Chapter 12 Errors and Bias

A. Introduction

12.1 The Consumer Price Index (CPI), like all other statistics, may be subject to general error that may occur during any stage of the estimation process but also errors that are unique to the CPI (e.g. substitution bias and quality change bias). This chapter first describes the general types of potential errors and the sources of sampling and non-sampling error that arise in estimating a population CPI from a sample of observed prices, and second reviews the arguments made in numerous studies that attribute bias to CPIs as a result of insufficiently accurate treatment of quality change, consumer substitution and other factors. It should be emphasized that many of the underlying issues discussed here are dealt with in much greater detail elsewhere in the manual.

B. Types of error

12.2 One of the main objectives of a sample survey is to compute estimates of population characteristics. Such estimates will never be exactly equal to the population characteristics. There will always be some error and the precision and accuracy of the estimate is affected by both sampling and non-sampling error. Table 12.1 gives a taxonomy of the different types of error. See also Balk and Kersten (1986) and Dalén (1995) for overviews of the various sources of stochastic and non-stochastic errors experienced in calculating a CPI. Two broad categories can be distinguished: sampling errors and non-sampling errors.

<table>
<thead>
<tr>
<th>Total error:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sampling error</td>
</tr>
<tr>
<td>Selection error</td>
</tr>
<tr>
<td>Estimation error</td>
</tr>
<tr>
<td>Non-sampling error</td>
</tr>
<tr>
<td>Observation error</td>
</tr>
<tr>
<td>Overcoverage</td>
</tr>
<tr>
<td>Response error</td>
</tr>
<tr>
<td>Processing error</td>
</tr>
<tr>
<td>Non-observation error</td>
</tr>
<tr>
<td>Undercoverage</td>
</tr>
<tr>
<td>Non-response</td>
</tr>
</tbody>
</table>

1. Sampling error

12.3 Sampling errors are due to the fact that an estimated CPI is based on samples and not on a complete enumeration of the populations involved. Sampling errors vanish if observations cover the complete population. As mentioned in previous chapters, statistical offices usually adopt a fixed weight price index as the object of estimation. A fixed weight index can be seen as a weighted average of partial indices of item groups, with weights being
expenditure shares. The estimation procedures that most national statistical offices (NSOs) apply to a CPI involve different kinds of samples. The most important kinds are:

- For each item group, a sample of items to calculate the partial price index of the item group;
- For each item, a sample of outlets to calculate the elementary price index of the item from individual price observations;
- For each item group, a sample of a day or a time span of the month when the data collection has to be carried out. Concerning this issue, the introduction of scanner data (that in general cover more than one week of a month) in CPI compilation could reduce the potential errors arising in the traditional data collection for this dimension of sampling;
- A sample of households needed for the estimation of the average expenditure shares of the item groups (some countries use data from national accounts instead of a household budget survey (HBS) to obtain the expenditure shares, as described in chapter 3).

Sampling error can be introduced at any of the stages of the sample selection process. The potential for sampling error is greater in the selection of outlets and even more so for products because there is no comprehensive frame from which to select units for sampling.

12.4 The sampling error can be split into a selection error and an estimation error. A selection error occurs when the actual selection probabilities deviate from the selection probabilities as specified in the sample design. The estimation error denotes the effect caused by using a sample based on a random selection procedure. Every new selection of a sample will result in different elements, and thus in a possibly different value of the estimator.

2. Non-sampling error

12.5 CPI surveys involve many operations, all of which are potential sources of non-sampling error. The non-sampling errors arise from the survey process, regardless of whether the data are collected from the entire universe or from a sample of the population. They can be subdivided into observation errors and non-observation errors. Observation errors are the errors made during the process of obtaining and recording the basic observations or responses. The most general categories of observation errors are overcoverage, response error, and processing error.

12.6 Overcoverage means that some elements are included in the survey which do not belong to the target population. For outlets, NSOs usually have inadequate sampling frames. For example, in some countries, a business register is used as the sampling frame for outlets, where outlets are classified according to major activity. The register thus usually exhibits extensive overcoverage, because it contains numerous outlets which are out of scope from the CPI perspective (e.g. firms that sell to businesses rather than to households). In addition, there is usually no detailed information on all the item sold by an outlet, so it is possible that a sampled outlet may turn out not to sell a particular item at all.

12.7 Response error results from the collection of incorrect, inconsistent, or incomplete data. Response error may arise because of the collection of data from inappropriate
respondents, deliberate distortion of responses, interviewer effects, misrecording of responses, pricing of wrong items, misunderstanding or misapplication of data collection procedures, or misunderstanding of the questions or survey needs and/or lack of cooperation from respondents. In price surveys, where the main price collection method is by persons who regularly visit outlets, they may return with prices of unwanted items.

12.8 **Processing error** occurs after the survey data are collected, during the processes that convert reported data to published estimates and consistent machine-readable information. Each of the processing steps, such as coding, data entry, transfer and editing (control and correction), can generate errors. For example, at the outlets, the price collectors write down the prices on paper forms or using dedicated software on a tablet. In the first case (paper-and-pencil data collection), after the collectors have returned home, a computer is used as the input and means of transmission for the price information. It is clear that this way of processing prices is susceptible to errors. The second case (computer assisted data collection) in less risky because it has built-in validation checks, but it could be susceptible to errors for other reasons such as the lack of adequate controls during the recording of the prices. Processing error also include a failure to identify true errors during regular micro- and macro-editing. Even when errors are discovered, they can be corrected improperly because of inadequate imputation and quality adjustment procedures. The occurrence of processing errors is strongly influenced by survey planning, and to some extent the survey’s resources (e.g., staff and budget, devices, training) and constraints (e.g., elapsed time between data collection and publication).

