Chapter 2

Methodology and Expert Judgement in Evidence-Based Policy: Lessons From the CPI Controversy

'How can that happen? Aren't paintings examined by experts?'

'Of course they are. But famous paintings come with a pedigree, a history, a string of learned opinions and endorsements, rather like precedent in law. When a painting has been accepted as genuine for a number of years, that's a very powerful recommendation. Experts are only human; they believe experts. If they're not expecting to see a fake—and if the fake's good enough—there is a better than even chance they won't spot it. Under normal circumstances, I'd have said Denoyer's Cézanne was genuine, because it's so beautifully done. But thanks to you, dear boy, I had my eye skinned for a fake.' Cyrus paused.

'And a fake is what I saw.'

(From Peter Mayle's *Chasing Cézanne*)

I

It is rare that a specialist controversy in economics carries over to the broader public. Measuring prices seems to be an exception to this general rule. Time and again, when details of the measurement procedure associated with the U.S. Consumer Price Index (CPI) have been debated among government statisticians and economists and academic economists, the public has showed a great interest. The most recent controversy was fuelled by the Final Report to the Senate Finance Committee, prepared by the Advisory Commission to Study the Consumer Price Index or Boskin Commission (after their chair, Stanford economist Michael Boskin), which announced that the CPI overstates consumer price inflation by 0.8-1.6 percentage

points annually. Following the publication of the report at the end of 1996, newspaper and television reports were filled with late-breaking stories on the CPI.

The outburst of public interest in academic economics was understandable and well warranted: measuring inflation is of immense significance for the wider public, both present and future generations. Since tax brackets and public spending programmes as well as many private contracts are inflation-indexed, the measured inflation rate affects disposable incomes for most employees and recipients of government benefits and pensions, government receipts, expenditures and national debt. To cite just one figure, the Congressional Budget Office estimated at the time of the report's publication that if the CPI were reduced by 1.1 percentage point per year (the Boskin commission's best estimate of the bias) the federal debt would be reduced by $691 billion in a decade.

Measuring prices is also of unrivalled importance for economic *analysis*. Many, if not most, empirical analyses of economic phenomena concern what economists call *real* variables, that is, variables deflated by an inflation index. Finding that our index seriously misrepresents the aspect of the phenomenon we are interested in potentially invalidates much of what we know about the economy, not only in terms of describing different economies and comparing them but also in terms of understanding the causal relationships responsible for the phenomena of interest. If we find a significant bias in the measured inflation rate many of our ideas and results regarding economic quantities such as income, growth, productivity, money and their interrelationships may have to be reconsidered (for instance, for the effect of a CPI-adjustment on what we know about consumption, see Triplett 1997). A transformation of that empirical knowledge, in turn, may have further policy consequences.

It is therefore indispensable that a hypothesis such as 'consumer price inflation in the urban areas of the United States 12 months to July 2006 was 4.1 per cent' or 'the U.S. CPI overstates consumer price inflation by 1.1 per cent annually' be based on good evidence. The aim of this Chapter is to provide a discussion of what it means that a hypothesis regarding the measurement of an aggregate economic variable is evidence-based in the context of this case. After a brief introduction to CPI measurement in section II and a summary and explanation of the Boskin recommendations in section III, I will criticise many of the recommendations made as based on less than valid evidence in the sense of Chapter 1. As we will see in section III, expert judgement played an influential role in the establishing the claim about the CPI's bias. This naturally raises the question whether and to what extent expert judgement can and should play a role in evidence-based economics. I will try to answer that question in section V. Section VI concludes.

II

Prior to some recent changes introduced in response to the Boskin Commission report, the CPI used to be a *price index*. A price index tracks the cost of purchasing a fixed basket of goods through time. If q_{i0} denotes the quantity good i in the base-period basket at time 0 and p_{i0}, p_{i1} denote its base and current-period price at times 0 and 1, respectively, then $L_{01} = \sum q_{i0}p_{i1} / \sum q_{i0}p_{i0}$ is the price index for the current period relative to the base period or *Laspeyres price index*.

In the U.S., the CPI is measured by the Bureau of Labor Statistics (BLS), which collects monthly price quotations on 71,000 individual goods and services at about 22,000 retail units in 88 regions throughout the country known as primary sampling

units. In addition to that, the BLS collects information from about 40,000 tenants or landlords and 20,000 homeowners for the housing components of the CPI. These individual goods and services are aggregated in two steps. In the first step, the individual prices are aggregated into 9,108 strata, one for each of 207 *items* in 44 *areas*. An item stratum is a low-level index for groups and services such as 'Men's suits, sports coats and outerwear' and 'Gardening and lawncare services'. Within each item stratum entry level items (ELIs) are defined. Some strata may have only one ELI (*e.g.*, 'Apples') while others can have a number of more or less heterogeneous ELIs (*e.g.*, 'Physicians' services'). An area is either an actual geographical region (32 areas correspond to locations in 29 cities) or a construct that represents smaller and mid-sized cities in several regions in the U.S. (that is, the remaining 12 areas are constructed from 56 primary sampling units).

