are likely to dominate the price index by virtue of large expenditure values, that is, because of their correspondingly large weights.

Table 4
Summary Comparison of US and UK Methodologies

	US	UK
HH Exp. Survey	$\hat{E}^{95}_{lc}$	$\hat{E}^{95}_{.c}$, $\hat{E}^{95}_{lh}$
Outlet Exp./Category	HH(POPS) $\tilde{E}^{y}_{ljc}$	Shops Survey (ARI) $\tilde{E}^{y}_{.kh}$
select item categories	2 ELI's *c*/item stratum *h*/PSU *l* *ppswr* $(\hat{E}^{95}_{lc} / \hat{E}^{95}_{lh})$	2 rep. items's *c*/section *h*/Region *l* *largest* $(\hat{E}^{95}_{.c} / \hat{E}^{95}_{.h})$
select outlets	8 outlets *j*/ELI $c \times$ PSU *l* *ppswr* $(\tilde{E}^{y}_{ljc} / \tilde{E}^{y}_{l.c})$	8 outlets *j*/rep. item $c \times$ *Region* $l - srs$ within shoptype *k*, $E^{y}_{ljc} > 0$
item within outlet/category	1 item *i*/*jc pps* $(E^{y}_{ljci} / E^{y}_{ljc.})$	1 variety *i*/*jc* $\max[\mathrm{Min}(q^{y}_{ji}, q^{y+1}_{ji})]$
elementary index	$\hat{I}^{y,\,y+1}_{lh} = \prod_{\substack{j\in l,\\ i\in c\in h\\ (i,\,j)\in s}} \left(\frac{p^{y+1}_{ljhci}}{p^{y}_{ljhci}}\right)^{s_{ljhci}}$	$\hat{I}^{y,\,y+1}_{lkhc} = \frac{\frac{1}{n}\sum_{i\in c,\ j\in k} p^{y+1}_{ljhci}}{\frac{1}{n}\sum_{i\in c,\ j\in k} p^{y}_{ljhci}}$
higher aggregation	$\hat{I}^{y,\,y+1}_{.h} = \frac{\sum_{l} \hat{E}_{lh}\, \hat{I}^{y,\,y+1}_{lh}}{\sum_{l} \hat{E}_{lh}}$	$\hat{I}_{c} = \sum_{l}\sum_{k} \hat{\hat{w}}_{lkhc} \hat{I}_{lkhc}$, $\hat{\hat{w}}_{lkhc} = f(\tilde{E}_{.kh}, \hat{E}_{l.h})$

Table 5
Comparison of US, UK Approaches

	Strengths	Weaknesses
US	Gather more information More use of information Satisfies classical sampling theory Gives regional (PSU) estimates Weighted estimators at lowest level More standardized operating procedure	Possible repetition in selection Ignores stratification of shops (that is, classification into chains)
UK	Relies on "Brute Force" principle Stratification of outlets Shops survey in field Uses variety of sources	Patchwork of weights Inconsistent in Centralized pricing aggregation? Unweighted and seemingly arbitrary estimator at lowest level

4. Results of Primary Study

Indexes comparing '95 to '96 are given in Table 6, for the population (1) as a whole (the three areas combined), (2) broken down by classes/major groups, and (3) broken down even further into item strata/sections. Four indexes are given which might be taken as the targets of estimation. Recall the discussion on targets which concludes Section 1.

Table 7 gives corresponding means, variances, and mean square errors for US and UK estimates, where the mean square error is computed with respect to the Fisher indexes. We observe the following:

1) For the all-items, classes, and item strata, the US estimates appear to approximate the *geomean* G. This confirms what we have suspected from other work (Dorfman *et al*. 1999), namely that the lowest level of aggregation dominates (we used a Laspeyres formula for higher level aggregation). The fact that G lies between the Laspeyres and superlative target provides some evidence that the US switch to this method of elementary aggregation was a step in the right direction.
2) There appears to be no clear order relation of UK *Section* estimates to their corresponding targets; for example, the Section 11 index is higher than the target L, while the Section 12 index is lower

than the superlatives, *etc*. As we aggregate up to the Major Group and All Items levels, however, the estimates clearly begin to approximate the superlatives *F* or *T*. (Dalén (1998) noted a similar result in aggregating cut-off samples.)

3) If we take the Fisher as the target, *even at* the section level, the root mean square error of the UK estimator is much lower than that of the US estimator. Given the relatively restricted nature of the UK sample design, it is not surprising that the UK estimator displays lower variance, but the form of the UK estimator would not lead one to expect it to unbiasedly approximate a Fisher index. Nonetheless, our results suggest that, at least for a population of purchases such as the one used in this study, the purposive, "brute force" methods of the UK (and many other countries) work well.

Similar results were found for the succeeding pairs of years through '99 – '00. Figure 6 shows the all items year-to-year *geomean* and Fisher for five pairs of years and the means across samples of the corresponding US and UK estimators. (Note the difference in scale between Figure 6 and Figures 1 through 5). It is readily seen that the U.S. estimator tends to track the population *geomean*. The UK estimator, tracking the Fisher, tends to overestimate in the later years, although it runs much closer to the Fisher than to the population *geomean*. It should be noted that we used increasingly out-of-date expenditure data, namely the '95 data, for purposes of sampling and estimation. It is possible that outmoded expenditure data are having a greater impact on the UK estimates than on the US estimates, perhaps by leading us to oversample expensive representative items or to focus on some group of shops that are increasingly pricey.