12.9 **Non-observation errors** are made when the intended measurements cannot be carried out. The most general categories of non-observation errors are undercoverage, and non-response error. **Undercoverage** occurs when elements in the target population do not appear in the sampling frame used for sample selection. The source of undercoverage error is the sampling frame itself. For instance, it is likely that there are delays in updating the outlet frame to include new in-scope units or to exclude mail order firms and non-food market stalls from their outlet sampling frame. Undercoverage means that some outlets where relevant items are purchased cannot be contacted.

12.10 **Non-response** is another category of non-observation error. Non-response errors may arise from the failure to obtain the required information in a timely manner from all the units selected in the sample. A distinction can be drawn between total and partial (or item) non-response. Total non-response occurs when selected outlets cannot be contacted or refuse to participate in the price survey. Another instance of total non-response occurs when mail or electronic questionnaires and collection forms are returned by the respondent and the price collector, respectively, after the deadline for processing has passed. Partial (item) nonresponse occurs when a responding unit does not complete an item or items on the survey questionnaire, or the responses obtained are unusable. Mail or electronic questionnaires and collection forms that are only partially filled in, scanner data with missing information concerning specific outlets or Global Trade Item Numbers (GTINs) in the sample and web-scraped prices where some information are not downloaded from the web, are examples of partial non-response. If the price changes of the non-responding outlets differ from those of the responding outlets, the quality of the price change estimates will be affected.
12.11 Another source of errors is the failure to measure the price actually paid. This failure may be caused by the use of list prices (e.g. for cars) and by the presence of discounts, coupons, bargaining etc., which are typically not accounted for. In many countries, the discounting of prices is becoming more common and the importance of discounted prices is increasing. Another source of error is due to the tendency of price collectors to choose an excessive proportion of regularly priced items in the price reference period, whereas the proportion of sales prices in the prices collected becomes proportional to their real share later in the year.

C. Measuring error

1. Estimation of variance

12.12 The variance estimator depends on both the chosen estimator of a CPI and the sampling design. The 2012 International Labor Organization (ILO) Survey of country practices gives an overview of the sampling methods that were applied in the compilation of CPIs by NSOs. It found that only one out of three use some sort of probability techniques for location selection, one in five for outlet selection, and only one in ten uses probability sampling for item selection. In the absence of probability techniques, so-called judgmental and cut-off selection methods are applied.

12.13 In view of the complexity of the sample designs in compiling a CPI (where the samples of locations, outlets and items are just partially connected), an integrated approach to variance estimation appears to be problematic. Therefore, it appears to be difficult to present a single formula for measuring the variance of a CPI, which captures all sources of sampling error. It is, however, feasible to develop partial (or conditional) measures, in which only the effect of a single source of variability is quantified. For instance, Balk and Kersten (1986) calculated the variance of a CPI resulting from the sampling variability of the HBS, conditional on the assumption that the partial price indices are known with certainty. Ideally, all the conditional sampling errors should be put together in a unifying framework in order to assess the relative importance of the various sources of error. Under rather restrictive assumptions, Balk (1989a) derived an integrated framework for the overall sampling error of a CPI.

12.14 There are various procedures for trying to estimate the sampling variance arising from a probability sampling design. For instance, assuming a cross-classified sampling design in which samples of items and outlets are drawn independently from a two-dimensional population, with probabilities proportional to size (PPS) in both dimensions, a variance formula can be derived. Where overall estimate of the sampling variance cannot be made, at the very least, a basic analysis should be conducted.

12.15 The main problem with non-probability sampling is that there is no theoretically acceptable way of knowing whether the dispersion in the sample data accurately reflects the dispersion in the population. It is then necessary to rely on approximation techniques for variance estimation. One such technique is quasi-randomization (see Särndal, Swensson and Wretman (1992, p. 574)), in which assumptions are made about the probabilities of sampling items and outlets. The problem with this method is that it is difficult to find a probability
model that adequately approximates the method actually used for outlet and item selection. Another possibility is to use a replication method, such as the method of random groups, balanced half-samples, jackknife, or bootstrap. This is a completely non-parametric class of methods to estimate sampling distributions and standard errors. Each replication method works by drawing a large number of sub-samples from the given sample. From each sub-sample, the parameter of interest can be estimated. Under rather weak conditions, it can be shown that the distribution of the resulting estimates approximates the sampling distribution of the original estimator. For more details on the replication methods see Särndal, Swensson and Wretman (1992, pp. 418–445).

2. Qualitative assessment of non-sampling errors

12.16 As estimating the quantitative impact of the non-sampling errors is more difficult, a qualitative assessment should be included. For instance, the coverage of the sampling frames as a proxy of the target populations can be described and provided (including gaps, duplications and definitional problems). The percentage of the target outlet samples from which responses or usable price data were obtained (i.e. the response rates) can be provided. Any known difference in the prices of responding outlets and non-responding outlets can be described, as an indication of the method of imputation or estimation used to compensate for non-response. Other examples of qualitative measures of non-sampling errors are indicators such as Implicit Quality Indices (IQIs), that compare indices with and without prices adjusted for quality changes. Similarly, the effects of editing can be measured by comparing the CPI estimates based on edited and non-edited data sets. As processing errors tend not to be well-reported or well-documented, they are seldom treated in the survey research literature. The occurrence of processing errors can be reduced through survey process improvements. Several categories of non-sampling errors provide the bulk of the bias issues discussed below.

3. Procedures to minimize errors

12.17 The estimation error can be controlled and minimized by means of the sampling design. For example, by increasing the sample size, or by taking selection probabilities proportional to some well-chosen auxiliary variable, the error in the estimated CPI can be reduced. The choice of an adequate sampling design for the CPI is an extremely complex matter (see Dorfman et al. (2006)). The target population is the set of all goods and services that are acquired, used or paid for by households from outlets in a particular time period. A proper probability sampling procedure selects a sample by a random mechanism in which each good or service in the population has a known probability of selection. In combination with a Horvitz-Thompson estimator, such a probability sampling design will produce an index that is (approximately) unbiased and precise.