Since a price index measures the cost of purchasing a *fixed basket of goods*, decisions have to be made how to treat cases where characteristics of goods change in one way or another. Here are some examples of relevant changes: some goods grow in their relative importance for consumers whereas others decline; the production of some goods discontinues whereas new goods are launched; the quality of goods improves or deteriorates; distribution channels change and new channels emerge; environmental factors of relevance for the consumer change. Already before the Boskin-induced alterations in the CPI-measurement procedure, the BLS employed an array of methods to deal with a changing world. Although the CPI was nominally a fixed-basket index, it was BLS practice to update its weights every decade or so. The weights themselves are obtained from the so-called Consumer Expenditure Survey (CES), which collects detailed information on all out-of-pocket expenditures from a national sample of households. Thus one could simultaneously tackle the substitution as well as the new goods problem, though admittedly crudely.

The weights obtained from the CES are used to aggregate higher-level price indices from indices for item strata. Below the level of the stratum, the weights used in the aggregation of individual prices stem from two sources. First, the Census Bureau conducts a so-called Point-of-Purchase Survey (POPS) which attempts to measure the distribution of expenditures across different retail outlets. Based on the results of the POPS the BLS selects a sample of outlets in a given area. The probability of being sampled is proportional to the share of that outlet in total expenditures in the area for the item selected. Second, BLS economists visit these stores and choose one or more specific items from which the broader category of items is to be priced. The probability to be selected for any given item is proportional to its estimated share in the outlet's revenue.

In this process, about 20 per cent of all items undergo sample rotation every year, that is, they are replaced by other, not directly comparable items. This means that full rotation takes place about every five years. For those items rotated into the sample, the BLS performs various quality-adjustment procedures.

Sample rotation is not the only source of item substitution. The BLS try to make sure that each month exactly the same kind of item is repriced. This is, however, not always possible because an item may be temporarily or permanently unavailable. In this case the BLS representative can judge the new item to be comparable to the old one, following certain standardised guidelines. If the item is judged to be non-comparable, one of a number of adjustment procedures will be applied. Quality changes are the subject of the following Chapter, for a more detailed discussion of these procedures, skip forward to section II of that Chapter.

This, in a nutshell, is the practice scrutinised and criticised by the Boskin Commission.

III

The Boskin Commission was an expert committee formed by the economists Michael Boskin of Stanford University, Ellen Dulberger of IBM, Robert Gordon of Northwestern University, Dale Jorgensen and the late Zvi Griliches, both of Harvard University. The experts made over a dozen specific recommendations regarding the measurement of the CPI, including the following (Boskin *et al.* 1996, section VIII):

- (Boskin Recommendation #1) The BLS should establish a cost of living index as its objective in measuring consumer prices.

- (from #3) A 'superlative' index formula should be adopted to account for changing market baskets, abandoning the pretence of sustaining the Laspeyres formula (specifically, the recommendation is to move to a 'trailing Tornqvist' index, *i.e.*, a regularly updated, weighted geometric mean of price relatives, at the stratum and ELI level, and to geometric means of price relatives at the elementary aggregation level)

- (#9) The BLS needs a more permanent mechanism for bringing outside information, expertise, and research results to it.

- (from #12) The BLS should develop a number of new data collection initiatives to make progress regarding the impact of new commodities and services and the

changing economic, social, and environmental climate within which the consumer is operating.

- (#15) Congress should establish a permanent (rotating) independent committee or commission of experts to review progress in this area every three years or so and advise it on the appropriate interpretation of the then current statistics.

- (#16) Congress and the President must decide whether they wish to continue the widespread overindexing of various federal spending programs and features of the tax code. If the purpose of indexing is fully and accurately to insulate the groups receiving transfer payments and paying taxes, no more and no less, they should pass legislation adjusting indexing provisions accordingly.

Not all of these recommendations are easily understandable to outsiders. In this section, I explain each recommendation in slightly more detail; the next section provides an analysis and criticism.