Results for the classes ("hot," *etc*.) were very similar for the US vis-à-vis the *geomean* and are not shown. Figure 7 shows the difference between the mean year-to-year UK estimates and the Fisher, for each of the four classes. It can be seen that the tendency to overestimate in the later years affects all four classes.

Overall, the UK estimators provide better estimates of the superlative Fisher target than do the US estimators. Table 8 gives the ratio of UK root mean square error to US root mean square error, for all five pairs of years, for all items, for groups, and for sections. There are a few anomalous places, notably in the '98 – '99 indexes where, for section 2 of "hot," and consequently for the entire class "hot," the UK estimates are appreciably worse. In general, however, the UK methods provide much better estimates. This is due in part to a tighter sampling structure (mainly because purposive/cutoff sampling is much more restrictive than random selection of the set of items which can enter the sample), yielding, not surprisingly, less variance. In part though, as well, it is due to a surprising tendency of the UK estimators to target the corresponding Fisher indexes, reducing bias. Since the UK estimators do not formally resemble the Fisher index, the reasons for their tendency to approximate it merit further study. We turn to this issue in the next section.

Table 6
Potential Target '95 – '96 Indexes

Description	*geomean*	Törnqvist	Fisher	Laspeyres
All	1.053	1.002	0.997	1.079
Classes/Major Groups				
1 – Hot	1.058	1.052	1.052	1.078
2 – Sugary	1.042	0.964	0.956	1.072
3 – Fruity	1.044	1.007	1.007	1.067
4 – Plain	1.069	1.027	1.027	1.092
Item Strata/Sections				
Hot – 11	1.043	1.044	1.044	1.057
Hot – 12	1.073	1.059	1.058	1.097
Sugary – 21	1.003	0.917	0.910	1.034
Sugary – 22	1.063	0.982	0.972	1.093
Sugary – 23	1.093	1.052	1.054	1.119
Fruity – 31	0.977	0.955	0.950	0.985
Fruity – 32	1.165	1.110	1.116	1.204
Plain – 41	1.067	1.021	1.021	1.094
Plain – 42	1.030	0.996	0.996	1.050
Plain – 43	1.104	1.063	1.062	1.125

Table 7
Simulation Results for '95 – '96 Indexes

Description	Target Index	U.S. Mean	U.S. Std. Dev.	U.S. RMSE	U.K. Mean	U.K. Std. Dev.	U.K. RMSE
All	0.997	1.057	0.016	0.062	1.002	0.011	0.012
Classes/Major Groups							
1 – Hot	1.052	1.059	0.031	0.032	1.045	0.022	0.023
2 – Sugary	0.956	1.046	0.030	0.095	0.971	0.023	0.027
3 – Fruity	1.007	1.053	0.035	0.058	0.986	0.027	0.034
4 – Plain	1.027	1.072	0.025	0.051	1.025	0.016	0.016
Item Strata/Sections							
Hot – 11	1.044	1.045	0.035	0.035	1.064	0.025	0.032
Hot – 12	1.058	1.072	0.049	0.051	1.027	0.035	0.047
Sugary – 21	0.910	1.004	0.050	0.106	0.850	0.045	0.074
Sugary – 22	0.972	1.070	0.051	0.111	1.089	0.030	0.121
Sugary – 23	1.054	1.095	0.044	0.060	1.026	0.027	0.039
Fruity – 31	0.950	0.979	0.020	0.035	0.932	0.020	0.027
Fruity – 32	1.116	1.178	0.084	0.104	1.077	0.059	0.071
Plain – 41	1.021	1.069	0.050	0.070	1.060	0.030	0.049
Plain – 42	0.996	1.033	0.035	0.051	0.987	0.031	0.032
Plain – 43	1.062	1.107	0.042	0.061	1.028	0.023	0.041

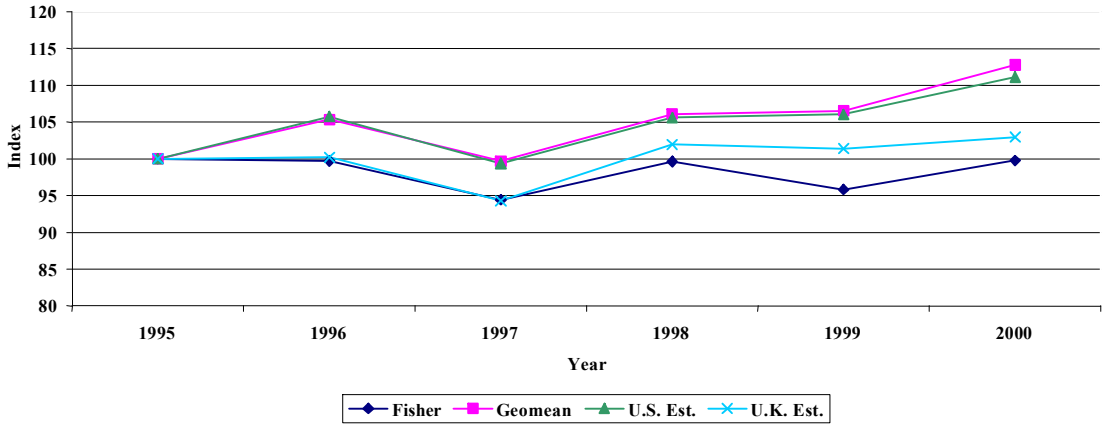

Figure 6. Index Targets and Estimates for All Cereals February-to-February Indexes and Index Estimates, 1995 = 100.