12.18 The probability sampling designs used extensively in survey practice are simple random (SI) sampling, and PPS sampling, with or without some form of stratification (more details are provided in Chapter 4). The advantage of SI sampling is its simplicity; it gives each population element the same probability of being included in the sample. PPS sampling has the advantage that the more important elements have a larger chance of being sampled than the less important ones. For instance, in one European NSO the outlets are selected with
probabilities proportional to some proxy for size, namely their number of employees. Unequal probability designs can lead to a substantial variance reduction in comparison with equal probability designs. In stratified sampling, the population is divided into non-overlapping sub-populations called strata. For instance, other European NSO the population of outlets is split into two outlet types (multiple and independent) to form different strata by region. In each stratum, a sample is selected by PPS sampling, or SI sampling. One of the reasons why stratification is so popular is that, if strata well constructed, it results in low variance of the price changes within a stratum. Stratification is a useful strategy to make the sample more efficient.

12.19 Because appropriate sampling frames are lacking, samples are frequently obtained by non-probability methods. Judgmental (or purposive) sampling is one form of non-random selection. In this case, an expert selects certain “typical” locations and outlets where data are to be collected. With skill on the part of the expert, a fairly good sample might result. A more sophisticated non-probability method is quota sampling. In quota sampling, the population is first divided into certain strata. For each stratum, the number (quota) of locations and/or outlets to be included in the sample is fixed. Next, the price collector in the field simply fills the quotas, which means in the case of outlet sampling that the selection of the outlets is ultimately based on the judgement of the price collectors. Another non-probability method is cut-off sampling, which means that a part of the target population is deliberately excluded from the sample selection process. In particular, this procedure is used when the distribution of the value of some auxiliary variable is highly skewed. For instance, a large part of the population may consist of small outlets whose contribution to total sales is modest. A decision may then be taken to exclude from the sampling frame the outlets with the lowest sales. Because the selection is non-random, non-probability methods usually lead to more or less biased estimates. Empirical results of research presented in De Haan, Opperdoes and Schut (1997) nevertheless show that non-probability selection methods do not necessarily perform worse, in terms of the mean square error, than probability sampling techniques.

12.20 Provided that the sampling design is given, the sampling variance of an estimated (all-items) CPI can in general be lowered by:

- Enlarging the samples of items and outlets;
- The application of suitable stratifications to the various populations (e.g. grouping items with respect to similarity of price changes).

12.21 It is important to allocate optimally the available resources both between and within the different CPI samples, since badly allocated samples may lead to unnecessarily high sampling errors. Dalén and Ohlsson (1995) shows that the error resulting from item sampling is relatively high compared with the error resulting from outlet sampling. In this case, it is worthwhile increasing the sample size of items and reducing the sample size of outlets. Beisteiner (2008) stress the importance of allocating resources to those areas where the effect on the quality of the all-items CPI is maximized, especially to goods and services with a high relative expenditure weight and to goods and services with high dispersion of prices. The paper presents a “ready-to-use” formula, the Neyman formula, for the allocation of the sample, which optimizes the precision of the CPI for given resources, as discussed in chapter 4.
A systematic analysis of sampling errors offers possibilities for improving or reducing cost. The problem of optimum sample allocation is usually formulated as the determination of the sizes of the samples of items and outlets, and their distribution over the strata that minimizes the sampling error of an all-items CPI, subject to the available budget.

The accuracy of the CPI could be improved by making use of scanner data, which collect more prices for more varieties of items on more days of the month than traditional data collection methods. Bradley (1996) discusses the potential for scanner data to reduce the sampling error of the corresponding official CPI component index. The use of scanner data also has a positive effect on the time dimension of sampling, covering a time span much longer than the one covered by the traditional data collection. In chapter 10 it is argued that the scanner data should cover the whole period for which CPI is constructed, rather than a sub-period. In some cases, the use of scanner data removes the need for sampling as a census of products can be used.

As already mentioned, a business register is usually not an adequate sampling frame for outlets, because it provides extensive overcoverage. Often, they include outlets that are no longer in business or have changed the activity. It is recommended to set up an appropriate sampling frame by enumeration of the main outlets within each sampled location. Such enumeration yields a list of all outlets in a location together with the item groups that belong to their assortments. A less expensive way to organize an outlet sampling frame is to ask the price collectors – who may be assumed to know the local situation well – to make a list of outlets where purchases are made by households. It is important, that information about the quality of the sampling frame, in terms of overcoverage or undercoverage, and its completeness for the target population is known.

The populations of items (and varieties) and outlets are continually changing through time. The composition of most item groups is not constant over time, because items disappear from the market and new ones appear. The outlet population also changes over time: outlets close, temporarily or permanently; new outlets emerge; the importance of some outlets diminishes or increases. The samples of items (and varieties) and outlets should be reviewed and updated periodically to maintain their representativity with respect to the current buying habits of the households. In many countries, these are reviewed and updated with annual periodicity.

Measurement errors by price collectors can be reduced by providing them with handheld computers or tablets for data entry that have integrated validation checks. In this way the validation and editing of observed prices can be executed at the point of price collection (i.e. in the outlet), by comparing the currently observed price quote with the previously observed one (by setting a limit on the percentage price change) and with the price quotes obtained from other outlets (by setting suitable upper and lower limits). Details are provided in chapter 5, section on handheld computers and tablets. Although using pricing forms that contain information on the previous period’s price can reduce response variance, it may as well cause response bias and lag. Before introducing handheld computers and tablets, proper usability testing and training for price collectors are required to avoid them being a source of error.

It is useful to appoint data collection supervisors to conduct quality assurance checks on the data collectors. It is also a good idea to organize regular meetings where price collectors and statisticians from the head office can share their experiences. In this way, the
statisticians will keep in touch with the conditions in the field, and may take the opportunity to provide more information about frequently made price collection errors and new representative products.