The most salient and important recommendation is that the BLS should establish a cost-of-living index (COLI) framework for the CPI. That the CPI should be a COLI is widely accepted among (U.S.) economists (though there are dissenters, see, *e.g.*, Deaton 1998), and indeed it was already recommended by the Stigler Committee when it studied the accuracy of the CPI in the late 1950s. In contrast to a price index, which measures the change in the cost of purchasing a fixed basket of goods, a COLI measures the cost of purchasing a given amount of 'utility' or 'welfare' or 'standard of living' (in discussions about the CPI, these terms are used interchangeably). Introducing this abstract quantity solves a number of conceptual problems simultaneously because it makes all goods commensurable, at least in principle.

Simply put, according to this framework, purchasing a good means to purchase a quantity of utility, and consumers know how different goods compare in terms of the amount of utility they are worth. Hence, there is a principle solution to all problems caused by the fact that the universe of goods traded in the market continually changes.

If prices change, consumers will react to that by shifting their purchases from now relatively more expensive to relatively cheaper goods—thereby maintaining their standard of living. Moreover, when aspects of the goods themselves change, old good and new good are now commensurable. If, say, an old fashioned TV set used to provide 1,000 units of utility on average and cost $500, and a new flat LCD screen now provides 1,500 units of utility but cost $1,500, the net price change is not 200 per cent but only 100per cent—the price change discounted by the additional utility provided.

In this framework, a Laspeyres index is inadequate. When relative prices change, consumers can optimise by switching from relatively more expensive to relatively cheaper goods at the same level of utility. Since in a Laspeyres formula the weights are fixed, the effect of this substitution behaviour is excluded. In other words, a Laspeyres price index *overstates* a (base-period utility) COLI.

An obvious problem is that statistical agencies do not know the level of utility attached to each of the goods in the basket. Hence, a 'true' COLI must be approximated on the basis of assumptions about the underlying utility functions of the consumers. A *superlative* index formula is one which approximates a true COLI using only price and quantity data (though only under a suitable set of assumptions about consumers' preferences).

For instance, under such assumptions, a Paasche index, that is, a current-period weighted price index $P_{01} = \sum q_{i1}p_{i1} / \sum q_{i1}p_{i0}$, with q_{i1} = current-period weights, can be shown to overstate a true (current-period utility) COLI (see for instance Diewert 2001: 172-3). This together with the earlier result that a Laspeyres index provides an upper limit for a true base-period utility COLI implies that there is a utility level in between the base- and current-period levels such that the COLI at that level of utility lies in between the observable Laspeyres and Paasche indices (*ibid*.). A Fisher index, that is, the geometric mean between a Laspeyres and Paasche index, is thus a superlative index. The Tornqvist index, similarly, is such a superlative index.

Many factors apart from narrowly defined economic goods and services may have an impact on consumers' utility levels. Just think of factors such as pollution, crime, longevity, health and many other general socio-economic aspects that do not neatly fall into product categories for which there are markets. One might also mention in- or decreasing product variety in this context, which may influence utility over and above the impact of the goods themselves. The Boskin Commission thus recommended using funds to study the effect of those factors.

Finally, the Commission recommended closer collaboration of the BLS with academic economists and a hearing of the economists when indexation decisions are made. The presupposition seems to have been either that the recommendations made by the Commission would not be fully implemented or that an upward bias, whose amount academic economists are in a better position to estimate, would remain even after full implementation.

The Boskin Commission estimated that the CPI overstates a COLI by 0.8 to 1.6 percentage points annually, their best point estimate being 1.1 per cent. Table 1 specifies the Boskin estimates for each individual category.

Sources of Bias	**Estimate**
Upper Level Substitution	0.15
Lower Level Substitution	0.25
New Products/Quality Change	0.60
New Outlets	0.10
Plausible Range	0.8 – 1.6

Table 2.1: The Boskin Estimate of the CPI Bias

Upper-level substitution refers to substitution *between* item strata, *e.g.* between 'Apples' and 'Pears'. Lower-level substitution is substitution *within* strata, *e.g.* between Royal Gala and Golden Delicious. The biases due to new goods and quality changes are jointly estimated, which makes sense since the distinction is quite arbitrary. Finally, the new outlets bias is due to the appearance of novel kinds of retailing channels such as discounters and the fact that consumers gradually shift from traditional to these new retailers.

In the period following the publication of the report, its findings were hotly debated inside and outside academic economics. Some critical voices regarding the Commission's results and methods (see Abraham 1997; Baker 1997) notwithstanding, many of the recommendations have been implemented subsequently (the remainder of this section largely follows Gordon 2000).

First, and foremost, the BLS agreed that the objective of the CPI is to measure a COLI (Bureau of Labor Statistics 1997). It is only consequent, then, to employ geometric weighting at the lower level because geometric weighting reduces the substitution bias. The BLS introduced geometric weighting for almost two thirds of product categories, effective with data from January 1999. At the upper level, the BLS switched from a period of about ten years to a two-year period in updating budget weights. This shift is thought to eliminate or at least reduce upper-level substitution bias.