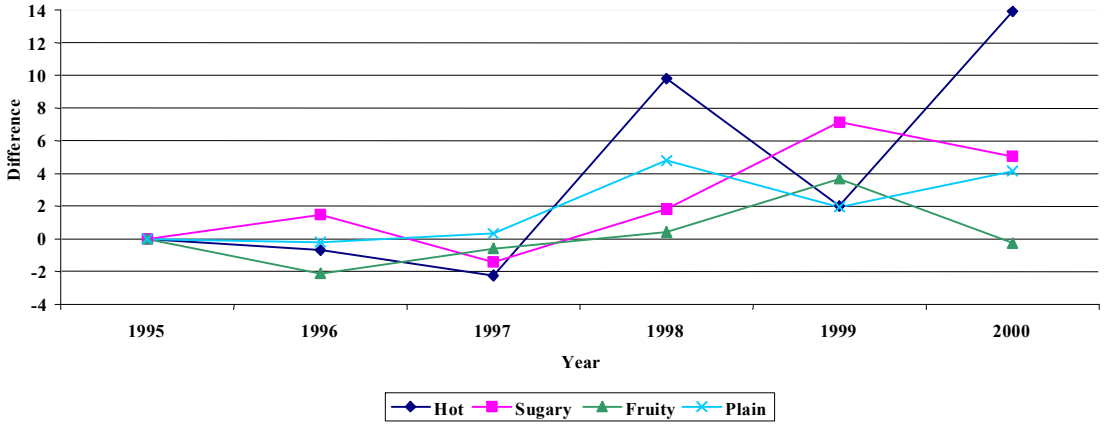

Figure 7. Differences Between U.K. Estimates and Population Fisher Indexes February-to-February Indexes and Index Estimates, 1995 = 100.

Table 8
Ratios of UK RMSE to US RMSE

Description	'95 – '96	'96 – '97	'97 – '98	'98 – '99	'99 – '00
All	0.196	0.192	0.419	0.548	0.288
Classes/Major Groups					
1 – Hot	0.713	0.517	0.483	1.437	0.589
2 – Sugary	0.286	0.336	0.314	0.522	0.282
3 – Fruity	0.595	0.508	0.308	0.501	0.405
4 – Plain	0.310	0.297	0.777	0.319	0.404
Item Strata/Sections					
Hot – 11	0.923	1.066	0.682	0.529	0.508
Hot – 12	0.920	0.850	1.169	1.860	0.842
Sugary – 21	0.702	0.392	0.421	0.595	0.330
Sugary – 22	1.092	0.426	0.380	0.341	0.365
Sugary – 23	0.650	0.455	0.448	0.925	0.851
Fruity – 31	0.778	1.059	0.637	0.581	0.618
Fruity – 32	0.683	0.809	0.314	0.457	0.356
Plain – 41	0.709	0.623	0.494	0.567	0.317
Plain – 42	0.642	0.511	1.117	1.092	1.005
Plain – 43	0.678	0.839	0.641	0.815	0.701

5. Follow-Up Study

There are four aspects in which the approaches of the UK and US differ: (1) the stratification structure, in particular, the reliance of the UK on different shops strata and, to an extent, on centralized sampling, (2) the aggregation and weighting structure, (3) the mode of sampling at different stages, and (4) the formula for elementary aggregates. This makes it difficult to disentangle the extent to which each aspect is contributing to the relative merits of US and UK index construction. In particular, as noted in the last section, it is a bit of a mystery why, especially at higher aggregations, the UK index estimator tends to target the superlative indexes.

In our follow-up study we focus on the lowest level of index construction, that is, on (3), the shop-representative item (ELI) level of sampling and on (4), the formulas for the elementary indexes. We compare the relative merits of different options, taking the within area elementary indexes as our targets. Aggregation to higher level indexes will be carried out uniformly for all alternative lower level options considered, using the true population expenditure shares. The importance of the method of construction of the elementary indexes is widely recognized; see Diewert (2004) and references; also Dorfman *et al.* (1999). The example discussed in Appendix B, with results given in Table 9, illustrates the decisive effect that the lowest level of index construction has on the index as a whole.

Thus, a likely important source of the difference in results of US and UK methodology lies in the sample estimation of the population elementary indexes. But this leaves open the question whether the differences arise because of differences in sampling method or in the formulas used in estimation, or in both. Thus we are interested in determining: (1) how judgment sampling (in this case, cutoff sampling based on *maxminq*) performs compared to probability sampling represented by *ppswr*, holding the estimator of the elementary indexes fixed, and (2) how estimators of elementary indexes compare when we keep the sampling method fixed. It will also be of some interest to determine what happens when *maxminq* sampling is based on data from the base and *previous* time period, rather than the base and current period.

5.1 Sampling Methods and Estimators at the Elementary Level

To explore these questions, we carried out further simulation studies. The data were the same Cereal Data used in the primary study (successive Februarys), but limited to the Independent Shops, *Chain 8*. This was done to make the study more manageable but also because, for the other chains, the UK elementary index estimators were more complicated than the simple *dutot*. Also, it is reasonable to expect price behavior to be most heterogeneous in this chain, so that inherent differences will be clearer. Chain 8 was the largest of the chains, comprising each year about 30% of the whole population, approximately 6,000 records.