12.28 It is important to check the collected price data for *processing errors* and, where possible, to correct these errors. This activity is called data editing. When editing is carried out on individual observations, it is called micro-editing. When the resources to spend on data editing must be minimized, while at the same time maintaining a high level of data quality, selective editing and macro-editing are possibilities. Selective editing is a form of traditional micro-editing, in which the number of edits is kept to a minimum. Only those edits that have an impact on the survey results are carried out. Macro-editing offers a top-down approach. The edits are carried out on aggregated data (for instance, the price index numbers of an item group) instead of individual records (for example, price observations). Micro-editing of individual records is then carried out only if macro-edits raise suspicion. In particular, attention should be paid to outliers among the observations (more information on data editing and use of algorithms is provided in chapter 5; a comprehensive description of statistical data editing procedures is given in De Waal, Pannekoek and Scholtus (2011)).

12.29 *Non-response* reduces sample size, results in increased variance, and usually introduces selection bias. Nonresponse rates, or missing observations, are often viewed as a proxy for the quality of a survey. While nonresponse rates are important, imputation rates alone provide no indication of nonresponse bias. There are three methods for the treatment of missing price observations. First, the corresponding price can be excluded from the data set of previous period prices, so that the set of previous period prices is “matched” with the set of current prices. Second, this matching can be achieved by using an imputed (or artificial) price for the missing one. The imputed price can be calculated by either carrying forward the previous price observation or by extrapolating the previous price observation using the change of other price observations for the same item. Third, there is the possibility to reweight the sample to minimize the effect of non-response error. The objective of reweighting is to inflate the weight given to the prices of the responding outlets. This compensates for those prices that are lost by non-response (for details, including advantages and disadvantages of each approach see chapter 7).

**D. Types of bias**

12.30 Bias is defined as a systematic tendency for the calculated CPI to diverge from some ideal or preferred index, resulting from the method of data collection or processing, or the index formula used. This section reviews several categories of systematic error, either in pricing or in index construction, that potentially can lead to bias in the all-items CPI. The emphasis here is on the categorization of different type of bias, along with some consideration of their likely size, but also on methods to reduce or eliminate these categories of errors. The question might arise of why such a discussion is necessary, since such issues as quality change, and the appropriate methods for handling them in the CPI, are dealt with at both a conceptual and operational level in other chapters (see chapter 6 and chapter 7 of the publication *Consumer Price Indices: Theory*).

12.31 The reason this chapter addresses the topic of CPI bias per se is the great surge in interest in price measurement problems during the mid-1990s. Especially in the United States
(US), the view became widespread that the CPI was subject to systematic upward biases because of the failure to deal adequately with consumer substitution, product quality improvements, and the introduction of new goods and services. Moreover, it was recognized, first, that the existence of such upward bias would have fundamental implications for the measurement of recent trends in output and productivity, and second, that the elimination of upward bias could substantially improve the government budget situation through reduced government expenditures and increased tax revenues (see, for example, Eldridge (1999) and Duggan and Gillingham (1999)). These findings led to a series of papers and reports on CPI measurement problems, often accompanied by point estimates of aggregate bias.

12.32 One of the most prominent examples of these quantitative studies of bias is that by the Advisory Commission to Study the CPI (US Senate, 1996). Responses and estimates by statistical agencies include those provided by Abraham et al. (1998), US Bureau of Labor Statistics (1998), Ducharme (1997), Edwards (1997), Fenwick (1997), Johnson et al. (2006), Lequiller (1997), Moulton (1996b), and Moulton and Moses (1997). Research undertaken has shown that it is difficult to both quantify and assess the direction of potential bias, and that the extent, the direction and even the existence of bias remain something that will depend upon the specific circumstances of each set of CPI estimates and cannot always be determined with certainty.

12.33 Two points are worth making at the outset with respect to measuring bias in CPIs. First, the issue has usually been addressed in the context of the cost of living index (COLI) (for the concept of COGI see chapter 1). That is, the CPI bias has been defined as the difference between the rate of increase in the CPI and the rate of increase in a true COLI. Many discussions on bias have taken as given that the COLI should be the CPI’s measurement objective. Somewhat different conclusions might be reached if the index objective was taken to be a fixed-basket price index. Notably, the gains in consumer welfare from a widening array of new goods, or the ability of consumers to substitute away from items with increasing relative prices, might be deemed irrelevant and an index that ignored those factors might not be judged biased on that account.

12.34 The second point is that CPI bias is not amenable to estimation with the same level of rigour as that used in CPI variance estimation. Since the COLI or other ideal target index is unobserved, analysts have been forced to rely in part on conjectures and on generalizations from fragmentary empirical evidence in order to quantify the extent of bias. The notable exceptions are with respect to substitution bias, when traditional Laspeyres indices and indices using superlative formulae can be computed using the same underlying price and expenditure data, and the differences construed as a measure of the upward bias from use of the Laspeyres formula.

12.35 Several different taxonomies of bias have appeared in the literature mentioned above. It is sufficient, however, to employ four categories roughly corresponding to those set forth in the best-known study, namely the Final report of the Advisory Commission to Study the CPI

(the Boskin Commission), established by the US Senate Finance Committee in 1995. These categories are: upper-level substitution bias; elementary aggregate bias; quality change and new goods bias; and new outlet bias.

12.36 These categories can be further broken down into two subgroups according to whether they refer to errors in individual price measurements or errors in computing index series. Quality change bias and new goods bias arise because of failures to measure adequately the value to consumers of individual goods and services that appear in (or disappear from) the marketplace. It should be recognized that discussions of “new goods” problems apply equally to all products, whether goods or services. At a conceptual level, it can be difficult to distinguish these two biases from each other. Operationally, however, quality change bias pertains to the procedures for comparing new products or models with the older products they replace in the CPI samples. In general, new goods bias can be thought of as applying to entirely new types of products, or products that would not enter samples routinely through forced replacement. New outlet bias, sometimes referred to as outlet substitution bias, is similar to new goods bias but is focused on the appearance of new types of stores or marketing methods that offer goods at lower prices or higher quality.