A change that had been planned already before the publication of the Boskin report was to increase the speed of sampling rotation. Now telephone surveys are used more extensively, allowing to increase the sample size and focus on categories where products turn over and new goods are introduced more rapidly.

A number of adjustments concern quality changes. The pricing of hospital services is treated differently as of January, 1997. In the old methodology, hospital visits used to be priced on a 'dollar per day' basis. By contrast, the new methodology obtains prices for a sample of specified treatments for particular diseases. Further, the use of hedonic methods was extended. In particular, television sets and personal computers are now priced using hedonic regression methods. Finally, pollution control measures receive a different treatment. Prior to the report, changes in automobile or petrol characteristics due to air pollution mandates used to be treated as quality changes. As of January, 1999, the BLS regards such changes as changes in price rather than quality. That is, pollution mandates are treated as implicit taxes.

IV

In this section, I assess whether and to what extent the Boskin estimates and recommendations were evidence-based, where by 'evidence-based' I mean 'grounded in a systematic empirical investigation of the phenomenon of interest'. The most salient and significant recommendation is that the objective in measuring consumer prices should be a COLI. The belief that the CPI should be a COLI is widely accepted among economists, in the U.S. and elsewhere. But economists provide little if any justification for this choice of measurement objective.

There is no such thing as 'the' consumer-price level. This is clear from the fact that different indices address different questions. A Laspeyres price index, for instance, answers 'How much does a consumer's income have to change between a base and a current period for the consumer to be able to purchase the same selection of goods?', while a COLI answers 'How much does a consumer's income have to change between a base and a current period for the consumer to be able to achieve the same level of utility?' Every index carries its application with it in this way. Hence, given a certain purpose, the right or correct index number is that one that is most effective at serving the purpose (Diewert 1996; Diewert 2001; Fisher 1921; Mitchell 1915; Reinsdorff and Triplett 2004). Now, prominent uses for the CPI are the following (Schultze and Mackie 2002: 192):

- as a compensation measure to calculate how much is needed to reimburse recipients of social security and other public transfer payments against changes in the cost of living and for formal or informal use in wage setting;
- for inflation indexation in private contracts;
- as a measure of inflation for inflation-indexed Treasury bonds;
- as a measure with which to index the income tax system to keep it inflation neutral;

- as an output deflator for separating changes in gross domestic product (GDP) and its components into changes in prices and changes in real output; and
- as an inflation yardstick for the Federal Reserve and other macroeconomic policy makers.

Because the indexation of federal programmes has received the greatest attention in the public and academia, let me focus on this purpose. The idea behind indexation is to insulate benefit recipients from rising price levels, which would reduce the value of the transfers if these stayed nominally the same. In the U.S., benefit indexation was introduced in 1969 by the Nixon administration. In his Special Message to the Congress on Social Security, the president said the following:

> The impact of an inflation now in its fourth year has undermined the value of every Social Security check and requires that we once again increase the benefits to help those among the most severely victimized *by the rising cost of living*.
>
> I request that the Congress remedy the *real losses* to those who now receive Social Security benefits by increasing payments by 10 per cent.
>
> Beyond that step to set right today's inequity, I propose that the Congress *make certain once and for all* that the retired, the disabled and the dependent never again bear the brunt of inflation. *The way to prevent future unfairness is to attach the benefit schedule to the cost of living*.
>
> This will instill new security in Social Security. This will provide peace of mind to those concerned with their retirement years, and to their dependents.
>
> By acting to raise benefits now to meet the rise in the cost of living, we keep faith with today's recipients. By acting to make future benefit raises automatic with rises in the cost of living, we remove questions about future years; *we do much to remove this system from biennial politics*; and *we make fair treatment of beneficiaries a matter of certainty rather than a matter of hope*.

Let us first focus on the idea that indexing Social Security payments to the CPI insulates recipients from 'rises in the cost of living'. If this is the ultimate aim behind the CPI, then the Boskin recommendation that the CPI be a 'cost-of-living index' seems to be validated (but *cf*. Diewert 2001 who regards the measurement of a cost-of-living index as an important purpose in itself and distinguishes this purpose from that of payment escalation).

But this inference is a bit too quick. The economists' technical notion of 'cost of living' as 'cost of achieving a given level of utility' is only one among many connotations the term has. The reader is invited to verify this by consulting a number of English dictionaries. Here are some salient meanings (*cf*. Banzhaf 2001):

- 'the cost of meeting certain basic requirements',
- 'the cost of purchasing goods which are included in an accepted standard level of consumption' and
- 'the cost of maintaining one's usual standard'.