The basic structure remained the same: 3 *psu*'s, 4 major groups/expenditure classes (hot, sugary, fruity, and plain), 10 sections/item strata, and 29 representative items/*ELI*'s. For each *ELI/representative item*, 3 outlets (one item per outlet) were selected, as opposed to 10 in the primary study above. For investigating *maxminq* based on previous time periods, the original 5 data sets, each using price and quantity data for a pair of years ('95/'96, '96/'97, *etc.*) were reduced to include only items that allowed "back matching";

that is, matching across three years to compare prices of items in outlets for '95/'96/'97, '96/'97/'98, *etc*. About 90% of the Chain 8 records allowed back matching. (In considering the results below, it is probably worth noting that the sample reduction could disproportionately impact the back matched *maxminq*). We shift our attention from the Fisher index to the superlative Walsh index, due to an astute suggestion of a referee, discussed in Appendix C.

Three estimators were used for elementary indexes: the ratio of averages (RA) (the *dutot*), the unweighted *geomean* (also known as the Jevons), and the average of ratios (*AR*). In the *pps* sampling of outlets, and then in the sampling of items within outlets, the size variable (expenditure) was assumed known (rather than being estimated, as in the main study). Besides *pps* with replacement (as in the US approach), and *maxminq*, we also investigated *pps* without replacement, on the suspicion it would be less variable than the with replacement version.

For each mode of sampling, within each *psu/ELI* combination, we took 500 samples. We calculated the mean square error of estimates with respect to a target *ELI* – level Walsh Index. Averages of *mse* across *ELI*'s were calculated for each mode of sampling/estimation, within each *psu.*

Table 10 shows the ratio of these averages to the average *mse* for the *maxminq/dutot* combination. For each estimator, for each *psu*, with one exception (*psu* 3, '99/'00), *maxminq* leads to lower *mse*, often by an appreciable margin. Sampling *pps* without replacement is second best. Holding the method of sampling fixed (comparing rows 1, 4, 7, then 2, 5, 8, *etc*. in Table 10), we note that with few exceptions, the *dutot* does better than the *geomean*, which does better than *AR*. These results suggest: (1) *maxminq* is better than *pps*(exp), and *pps*(exp) is better than *ppswr*(exp). (2) The *dutot* is more efficient than the *geomean*, and the *geomean* is more efficient than an average of ratios. There is a beneficial synergism between *maxminq* sampling and the *dutot.* Biases and variances were also studied, and the results (not shown) tended to follow the same pattern.

Table 9
Population Indexes '95 – '96, Chain 8

Description	Laspeyres	*geomean**	Fisher	Walsh	Laspeyres of Walsh Elementary
All	1.129	1.091	1.028	1.030	1.040
Classes/Major Groups					
1 – Hot	1.161	1.115	1.080	1.082	1.084
2 – Sugary	1.129	1.088	1.007	1.012	1.025
3 – Fruity	1.084	1.054	0.997	1.005	1.015
4 – Plain	1.135	1.101	1.046	1.042	1.050
Item Strata/Sections					
Hot – 11	1.157	1.117	1.088	1.089	1.090
Hot – 12	1.164	1.113	1.072	1.075	1.079
Sugary – 21	1.086	1.045	0.962	0.970	0.992
Sugary – 22	1.187	1.142	1.055	1.056	1.058
Sugary – 23	1.117	1.091	1.034	1.039	1.043
Fruity – 31	1.003	0.992	0.949	0.965	0.966
Fruity – 32	1.228	1.172	1.100	1.091	1.102
Plain – 41	1.212	1.161	1.091	1.080	1.090
Plain – 42	1.048	1.030	0.997	0.997	0.998
Plain – 43	1.136	1.107	1.048	1.046	1.056

* Weighted by base period expenditure.

Table 10
Standardized Average Relative Mean Square Error Across ELI's, Reduced Populations, Chain 8

	psu 2				*psu* 3				*psu* 4			
estimator/sampling method	'96 – '97	'97 – '98	'98 – '99	'99 – '00	'96 – '97	'97 – '98	'98 – '99	'99 – '00	'96 – '97	'97 – '98	'98 – '99	'99 – '00
dutot/*maxminq*(UK)	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00	1.00
dutot/*ppswor*	1.73	1.70	1.68	1.91	1.23	1.82	1.35	2.24	1.22	1.06	1.12	0.93
dutot/*ppswr*	2.13	2.10	1.91	2.14	1.42	2.10	1.46	2.67	1.45	1.23	1.36	1.07
geomean/*maxminq*	1.20	1.16	1.16	1.06	1.06	1.14	1.08	1.05	1.10	1.11	1.12	0.96
geomean/*ppswor*	2.08	1.88	1.98	2.27	1.33	1.94	1.47	2.59	1.33	1.09	1.28	0.97
geomean/*ppswr* (US)	2.49	2.29	2.18	2.53	1.58	2.23	1.58	3.09	1.59	1.30	1.52	1.12
AR/*maxminq*	1.42	1.32	1.31	1.14	1.24	1.03	1.30	1.05	1.11	1.20	1.21	1.07
AR/*ppswor*	2.81	2.35	2.49	2.85	1.70	2.31	1.77	3.43	1.57	1.30	1.42	1.17
AR/*ppswr*	3.23	2.77	2.66	3.08	2.03	2.58	1.87	3.96	1.83	1.49	1.66	1.30
dutot/*maxminq*, prior *q*'s	1.12	1.19	1.19	1.41	1.56	1.42	1.69	1.51	1.20	1.02	0.85	1.48

That the *dutot* sample index can target the Walsh population index (and hence indirectly any superlative index), when consistently largest sellers are sampled is, we suggest, the result of a very simple, "brute force" mechanism: to the extent that the Walsh can be represented by a small sample of items, it is best represented by those with the consistently largest quantities, and it is these items that the *maxminq* sampling scheme virtually always supplies. In Appendix C we discuss an alternative explanation for the good performance of the *maxminq/dutot* combination.