12.37 The other categories of bias refer to the procedures for constructing index values from component series. As noted throughout this manual, CPI construction can be thought of as taking place in two steps, or at two levels. At the lower level, individual price quotations are combined; at the upper level, these basic indices are aggregated together. Corresponding to these two levels are two forms of potential bias. Elementary aggregate bias involves the averaging formulae used to combine price quotations into basic indices. Upper-level substitution bias applies to the formulae used to combine those elementary aggregates into higher-level indices. These components of potential bias, and the means used to measure them, are discussed in more detail below.

1. Components of bias

Upper-level substitution bias

12.38 Upper-level substitution bias is perhaps the most widely accepted source of CPI bias, and the kind with which economists are most familiar from textbook expositions of price index theory and practice. Simply stated, it arises when CPIs employ the Laspeyres formula (see chapter 3 of the publication Consumer Price Index Theory) or a similar method that uses a fixed base like the Lowe and Young formulas, which is well known to provide an upper bound on a COLI under certain assumptions about consumer behaviour. The assumption in the definition of the Laspeyres-type price index is that substitution among goods is zero, which is against one of the cornerstones in the theory of demand. As noted in paragraph 12.38 above, quantitative measures of upper-level substitution bias can be generated by comparing Laspeyres-type price indices to Fisher ideal, Tornqvist or other superlative indices. Under certain assumptions about, for example, constant preferences, these will stand as relatively precise bias estimates.

12.39 Genereux (1983) and Aizcorbe and Jackman (1993) provide such index comparisons and estimates of upper-level substitution bias using actual CPI index series for Canada and
the US, respectively. Other early studies by Braithwait (1980) and Manser and McDonald (1988) estimate the substitution bias in US national account indices. In lieu of superlative indices, the Braithwait study uses estimated exact COLIs based on demand system estimation. A similar estimate for the Netherlands is provided by Balk (1990). In these studies, and in the more recent analyses of US CPI data by Shapiro and Wilcox (1997a) and Cage et. al. (2003), the existence of an upward bias from the Laspeyres formula is demonstrated consistently. The biases in the annual index changes in individual years are relatively small, typically 0.3 percentage points or less, and depend empirically on such factors as the distance from the Laspeyres base period, the level of index detail at which the alternative formulae are applied, and whether the superlative index is of the fixed base or chained variety.

12.40 The major differences between Laspeyres and superlative indices derive from the variation in relative prices over the period being compared, and from the shift in quantities consumed towards those index categories that have fallen in relative price. This leads to several conclusions:

- If index movements are characterized by continuing, uniform drift in relative prices over time, with accompanying drifts in consumption, the size of the annual Laspeyres bias will tend to increase with the distance from the base period. The estimates of the upper level substitution bias presented in ABS (2017) show that average annual substitution bias is 0.11 one year after a re-weight of the CPI, increasing to 0.20 in the sixth year. Greenlees (1997) notes that there is little evidence for this phenomenon in the US; see also Szulc (1983).

- Under the same circumstances, reducing the expenditure weight chaining interval will work to reduce the upper-level substitution bias in the Laspeyres-type CPI. The more frequent chaining will increase the weight given to indices that are falling in relative price, thereby reducing the rate of CPI growth. Conversely, if there is “bouncing” in relative index movements, frequent chaining can lead to an upward “chain drift” in a Laspeyres index.

- Upper-level substitution bias will tend to be larger during periods of higher inflation, if these periods also have greater relative price variation. Little empirical evidence exists on this point, however.

12.41 The concept of upper-level substitution bias has been derived and discussed in the context of COLI theory, but an equivalent bias may be defined from the perspective of the fixed-basket price index. If the Fisher ideal or other superlative index is judged preferable on the basis of its symmetric treatment of base period and current period expenditure patterns, then the difference between that index and a Laspeyres could be interpreted as a measure of representativity bias. A similar argument could be applied with respect to lower-level substitution bias within elementary index cells.

12.42 Lebow and Rudd (2003) have defined and estimated another category of bias related to upper-level aggregation. They concluded that the HBS weights used in the US CPI were subject to error because of, for example, underreporting of alcohol and tobacco expenditures. This will lead to a weighting bias if the errors in relative weight are correlated with component index changes (sources for, and problems in, expenditure weight estimation are discussed in detail in chapter 3).
Elementary aggregate bias

12.43 Elementary aggregate bias arises from the use of an inappropriate method for aggregating price quotations at the very lowest level of aggregation. An elementary index in the CPI is biased if its expectation differs from its measurement objective. It can take two forms; formula bias or lower-level substitution bias. The index suffers from formula bias if, as a result of the properties of the formula, the result produced is biased relative to what would have been the result if a price change of fixed-basket could have been estimated. The index suffers from lower level substitution bias if it does not reflect consumer substitution among the items contained in that elementary aggregate. Lower level substitution error is only relevant where the measurement objective is a COLI. Thus, given any elementary index formula, the two forms of bias can be distinguished according to the objective of the elementary index.

12.44 Chapter 9 of this manual and chapter 5 of the publication Consumer Price Index Theory discuss the characteristics and provides detail, an illustration, and the relative merits of the use of different elementary index number formulae. A key finding is that the Carli formula or the arithmetic average of price ratios is unsuitable for a CPI because it is liable to lead to substantial drift in the results, especially when used in its chained form. Therefore, the recommendation that Carli formula should not be used, especially in its chained form. The problems with elementary aggregate bias and the methods chosen to address them are discussed, for example, by Reinsdorf (1998), Reinsdorf and Moulton (1997) and Moulton (1996b).