Each of these has different implication for the associated measurement procedures, mainly, of course, as regards the choice of the basket of goods. The first and second definitions contrast with the subjective nature of the economists' conception and call for intersubjectively accepted conceptions of 'basic requirements' and 'standard of consumption', respectively. While the first includes only essential goods such as food, clothing and shelter the second covers any definition as long as it is somehow socially legitimised. The third may include a household's standing *relative to others*, and thus come closer to an income rather than a price index.

Different definitions have different consequences for the wellbeing of benefits recipients. This shifts the discussion from 'what is the right index number?' to 'what

is the right concept of "cost of living"?'. At this point we leave the territory of solid empirical facts and enter the realm of evaluations. For no matter how much we investigate prices and goods, their traded quantities, their qualities and how they change over time, such analysis will not give us any clues as to the correctness of our concept of the cost of living. What kind of investigation should we seek instead?

The answer is, in my view, that this question should be settled in the exact same way as other questions regarding the aims of a society are settled. Societies must make up their minds about such decisions all the time: What's the age of consent? Should there be capital punishment? Should we allow smoking in bars? How do we treat users and dealers of recreational drugs? Should there be a national speed limit? Ought we to allow stem-cell research? And so on and so forth. Since this is not a treatise in applied political philosophy, I hesitate committing to a definite answer to our specific question (or any of the above, for that matter). But from the point of view of the methodology of economics it is important to see that such a question cannot be addressed by scientific means alone. There simply is no naked fact there, waiting to be picked up by a scientist. Rather, just as society decides whether to support certain groups in the first place, it should decide in the same way about concept of cost of living because that concept determines the cash value of the support.

Absent a decision about the correctness of the concept of cost of living by society (be it by explicit vote, some representational mechanism, a benevolent dictator, a philosopher king or what have you), it is thus hard to argue that the COLI is necessarily the wrong choice (but consider the discussion in the next Chapter). In what follows I will, for the time being and the sake of the argument, take the COLI concept as given.

If the CPI ought to be a COLI, it is correct to try to estimate the substitution bias, which arises from the fact that there is no substitution whatsoever in a fixed-basket price index. We need to be alert though. A superlative index approximates a true COLI only when certain assumptions can be made. The approximation result presupposes the truth of general equilibrium theory, that consumers' preferences are stable over time and have a certain form, and their relatively homogenous distribution in the group considered. None of these assumptions are necessarily or universally true or even true for the most part. The COLI approach implies that item substitution is caused by no other factor than price changes. In fact, item substitution may also be due to many other factors such as changes in tastes, limited availability, chance and error (see for instance the discussion about the difference between forced and voluntary substitution in Reinsdorff and Triplett 2004: 11 as well as the discussion below). Assuming general equilibrium theory tacitly supposes that these other factors are not at work but whether they are or not can and can and should be determined empirically.

The most controversial issue is, however, the treatment of quality changes. Including the bias due to new outlets the error in this area estimated by the Commission amounts to 0.7 per cent of the 1.1 per cent total. As mentioned above, the BLS employ an array of tools to deal with quality changes. Given these tools are already in place, the Boskin Commission is effectively saying that the BLS methods are insufficient to determine the exact effect of 'true' quality changes on utility. But how would the Commission know? The answer must be that they (think that they) employ superior methods to determine the impact of quality changes. This is surprising because they did not have any special research funds at their disposal to empirically investigate the issue. Hence, the Commission based their results on extrapolations from previous research and guesswork.

In the Final Report, the estimates of the bias are organised in product categories following the seven main categories of the CPI. Let us look at the arguments given in slight detail (all quotes and citations are from the relevant sections in Boskin *et al.* 1996).

1. *Food and Beverage*. The Boskin Commission notes that there is little if any published evidence on the impact of quality changes in this category. They argue, however, that this category has experienced great improvements over the past 30 or so years due to increased product variety, increased freshness of the produce and a greater selection available at supermarkets, eliminating the need to do time-consuming trips to specialised shops. They reckon: 'A conservative estimate of the value of extra variety and convenience might be 10 percent for food consumed at home other than produce, 20 percent for produce where the increased variety in winter (as well as summer farmers' markets) has been so notable, and 5 percent for alcoholic beverages where imported beer, microbreweries, and a greatly improved distribution of imported wines from all over the world have improved the standard of living. Increased variety and convenience in food away from home, in every price category from McDonalds to luxury restaurants…, can also be credited with a 10 percent premium', which amounts to a total bias in this category of 0.4 per cent per annum (weighted with CPI shares and converted into annual geometric growth rates).