Average mean square errors were also calculated for the *maxminq/dutot* combination based on *previous* values of q, that is on q_i^{y-1}, q_i^y. Results are given in the last row of Table 10. There is an anticipated weakening compared to the updated *maxminq/dutot*, but the results still compare favorably to the other options. We study this further in subsection 5.2.

5.2 Effect of Lagged Quantities on *maxminq* Sampling

To put the results of Section 4 in perspective, we need to inquire what the effect is of using lagged q's in *maxminq*. The reason is simple: although at first sight, using base and current period quantities seems the obvious way to capture the UK's idea of persistent items, nonetheless, this involves using information (the current period quantities) which was not used in simulating US sampling. Perhaps this gives the UK methodology an unfair edge.

We therefore compared the US approach, viz. *ppswr* (with size variable being base period expenditure) and *geomean* at the elementary level, with the UK approach represented by *maxminq – dutot*, but now with *maxminq* based on quantities q_{y-1} and q_y. Data sets were reduced slightly to guarantee that we would have matching data for three consecutive years. Aggregation to upper level indexes used actual population expenditures for both US and UK.

Table 11 gives results for the All Cereals Indexes for chain 8, comparing biases, standard deviations, and root mean square errors with respect to the population Walsh. As expected, the results are not as good as those obtained by using current q's. Nonetheless, with respect to all three accuracy measures (bias, standard deviation, and root mean square error), the UK *maxminq/dutot* combination still does better than the US approach representing probability sampling.

For finer categories, Table 12 gives the ratios of mean square errors obtained under the UK method with lagged q's to those obtained under the US method. Although they are generally larger than those in Table 8, they still suggest that the purposive sampling approach of the UK is better.

6. Discussion

We have presented a comparison of two fundamentally different approaches to sample design and inference for a consumer price index. The inescapable conclusion is that, in the population we studied, the "UK" approach, which involves tighter stratification and, more importantly, more restrictive judgment sampling within strata than the probability sampling of the "US" approach, does better in estimating a target superlative index.

This is shown to be the case, whichever low level price index estimator (the *dutot*, or *geomean*, or the average of ratios) is employed, although the *dutot* (ratio of averages) performed best.

The UK approach does better for two reasons: (1) its tighter sampling, restrictive of items selected (for example, see Table 13 described in Appendix C), leads, not surprisingly, to lower variance, an observation made already in de Haan *et al*. (1999), and (2) the *dutot* sample indexes target the superlative indexes under dominant market sampling, which *was* surprising and called forth the investtigation described in Section 5. On the other hand, the US approach yielded an index estimator which could be described as unbiased, but it was unbiased for the (wrong) population *geometric* index weighted by first period expenditure. Thus it tended to run considerably higher than the target superlative index (whether Fisher, Walsh, or Törnqvist).

If sample sizes were allowed to increase, we could anticipate that the variances of both the US and UK would decrease, but the UK variance would remain lower. The bias of the US estimator for the superlative target would be unaffected by increased sample size, so that the relative mean square error of the UK approach would be increaseingly lower.

In practice, of course, period 2 quantities are not available at the time of sample selection (at period 1), and as part of our follow-up study we give some measure of the partial degradation that arises from using past quantities: it is not severe enough to undo the conclusion of better UK performance. Furthermore, the field economist's judgment as to the best seller might be able to invoke data more recent than a year earlier. Thus the actual effect might be somewhere between the lagged and non-lagged versions of *maxminq* which we have used. In practice, however, US field economists may often sample items within outlets based on an estimate of expenditure share that is really a smoothed average of base period *and* recent expenditure shares. This may attenuate the bias we have seen in our simulations, where only the base period expenditures were used for within-store sampling.

Table 11
Biases, Standard Deviations, and Root Mean Square Error (Each Multiplied by 1,000), in Estimating Population All Cereals Walsh Index, Chain 8, Based on Three Approaches to Sampling/Estimating Elementary Indexes*

	(a) Bias				
	'95–'96	'96–'97	'97–'98	'98–'99	'99–'00
dutot/maxminq	29	15	–13	33	2
dutot/maxminq, prior q's	–	46	32	82	36
geomean/ppswr	78	62	66	82	66
	(b) Standard Deviation				
	'95–'96	'96–'97	'97–'98	'98–'99	'99–'00
dutot/maxminq	16	13	11	14	12
dutot/maxminq, prior q's	–	14	12	15	14
geomean/ppswr	22	18	17	18	20
	(c) Root Mean Square Error				
	'95–'96	'96–'97	'97–'98	'98–'99	'99–'00
dutot/maxminq	33	20	17	36	12
dutot/maxminq, prior q's	–	48	34	83	39
geomean/ppswr	80	65	68	84	68

* At ELI/Representative Item level. To get overall index estimates, the elemenentary index estimates were aggregated using known population expenditures.