12.45 The ratio of arithmetic averages (Dutot) and geometric mean (Jevons) formulae eliminate formula bias as defined here. Their expectations differ, however, when item prices do not change at a uniform rate. The differences provide one way of evaluating the potential importance of lower-level substitution bias. The geometric mean formula is exact for a COLI if consumers follow the Cobb-Douglas behavioural model (assuming that consumers adjust the relative quantities they consume inversely in proportion to the changes in relative prices so that expenditure shares remain constant), whereas the formula based on the ratio of arithmetic averages corresponds to zero-substitution behaviour. Thus, if the goal is to approximate a COLI, the geometric mean formula is judged preferable.

12.46 Scanner data provide new opportunities for measuring and addressing elementary aggregate bias. The availability of both prices and quantity information in scanner data remove the need for unweighted index formula, at least for those items where unit values are available, and it allows the calculation of elementary CPI indices by employing superlative formulae. By using scanner data Gabor and Vermeulen (2015) compute product category level elementary price indices using nine different index formulas (Carli, Dutot, Jevons, Laspeyres, Paasche, Fisher Ideal, Lowe, Geometric Lowe and expenditure weighted Jevons) and compare the resulting indices with the Fisher index. The main findings are that across product groups mean levels of annual elementary index bias vary between -0.53 percentage points and 0.55 percentage points depending on the index.

12.47 Haan and Heymerik (2009) have identified a problem associated with the use of scanner data, especially when bilateral superlative price indices are used. The high-frequency chaining, used to handle the problem of high attrition rate of items, can create drift in the index series when prices and quantities change or bounce arising from sales. Therefore, new
methods have been developed for price measurement based on scanner data. The approach proposed in Ivancic, Diewert and Fox (2009) that provides drift free, superlative-type indices through adapting multilateral index number theory seems to provide a solution to this problem. The methods proposed, however, pose some practical challenges and require some more evidence before becoming widely accepted. For overview of methods for price measurement using scanner data see Chessa Verburg and Willenborg (2017) and chapter 10 on scanner data.

12.48 The method used by the NSO for sampling items within a category will determine the effectiveness of formula choice in dealing with lower-level substitution bias. For example, if only a single representative item is chosen to represent the category, the index formula will fail to reflect the consumer response to any relative price change in the universe of items. Therefore, a larger sample of representative items should yield a smaller sampling variance for a given elementary index. More generally, the geometric mean formula index suffers from an upward bias in very small samples (less than 5 observations), so lower-level substitution bias may be underestimated in empirical comparisons of the geometric mean to other index formulae. White (1999) discusses the relationship between sampling error and bias estimates. McClelland and Reinsdorf (1999) also study the impact of small sample sizes on the index and conclude that it has the effect of rising the expected values of an index based on non-linear formulas, especially the geometric mean formulae. More extensive use of scanner data may mitigate the problem of small sample all together given that sample sizes in a typical scanner data set are large. In some cases, the use of scanner data may remove the need for sampling.

12.49 The impact of formula choice can be estimated with some degree of precision over a given historical period. Any corresponding bias, however, can be estimated only by assuming that the geometric mean or other functional form successfully approximates the index’s measurement objective.

12.50 As implied by the above discussion, the importance of elementary aggregate bias will vary by country, depending on the particular index formulae used, the degree of heterogeneity within index strata, and the sampling methods employed. Also, as with upper-level substitution bias, elementary aggregate bias will vary with the overall level of inflation in the economy if absolute and relative price changes are correlated.

12.51 The performance of any formula for elementary aggregate calculation will also be affected by the methods used by the NSO to handle special situations, such as seasonal goods and other products that are temporarily unavailable. Armknecht and Maitland-Smith (1999) discuss how the failure to impute missing prices can lead to bias in the modified Laspeyres and other index formulae.

Quality change and new goods bias

12.52 Discussion of potential CPI biases arising from inadequate quality adjustment has a long history. For example, the Stigler Committee report on US price statistics (Price Statistics Review Committee, 1961) indicated that “if a poll were taken of professional economists and statisticians, in all probability they would designate (and by a wide majority) the failure of the price indices to take full account of quality changes as the most important defect of these
indices”. In most studies of bias, unmeasured or mismeasured quality change is also the
largest contributor to the total estimated bias. Just as quality adjustment is widely recognized
as an extremely difficult process, however, it is correspondingly difficult to measure any
quality change bias.

12.53 Unlike substitution bias, which can be estimated by comparison of alternative
formulae, quality change bias must be analyzed on a product-by-product basis. Products and
their associated index components will experience widely varying rates of quality change
over time. Moreover, the methods used for quality adjustment will also vary. Whereas the
linking method may dominate in terms of frequency of use, important index components may
employ production cost, hedonic adjustment, or the other methods described in chapter 7 of
this manual and chapter 7 of the publication Consumer Price Index Theory.

12.54 A crucial point to recognize is that the direction of overall quality change does not
imply the direction of any quality change bias. Non-experts sometimes assume that little or
no quality adjustment is carried out in the CPI, and that it therefore must overestimate price
change in view of the many demonstrable improvements over time in the quality of goods
and services. Rather, for any component index, the issue is whether the direct or indirect
method chosen for quality adjustment overestimates or underestimates the relative quality of
replacement items in the CPI sample. The resulting bias can be either positive or negative.

12.55 Empirical evidence on quality change bias has been based largely on extrapolation
from individual studies of particular products. These individual studies may involve, for
example, comparisons of hedonic regression indices to the corresponding CPI series or
estimates of the value of some product improvement that is ignored in CPI calculations.
Although the majority of such studies have suggested upward rather than downward bias, the
reliance on fragmentary evidence has led to criticism by observers who point to evidence of
quality declines that have not been subjected to systematic analysis.

12.56 Overall quality trends can also be a matter of subjective valuation, especially for
services. New technology has led to unambiguous improvements in the quality of many
consumer durables and other goods. By contrast, in service sectors such as mail delivery,
public transport and medical care, it can be difficult to evaluate changes in quality. Airline
travel, for example, has become safer and faster but perhaps less comfortable and reliable in
recent decades, and the lack of cross-sectional variation in these characteristics makes the use
of hedonic quality adjustment problematic.