2. *Housing*. The most important component within housing is shelter, and the Boskin Commission thinks that the BLS underestimates the value of appliances such as modern refrigerators, stoves/ovens as well as air conditioning and improved plumbing. Their 'conservative' estimate of this bias is 10 per cent over 40 years or 0.25 per cent annually. The Commission further argues that the BLS misses important

quality improvements in the non-shelter components of housing such as telephone and cable TV services. New communication technologies have improved clarity and reliability of phone calls, and 60 per cent of households now have cable TV along with its improved picture quality and variety of channels. The 'conservative' estimate in this category is 10 per cent per decade or one per cent annually. Unlike most other product categories, that of appliances and consumer electronics has been systematically researched—incidentally by one of the Commission members, *viz*. Robert Gordon (Gordon 1990). On the basis of this research the Commission estimates the bias to be 3 per cent for appliances, 4 per cent percent for radio-TV, including VCRs and camcorders, and 15 per cent for personal computers, which amounts to 0.1 percentage point annually after weighting.

3. *Apparel*. In this category too Robert Gordon has conducted a systematic study (Gordon 1996). On the basis of year-by-year comparisons of identical items from the Sears catalogue, Gordon estimates that the CPI overstates true apparel inflation by 1.92 per cent annually. Yet, the Commission 'conservatively' concludes that the bias is 1 per cent per year for apparel.

4. *Transportation*. Within the transportation component of the CPI new and used cars occupy a central place. In their calculation of the bias, the Commission treats mandatory anti-pollution devices as price rather than quality increases. Curiously, mandatory safety equipment such as seatbelts and crash-resistant bumpers continue to be treated as quality improvements 'since our feeling is that consumers see the connection between their own safety and the devices more directly than they do between anti-pollution devices and air quality'. This is irrelevant, however, for the determination of the current bias of the CPI since most of these new regulations were introduced in the period before 1983. But since 1983 the longevity of cars has

improved markedly, another fact overlooked by the BLS. The Commission argues that cars should be treated on a rental-equivalent basis similar to owner-occupied housing. If the lifespan of a car increases, the annual cost goes down as the same purchasing price can be distributed among more years. The bias due to this factor is estimated to amount to 0.59 per cent annually for the post-1987 period. Similar considerations apply to transportation-related products such as petrol. The Commission thinks there is a bias in this category of 0.25 per cent due to increased convenience from automatic credit-card readers built into pumps. Finally, Baily and Gordon 1988: 416, estimated a substantial bias due to the failure of the BLS to take discount air fares into account.

5. *Medical Care*. There are three primary categories in this area: prescription drugs, professional medical services and hospitals. The CPI for drugs is upwardly biased because of the treatment of generic drugs (prior to 1995) and the introduction of new drugs, according to the Commission. This conclusion is based on research by Commission member Zvi Grilliches and his collaborators (Bernt *et al.* 1996; Griliches and Cockburn 1994). The bias is estimated to be two per cent annually. In the physicians/hospital services category there is a bias due to the treatment of services on the based on input rather than outcomes (*cf*. above discussion). According to a number of studies of particular treatments, which assess the bias due to the BLS practice to measure 'days in hospital' rather than 'services pertaining to a specific condition', the CPI overstates true inflation by roughly 4.5 per cent. The Commission regards three per cent as a conservative estimate.

6. *Entertainment*. The Commission finds no bias in newspapers and magazines but assumes that sporting equipment and toys are subject to a slightly smaller bias than

Gordon found for appliances (two per cent per year as compared to three per cent for appliances).

7. *Other Goods and Services*. This category includes small personal care appliances such as hair dryers, which are subject to the same bias as other appliances according to the Commission. Further, personal financial services are 'conservatively' estimated to suffer from an annual bias of two per cent, due to improved technology such as cash machines.

The results found in each category can and have been questioned on the basis considerations very similar to the ones given by the Commission. For instance, the Commission members argue that the quality of the available food has constant improved over the past 30 years due to better variety, refrigeration technology and so forth. With the same degree of justification we can argue that the quality of our food has deteriorated in this period. The flip side of the greater convenience that accompanies the larger market share of big supermarkets is a greater concentration of business, which means greater market power of retailing companies, which often use this power to put pressure on food producers. This, in turn, forces food producers to cut costs and thus to use cheaper ingredients, more intensive farming methods and so on. Food scandals like BSE are only one effect of this process. Another is the use of cheaper inputs like hydrogenated vegetable fats instead of good oils, butter or lard and the use of animal waste products instead of muscle meat. The health effects of these and similar factors are well known, and looking at U.S. obesity figures lets one wonder whether the quality of nutrition really improved by as much as the Boskin Commission reckons.