Table 12
Ratios of UK RMSE to US RMSE, Chain 8, Walsh Targets:
maxminq Using Lagged *q*'s & *dutot* Versus *ppswr*(Expenditure) & *geomean*

Description	'96–'97	'97–'98	'98–'99	'99–'00
All	0.748	0.498	0.993	0.567
Classes/Major Groups				
1–Hot	1.539	0.495	1.280	0.765
2–Sugary	0.563	0.676	0.941	0.797
3–Fruity	0.409	0.323	0.463	0.852
4–Plain	0.915	0.560	1.164	0.359
Item Strata/Sections				
Hot–11	0.748	0.607	0.660	0.657
Hot–12	1.695	0.599	1.333	0.843
Sugary–21	0.757	0.593	1.136	0.924
Sugary–22	0.370	0.776	0.751	0.671
Sugary–23	0.479	0.785	0.796	0.508
Fruity–31	0.570	0.443	0.678	1.008
Fruity–32	0.526	0.350	0.277	0.674
Plain–41	1.167	0.509	1.395	0.397
Plain–42	0.623	0.411	0.918	0.624
Plain–43	0.919	1.171	0.668	0.560

Table 13
Items Selected by *maxminq* and *pps*($\sqrt{q_y q_{y+1}}$) in 500 Samples

'95 – '96, Chain 8, *psu* 2, ELI 105											
pps	items selected	2889	2803	1564	2763	1558	2242	2344	2776	760	2850
	% of samples in which selected	43.2	32.2	10.4	5.4	3.87	1.53	1.33	0.87	0.8	0.4
maxminq	items selected	2889	2803								
	% of samples in which selected	80.87	19.13								
'95 – '96, Chain 8, *psu* 3, ELI 401											
pps	items selected	1731	2378	2866	1742	2922	2375	2528	403	871	
	% of samples in which selected	33.27	18.8	12.8	12.73	9.47	4.6	4.27	2.8	1.27	
maxminq	items selected	2378	1731	2866	1742						
	% of samples in which selected	46.27	24.47	15	14.27						
'99 – '00, Chain 8, *psu* 4, ELI 401											
pps	items selected	1731	2866	1742	2378	2922	2528	403			
	% of samples in which selected	30.07	21.93	14.3	11.07	9.53	6.8	6.27			
maxminq	items selected	1742	2866	2922	1731						
	% of samples in which selected	34.27	30.87	18	16.87						

It is generally accepted that the non-randomization approaches are intrinsically cheaper. For example, there are typically fewer outlets to visit, and price collection within outlets is less labor intensive. Thus, for a given budget we can expect the UK approach to be more efficient, compared to US probability sampling, than the present study suggests.

It would be salutary to expand this study to scanner data for products other than cereals. In particular, items with more volatile price movements would be of great interest. To some extent, the good behavior of *maxminq/dutot* may be related to the surprising closeness of the population *dutot* to the superlatives (as seen in Table 1). How typical is such closeness, and, if it is absent, will the good sampling behavior persist?

One final *caveat*. It may be a good idea in practice to inject a dose of randomness at some stage or stages of the sampling process, and in particular be a bit cautious about centralized sampling – not for statistical reasons, but to guarantee fairness and the appearance of fairness (Reinsdorf and Triplett 2005, Section II; Royall 1976).

Acknowledgements

The opinions expressed in this paper are those of the authors and do not represent US Bureau of Labor Statistics or Bureau of Transportation Statistics policy. The authors thank David Richardson and Lyuba Rozental for providing us with the cereal data and for timely assistance, Sonja Mapes and Scott Pinkerton for their work on the classification of cereals into types, and Mick Silver, Adrian Ball, and Dawn Camus for providing understanding and materials on the United Kingdom's RPI methods. The authors wish also to thank three referees and an associate editor for many insightful comments and for encouraging us to expand the study, and J. De Haan, M. Reinsdorf, and B. Moulton for their helpful suggestions. We especially wish to acknowledge the encouragement of the late M.P. Singh whose suggestions as Editor guided the final course this paper has taken.

Appendix A
Targets – Population Indexes

Laspeyres $$L = \frac{\sum_i q_i^y p_i^{y+1}}{\sum_i q_i^y p_i^y}$$

Paasche $$P = \frac{\sum_i q_i^{y+1} p_i^{y+1}}{\sum_i q_i^{y+1} p_i^y}$$

Walsh $$W = \frac{\sum_i \sqrt{q_i^y q_i^{y+1}}\, p_i^{y+1}}{\sum_i \sqrt{q_i^y q_i^{y+1}}\, p_i^y}$$

Fisher $$F = \left\{ \frac{\sum_i q_i^y p_i^{y+1}}{\sum_i q_i^y p_i^y} \frac{\sum_i q_i^{y+1} p_i^{y+1}}{\sum_i q_i^{y+1} p_i^y} \right\}^{1/2} = \sqrt{LP}$$

Törnqvist $$T = \prod_i \left(\frac{p_i^{y+1}}{p_i^y} \right)^{s_i^{y,\,y+1}},$$

where

$$s_i^{y,\,y+1} = \frac{1}{2}\left(\frac{p_i^y q_i^y}{\sum_i p_i^y q_i^y} + \frac{p_i^{y+1} q_i^{y+1}}{\sum_i p_i^{y+1} q_i^{y+1}} \right)$$