12.57 Digitalization of economy, if not properly captured could also be a source of bias.
Reinsdorf and Schreyer (2017) identify three possible sources of distortion that digital
economy can cause, one of which “incomplete adjustment for quality change, i.e., the
treatment of new, and typically improved, varieties of existing digital products; the treatment
of new digital products that replace existing non-digital products; and improved variety
selection of digital and non-digital products” yield overestimation of inflation. By using
weights in an average consumption basket for OECD countries, from the OECD’s purchasing
power parities program, the inflation was overestimated by 0.28 percentage points due to
possible under adjustment for quality changes in digital products such as computers, ICT
equipment and telecommunication services.

12.58 New goods bias, like elementary aggregate bias, can be divided conceptually into two
components. The first concerns the failure to bring new products into the CPI sample with
sufficient speed. This can lead to upward bias if those new products later experience large price reductions that are not reflected in the index. The second component is the welfare gain that consumers experience when a new product appears. This may not be viewed as a bias, however, when the COLI is not accepted as the CPI’s measurement objective.

12.59 As discussed in chapter 7, “new goods” can be: products that replace predecessor items, for example cloud storage areas replacing physical storage devices; new varieties of an existing product that widen the range of consumer choice, such as non-alcoholic beers and ethnic restaurants; or products that represent entirely new categories of consumption, such as multi task robots for cooking or smartphones.

12.60 Like quality change bias, new goods bias has sometimes been estimated primarily by generalization from individual product evidence. A frequent approach has been to measure the price change for a product or category during a period prior to its entry into the CPI sample. Studies by Hausman (1997, 1999) of breakfast cereals and mobile phones provided quantitative measures of the consumer surplus gain from the new products, but this complex econometric approach has not been applied widely. Some of the Boskin Commission’s estimates of new product bias, notably those for food, were necessarily based on conjecture.

12.61 Also, like quality change bias, new goods bias could be negative if the range of products decreases, if valuable consumer goods disappear from the market, or if the index fails to capture phases of rapid price increase for items. Most observers, however, seem to agree on the direction of bias as upward, and that the uncertainty concerns the magnitude. The extent of the new product bias depends on the importance of the new products in terms of the proportion of consumer expenditure spent on new products not yet introduced in the CPI basket, and the extent of the price decline from the initial price.

12.62 One of the risks of downward bias in CPI is associated with producers that reduce the package size of household goods keeping the price stable (“shrinkflation”) or repacking the old product. This phenomenon is strictly linked with the minor changes in product packaging or product characteristics (so called “product relaunches”). These should be properly handled, through the use of a unit value approach, in particular in scanner data, because the product relaunched in most of the cases presents a new GTIN but it is directly comparable with the product before the relaunch. When scanner data are used in CPI and big amount of data are processed on a weekly basis, it is not possible to observe and report all changes incurred in the size or in the characteristics of the product and assess the comparability of the previous and the replacing product. Therefore, automatic procedures have to be carefully implemented to link different GTINs in contiguous months and correctly estimate the price change, avoiding the risk of bias that usually is downward.

New outlet bias

12.63 Conceptually, new outlet bias is identical to new goods bias. It arises because of the failure to reflect either price changes in new outlets not yet sampled, or the welfare gain to consumers when the new outlets appear. The explanation for its existence as a separate bias category is twofold. The first reason is historical: new outlet bias was identified by Reinsdorf (1993) as a potentially major explanation for anomalous movements in the US CPI. Second,
the methods used to sample and compare outlets differ from those used with products, and the problems in controlling new outlet bias are somewhat different.

12.64 A failure to maintain a current outlet sample can introduce bias because the new outlets are distinctive in their pricing or service policy. Reinsdorf (1993), and more recently Hausman and Leibtag (2004, 2005), focus on the growth of discount stores. It should be noted, however, that the problem could also be geographical in nature; it is important to employ outlet sampling frames that reflect new as well as traditional shopping locations, although the wide spread and increasing weight of online outlets have changed the dimension of this issue.

12.65 One way that new products enter the CPI sample is through forced replacement, when exiting or less successful products disappear from shelves. Outlet disappearance is less frequent, and agency procedures may not provide for automatic replacement. Moreover, when a new outlet enters the sample there are no standard procedures for comparing data at the new and old outlets. Thus, the index will not incorporate any effects of, for example, lower price or inferior service quality at the new outlet.

12.66 Reinsdorf (1993) estimated the degree of new outlet bias by comparing average prices at outlets entering and disappearing from US CPI samples. There has been little or no empirical work, however, on the measurement or consumer valuation of outlet quality such as product variety, location, car parking, and customer services. As a consequence, there is little evidence on which to evaluate the accuracy of new outlet bias estimates.

12.67 Greenlees and McClelland (2011) confirms the potential importance of new outlets bias in the CPI. This study provides new and detailed evidence on the impact of the appearance and growth of new types of product outlets on the CPI. Using BLS-collected price data for 2002-2007, the authors observed a continuous increase in the market share of discount department stores and warehouse club stores, and significantly lower prices at discount department stores and at warehouse club stores than at large grocery stores. The increasing shares of lower-priced store categories reduced the average prices collected by the BLS. Changes in the distribution of outlets within categories also led to a substantial decline in average prices. Hausman (2004) also documents the growing role of discount outlets and provides a specific example of outlet bias.

12.68 Like with new goods bias, most studies seem to agree on the direction of bias as upward. The extent of the new outlet bias depends on (i) the components of the CPI basket that are likely to be affected, (ii) the change in market share of new outlets for these items, and (iii) the percentage difference in quality-adjusted prices between new outlets and old outlets. Estimates of the size of outlet-substitution bias must take into account the fact that the market price of an item depends on both the quality of the item and the quality of the retail outlet where it is purchased, based on such factors as the level of service and the convenience of the location.