To be sure, I do not want to argue that the Commission's *results* in the food and beverage category are necessarily mistaken. Maybe the factors cited by the Commission outweigh the just mentioned opposing factors by exactly the amount estimated. The problem is that we cannot tell—on the basis of the evidence given.

The exact same reflections apply to all categories. Maybe the average quality of housing improved over the past decades. At the same time the U.S. experienced the phenomenon of the 'urban sprawl'—suburbanisation—with all its potentially negative effects on wellbeing. Maybe cars have much better durability than they used to. But maybe we only use them for more years because we're forced to by income constraints or because we simply like to. Maybe the prescription drugs available today are much better than those 20 years ago. But maybe the pharmaceutical industry convinces us to suffer from diseases we would not have were it not for the drugs that treat them. We need to recall that the relevant quantity is not an objective fact about quality changes of various products. It is rather the effect of perceived quality changes on subjective wellbeing of consumers, on consumers' 'utility'. Guesswork in this area cannot substitute systematic research into consumers' opinions.

In summary, there is little systematic evidence for the correctness of the Boskin estimate of a 1.1 percentage point upward bias of the CPI. Accepting that the CPI should be a COLI, a judgement for which there is itself little supportive evidence, it is correct to calculate the bias that is due to consumers' substitution behaviour. The technique employed by Boskin *et al*. presupposes that all substitution is due to relative price changes, an assumption that lacks plausibility and should be substantiated empirically. Moreover, bias due to substitution makes up only 0.4 per cent out of the 1.1 per cent total. The remainder is due to changes in the quality of

goods and retailing channels, and the impact of such changes on consumers' utilities should be estimate empirically and not guessed.

It looks therefore as if the Boskin estimates and recommendations were based on dogma rather than the appropriate political process; and on guesswork rather than systematic empirical investigation. But perhaps this conclusion is too hasty. Don't many socio-political decisions require expertise rather than explicit vote, say, and judgement rather than standardised empirical test? The following section takes this suggestion serious and examines the prospects for basing such decisions on expert judgement.

V

Let us now return to the Nixon quote and focus on a different aspect. Here is the last sentence again: 'By acting to make future benefit raises automatic with rises in the cost of living, we remove questions about future years; *we do much to remove this system from biennial politics*; and *we make fair treatment of beneficiaries a matter of certainty rather than a matter of hope*.' In other words, the underlying aim of benefit indexation is to substitute *political judgment* with an automated or *mechanical procedure*.

In this *Trust in Numbers* (Porter 1995), Ted Porter argues that one driving factor behind the increasing mechanical quantification of political matters is the attempt to overcome mistrust in expert judgement. Much of the history of CPI measurement is an almost paradigmatic case study for this process. The CPI was first introduced in 1919 (though under a different name), following the Shipbuilding Labor Adjustment Board's decision to escalate wages by a price index. Federal programmes were first

CPI-indexed in 1969 (see Nixon's Special Message above), and the number of indexed programmes has since risen greatly.

Recall the Boskin Commission's recommendation number 15: 'Congress should establish a permanent (rotating) independent committee or commission of experts to review progress in this area every three years or so and advise it on the appropriate interpretation of the then current statistics'. One way to interpret this recommendation is to say that the Boskin Commission wants to resist this process of mechanisation or objectification of socio-political decisions. Porter points out in his book that such counteracting tendencies have been common in the past. Importantly, he shows that there is often a struggle between two conflicting aspirations: our desire to make *transparent* decisions on the one hand (by applying mechanical decision rules, which, in principle, can be checked by the public) and our desire to make *accurate* decisions on the other (by invoking experts).

We mistrust experts because they are humans and as such sometimes err. They will always act on some interests, and these interests may conflict with the purpose of the investigation. Moreover, expert judgement has an irreducible subjective element. This creates a problem for the public accountability of the decision making. An explicit procedure is transparent and can, in principle, be checked. Human judgement, by contrast, cannot be rationalised in the same way. Using it may create the impression that scientific decisions are up to a committee of elitist insiders whose exact working is foreclosed to ordinary people.