Geometric Mean $$G = \prod_i \left(\frac{p_i^{y+1}}{p_i^y} \right),$$

where

$$w_i = s_i^y = \left(\frac{p_i^y q_i^y}{\sum_i p_i^y q_i^y} \right)$$

or

$$w_i = 1/N$$

Unit Value $$U = \frac{\sum_i q_i^{y+1} p_i^{y+1} \Big/ \sum_i q_i^{y+1}}{\sum_i q_i^y p_i^y \Big/ \sum_i q_i^y}$$

dutot $$RA = \frac{\sum_i p_i^{y+1} \Big/ N}{\sum_i p_i^y \Big/ N}$$ ("Ratio of Averages")

Average of Ratios $$AR = \frac{\sum_i p_i^{y+1} \Big/ p_i^y}{N}$$

Appendix B
An Example Illustrating the Importance of Lowest Level Aggregation

We here present a simple example to illustrate the importance of the method used for constructing the elementary indexes. We compare population Walsh indexes to indexes resulting from aggregating elementary Walsh indexes according to a Laspeyres formula instead. The reason for focusing on the Walsh is given in Appendix C. The "pure" Walsh index is

$$\mathrm{W} = \frac{\sum_i \sqrt{q_i^y\, q_i^{y+1}\, p_i^{y+1}}}{\sum_i \sqrt{q_i^y\, q_i^{y+1}\, p_i^{y}}} = \sum \tilde{s}_h\, W_h^{y,\,y+1},$$

where the $W_h^{y,\,y+1}$ are the h^{th} elementary Walsh indexes and

$$\tilde{s}_h = \frac{\sum_{i\in h} \sqrt{q_i^y\, q_i^{y+1}\, p_i^{y}}}{\sum_i \sqrt{q_i^y\, q_i^{y+1}\, p_i^{y}}}$$

are proper Walsh aggegration weights. To this we compare a Laspeyres aggregation of elementary Walsh indexes ("ersatz Walsh"), $L_W^{y,\,y+1} = \sum\sum\sum s_h\, W_h^{y,\,y+1}$, where the s_h are standard base period weights.

The results are given in Table 9. We do see a perceptible difference between the actual population Walsh and the Laspeyres aggregate of elementary Walsh indexes: the latter tends to run slightly higher. However, these differences are on a par with the differences between them and the Fisher. They are minor compared to the gap between the *geomean* or Laspeyres indexes and the superlatives. This sort of result verifies that sound procedure at the lowest level is a key part of index construction.

Appendix C
The *maxminq/dutot* Combination

Why does the *maxminq/dutot* combination work so well, seeming to lead to unbiasedness for the superlative indexes?

A referee notes that *maxminq* sampling bears a strong resemblance to sampling *pps* with size variable $\sqrt{q_i^y\, q_i^{y+1}}$; for *ppswor* $(\sqrt{q^y q^{y+1}})$, the *dutot* is approximately unbiased for a Walsh target index, and so, indirectly, for any other superlative index.

Indeed, for the expectation of the numerator of the *dutot*, under this probability sampling scheme, we have

$$E_\pi\left(\sum_{i\in s} p_i^{y+1}\right) = E_\pi\left(\sum_{i'\in U} I_{i'}\, p_{i'}^{y+1}\right) = \frac{n}{\sum_{i'} \sqrt{q_{i'}^y\, q_{i'}^{y+1}}} \sum_{i'} \sqrt{q_{i'}^y\, q_{i'}^{y+1}}\; p_{i'}^{y+1},$$

where $E_\pi(\,)$ signifies expectation with respect to the sample design and $I_{i'}$ is a random indicator taking the values 1 or 0, as i' is in the sample or not. We get a similar expression for the denominator. The ratio of these two expected values is the Walsh. Therefore, apart from the usual (mild) ratio bias, which can be shown to be typically positive, the *dutot* does indeed target the Walsh, under this *pps* scheme.

We need to ask: do the two modes of sampling actually tend to have a sizeable overlap in what items get picked? For each run, for each *psu l*, *ELI c*, three items were selected either by *maxminq* or by *ppswor* $(\sqrt{q^y q^{y+1}})$ of items within *lc*. Table 13 gives the percentage of times (over 500 runs) different items make it into the sample, for some arbitrarily selected representative cases. We conclude, not entirely without surprise, that: (a) *pps* sampling leads to a wider spread of items selected, (b) the items selected by *maxminq* are a subset of those from *pps*, (c) there is a certain amount of correlation of "dominant items", that is, of those items that tend most to be selected by either method. In short, *maxminq* and *pps* $(\sqrt{q^y q^{y+1}})$ appear to be related, but loosely.

To get further insight into the relationship between the two sampling methods, we calculated bias and mean square error estimates, with respect to the Walsh population index, for the *dutot* index for each ELI, both for *maxminq* and *pps* $(\sqrt{q_y\, q_{y+1}})$ sampling. The bias and MSE estimates were based on 500 runs for each sampling method. Summary statistics were calculated across ELI's for each pair of years and each *psu*. Table 14 gives the percentage of ELI's for which the *dutot* elementary indexes are positively biased for each mode of sampling. As anticipated, *pps* sampling tends to result in positive bias; we find that *maxminq* is equally biased positive and negative.