2. Summary of bias estimates

12.69 The 1996 Boskin Commission report gave a range of estimates for the total upward US CPI bias of 0.8 to 1.6 percentage points, with the point estimate being 1.1 percentage point. This total reflects the straightforward summation of the component bias estimates. As
reported in US General Accounting Office (2000), however, changes in CPI methods subsequent to 1996 led the Boskin Commission members to reduce their estimates of total bias. Lacking evidence to the contrary, additivity of biases has been assumed in most such studies. Shapiro and Wilcox (1997b) provide probability distributions and correlations of their component bias estimates, yielding an overall confidence interval for the total bias. Most detailed studies of bias also conclude that the CPI bias is in an upward direction, although there have been numerous criticisms of that conclusion. For example, Brown and Stockburger (2006) estimated that the hedonic quality adjustment methods in apparel have had both upward and downward impacts at different points in time and for different categories of clothing in the US.

12.70 It is apparent that NSOs cannot compute or publish CPI bias estimates on a regular basis. Many of the same obstacles that prevent the elimination of bias also stand in the way of estimating bias. These include the lack of complete data on product-level consumer preferences and spending behaviour, and the inability to observe and value all differences in quality among items in the marketplace. Without such information it is impossible to calculate a true cost of living index, and similarly impossible to measure the divergence between its rate of growth and the growth rate of the CPI.

12.71 NSOs have been reluctant to provide their own estimates of CPI bias. In some cases, they have accepted the existence of substitution bias, recognizing that the use of a Laspeyres formula implies that the CPI usually will overstate price change relative to a COLI estimated by a superlative index such as the Fisher. NSOs have, however, been reluctant to draw even qualitative conclusions from the fragmentary and speculative evidence on quality change, new products and new outlet bias.

12.72 The CPI bias may appear to a different extent in different countries. Hanousek-File (2001) show that bias was especially high in countries in transition, and during the period of high inflation. They argue that the substitution bias increases with the increase of the variance of relative price changes. They argue that bias arising from substitution increases in line with the level of inflation, because as the rate of inflation increases, so does the variance of relative price changes.

3. **Procedures to minimize bias**

12.73 Although it is almost impossible to eliminate sources of bias, some measures can be taken to minimize them. These include:

(i) Use appropriate formulas in compiling elementary aggregate indices, in particular use of the geometric mean (Jevons) formula where appropriate or the ratio of arithmetic prices (Dutot) formula.

(ii) Review and update weights and CPI baskets frequently, but at least once every five years. Given that significant part of the total measurement bias in the CPI may be caused by the fixed nature of the CPI basket, the item-substitution bias and some of the new products bias could be reduced by increasing the frequency at which weights are updated. For some categories, it may be necessary to update the weights more frequently as such weights are likely to become out of date more quickly than higher-
level weights. In periods of high inflation, the weights should be updated even more frequently. Scanner data may help in this, at least for in some areas such as food.

(iii) Use a superlative index formula rather than the Laspeyres, if current-period weighting data can be obtained on time. Where Laspeyres-type formula is used, the upper level substitution bias can be reduced by more frequent updating of expenditure weights, implementing them with minimal time lag. Other options might be to use formulas that allow substitution or assumptions about substitution between item groupings to be entered.

(iv) Closely monitor and update price samples to reflect changes in the outlets from which households purchase. For example, there is clearly a need to plan for the inclusion in CPIs of purchases from outlets operating exclusively online but also from discount outlets, factory outlets, etc. whose importance has been increasing.

(v) Include new goods into the CPI as soon as possible. For a fixed-weighted index such as Laspeyres, there would also be a need to update the fixed weights to allow for the inclusion of the new goods if they are substituting for all goods in general, or to adjust the weights within an item grouping if the new good is substituting for specific items. For example, one could argue that CDs were a new good, but as they were substituting for records and tapes, they could be introduced into the item grouping for records and tapes, and weights between these items adjusted accordingly.

(vi) Make sure the most appropriate quality adjustment methods are applied. Whenever suitable and where price-determining characteristics are quantifiable, make greater use of the hedonic technique to adjust for quality change.

(vii) Make greater use of the scanner data to deal with quality change, substitution and new products. They contain detailed and timely information on the prices and quantities of all consumer transactions. The role for scanner data cannot be understated, given its ability to track market trends and detect the emergence of new products on the market, helping to reduce the lag in introducing new goods into the CPI basket. The use of scanner data also makes it possible to compile superlative price indices at detailed aggregation levels since prices and quantities are available.

E. Conclusion

12.74 The CPI is subject to various types of errors and biases that affect precision and accuracy of the CPI estimates. Several potential sources of errors and bias have been identified in the CPI and addressed, though the debate continues over to what extent and in what direction bias may still exist and the ways in which its accuracy can continue to be increased.

12.75 In order to ensure public confidence in a CPI, a detailed and up-to-date description of the methods and data sources should be published. The document should include, among other things, the objectives and scope of the index, details of the weights, and last but not least, a discussion of the accuracy of the index. A description of the sources and magnitude of the sampling and non-sampling errors (coverage, non-response rates, etc.) in a CPI provides users with valuable information on the limitations that might apply to their uses of the index. Examples of handbook of CPI methods are that published by the US Bureau of Labor
Statistics (2015), and Australian Bureau of Statistics (2009), which devote a section to the varieties and sources of possible error in the index.

12.76 Improving precision and accuracy of the CPI will take both time and resources. Resources should be allocated to those areas where the effect on the quality of the all-items CPI is maximized, especially to goods and services with a high relative expenditure weight and to those with high dispersion of prices. Further uses of scanner data can help NSOs deal with the quality changes, outlet substitution and new goods bias problems. An investigation into factors affecting consumer choice and an expanded HBS would help identify consumer preferences for different outlet types and to improve accurate price measurement in areas where quality change is rapid. Opportunities presented by technology, such as the use of computer-assisted techniques, scanning and web scraping techniques, can minimize processing errors.