Table 14
Percentage of ELI's for Which the *dutot* has Positive Bias for a Walsh Target, for Two Sampling Schemes

	pps$(\sqrt{q_y q_{y+1}})$			*maxminq*		
	psu 2	*psu* 3	*psu* 4	*psu* 2	*psu* 3	*psu* 4
'95–'96	75.0	86.2	75.9	64.3	61.1	61.1
'96–'97	60.7	72.4	65.5	53.6	65.5	51.8
'97–'98	65.5	75.9	78.6	41.4	27.6	42.9
'98–'99	72.4	75.9	70.4	48.3	75.9	40.8
'99–'00	89.7	72.4	75.9	48.3	20.7	44.9

Table 15 (a) gives the percent of ELI's in which the absolute bias from using *maxminq* is bigger than that from *pps* ($\sqrt{q_y\, q_{y+1}}$). In this regard, *pps* sampling is better. However, Table 15 (b) gives the percentage of ELI's in which *maxminq* yielded a larger mean square error, and here *maxminq* does better in all but two time periods/*psu*'s. We regard the mean square error criterion as the more decisive, especially given the bi-directionality of *maxminq*'s biases.

Table 15
Percentage of ELI's for Which the *dutot*'s Bias and Mean Square Error for a Walsh Target is Less for Probability Proportional to Size (Size Variable = $\sqrt{q_y q_{y+1}}$) than for *maxminq* Sampling

	(a) *Bias of pps less*			(b) *MSE of pps less*		
	psu 2	*psu* 3	*psu* 4	*psu* 2	*psu* 3	*psu* 4
'95 – '96	82.1	93.1	86.2	32.1	58.6	41.4
'96 – '97	89.2	96.6	100.0	35.7	37.9	27.6
'97 – '98	89.7	86.2	100.0	41.4	24.1	64.3
'98 – '99	89.7	82.8	92.6	41.4	37.9	40.7
'99 – '00	89.7	96.6	41.4	34.5	31.0	37.9

We conclude that the good effects of *maxminq* sampling combined with the *dutot* estimator are *not* explainable in terms of approximate mimicry of *pps* sampling. They behave differently; and overall m*axminq* seems to be somewhat *better* than *pps* ($\sqrt{q_y\, q_{y+1}}$).

We can see no alternative to explain why the *dutot* sample index should target the Walsh population index when the consistently largest sellers are sampled than that of this "brute force" mechanism: to the extent that the Walsh can be represented by a small sample of items, it is best represented by those with the consistently largest quantities, and these items are the ones the *maxminq* sampling scheme supplies.

References

Balk, B. (1999). On the use of unit values as consumer price subindices. *Proceedings of the Fourth Meeting of the International Working Group on Price Indices*, BLS, Washington, D.C.

Balk, B. (2003). Price indexes for elementary aggregates: The sampling approach. *Proceedings of the Seventh Meeting of the International Working Group on Price Indices* (*Ottawa Group*), Paris.

BLS Handbook of Methods (2005). http://stats.bls.gov/bls/descriptions.htm.

Consumer Price Indexes Technical Manual (2005). Office for National Statistics, London, http://www.statistics.gov.uk/downloads/theme_economy/CPI_Technical_Manual_2005.pdf.

De Haan, J., Opperdoes, E. and Schut, C. (1999). Item selection in the consumer price index: Cut-off versus probability sampling. *Survey Methodology*, 25, 1, 31-41.

Dalén, J. (1998). Studies on the comparability of consumer price indices. *International Statistical Review*, 66, 1, 83-113.

Diewert, E. (1997). "Commentary" [on 'Alternative Strategies for Aggregating Prices in the CPI' by M.D. Shapiro and D.W. Wilcox]. Federal Reserve Bank of St. Louis Review, 79, 3, 27-37.

Diewert, E. (2004). Index number theory: Past progress and future challenges. Presented at the SSHRC Conference on Price Index Concepts and Measurement, Vancouver, Canada, at http://www.econ.ubc.ca/diewert/concepts.pdf.

Dorfman, A.H., Leaver, S.G. and Lent, J. (1999). Some observations on price index estimators. *Proceedings of the Federal Committee on Statistical Methodology Research Conference*, *Monday B Sessions*, 56-65.

Reinsdorf, M., and Triplett, J.E. (2005). A review of reviews: Ninety years of professional thinking about the consumer price index. To appear, *Proceedings of the June 2004 NBER-CRIW Conference on Price Indexes*, Vancouver.

The Retail Prices Index Technical Manual (1998). (Ed. M. Baxter, The Stationary Office, London, at http://www.statistics.gov.uk/downloads/theme_economy/RPI_TECHNICAL_MANUAL.pdf.

Richardson, D.H. (2000). Scanner indexes for the CPI. *Proceedings of the Conference on Scanner Data and Price Indexes*, NBER, Cambridge, http://www.nber.org/books/.

Royall, R.M. (1976). Current advances in sampling theory: Implications for human observational studies. *American Journal of Epidemiology*, 104, 463-473.

Särndal, C.-E., Swensson, B. and Wretman, J. (1992). *Model assisted Survey Sampling*. Springer, New York.

Valliant, R., Dorfman, A.H. and Royall, R.M. (2000). *Finite Population Sampling and Inference*. New York: John Wiley & Sons, Inc.