There are thus methodological reasons as well reasons having to do with the public interest in deriving scientific decisions to limit the role of expert judgement. And yet, there are few areas in science (and many other areas of inquiry) that are characterised

by problems simple enough to be addressed adequately by purely mechanical procedures. Porter describes one case where the experts were able to sustain the pressure to generate mechanical, quantitative evidence from the government bureaucracy for a comparatively long time: the Victorian 'gentleman actuaries'. An argument frequently used by the actuaries was that company principles, investments made and lives insured differ too much from case to case such that a set of standardised rules would necessarily lead to bias (in fact, this claim by the Victorian actuaries seems to be supported by evidence, see Ericsson and Lehmann 1996). What was called for was thus 'judgement and discretion'. But the necessary expertise was not based on particular elite from which the actuaries were recruited or even their academic training. Rather, it was the acquaintance with the local particularities of the business concerned that mattered. Porter summarises a lecture by actuary Henry Porter thus (*op. cit.*: 106):

> Men of experience recognize the crucial importance of 'judgment' in actuarial practice. Porter's lecture was a paean to 'judgment and experience,' which 'cannot be taught' but only acquired through an apprenticeship, as in all professions.

Stories about the expert wine taster, chicken sexer, chess or bridge player or arts dealer are legend. The chicken sexer merely looks at the back of the day-old chick and knows its sex on the basis of very subtle cues without really being able to explain how he came to the decision (Horsey 2002); the chess player 'sees' the next move without going through a complicated algorithm that could be instantiated on a machine; countless tests have been developed to tell a fake from a genuine piece of art—but they can all be tricked more easily than the trained expert.

The same is true of many areas in science. Experts reliably outperform novices in cases of difficult medical diagnoses, especially when diagnoses are visual, for

example diagnoses from X-rays or diagnoses of skin disorders. Experienced physicists exceed younger colleagues in problem solving because they have developed superior mental representations of paradigmatic situations and so on.

The trade off pointed out by Porter then seems to be real. Many domains of science and everyday inquiry involve a tacit element that often resists formalisation into a standardised rule. Experts can fill in the gap but at the cost of bringing in methodological and normative problems.

In order to trade off the conflicting desiderata adequately, a reasonable strategy may be to require that experts meet certain tests if their judgement is to be accepted as evidence. We calibrate a measurement instrument against a standard or investigate its proper functioning empirically in other ways. Analogously, we want to ascertain that experts 'work well' when their judgments are called for. In terms of the concepts introduced in Chapter 1, we can say that any expert judgement is prima facie evidence for a claim; the judgement is valid or sound evidence only if the experts are free of bias—if they pass the tests. The following set of principles is intuitively appealing (indeed, see the six conditions in Walton 1989: 60, which overlap with the following to some extent):

- **Subsidiarity Principle**. *Experts should be invoked only when there is reason to believe that they perform better than mechanical rules or that mechanical rules cannot be applied in the context of a given task.*

  If it is true that there is a trade-off between accuracy and transparency, we should seek to get at least one pole of the trade-off. It hardly makes sense to seek evidence from a procedure that is both inaccurate and intransparent.

- **Supporting Evidence Principle**. *In so far as possible, experts should underpin their judgement with supporting evidence.*

  Experts should judge *with* numbers rather than *instead of* numbers (Levy 2001; see also Ted Porter's response in Porter 2001). A prudent physician will, when making a complex medical diagnosis, support his or her judgement with external evidence in as much as this is possible and reasonable. Similarly, experts in any area should not ignore any piece of information that can improve the accuracy of a judgement.

- **Relevant Expertise Principle**. *We should seek advice only from an experts who is a specialist in the relevant area.*

  Why do experts sometimes outperform laypeople and standardised procedures? This is often because of an intimate acquaintance with the relevant subject area, that is, because of superior experience with the matter at hand. There is no reason to suppose that experts in *some* area produce accurate judgements in *any* area. We should not trust authorities per se but rather the years of training and practice in a given subject.

- **Democratic Principle**. *Expert advice should be made on the basis of the goals and values of the advice seeker, not the advisor.*

  Suppose you have a certain condition and there are two medical treatments to choose between. One promises a longer life on average but more pain; the other is risky but if it works there are good chances that you'll be pain free. Consulting your doctor, he can tell you his best guess what the figures are in your case, about

the effects and side effects of each treatment and so forth. But he cannot take for you the decision whether you prefer a higher risk or dying or a longer but more painful life. This is a value-decision that should be up to you, the patient. The same holds for all instances of client-expert relationships. Experts are experts about the likely consequences of alternative courses of action; but the aims the decision about the course of action is going to pursue should be chosen by the client.

- **Impartiality Principle**. *As a rule, the expert should not have a stake in the matter considered*.

  This can be used as a catch-all principle seeking to avoid obvious instances of bribery, partisanship, sectarianism and so on. Of course, there is no guarantee that the medical expert who is paid by the pharmaceutical industry will misjudge the efficacy of a new drug; nor is there a guarantee that the superannuated catholic priest ill advises a teenager about risks and virtues of birth control. But judgements by partial experts should be taken with special care and, if possible, avoided.

These seem intuitive enough. And yet all five principles appear to have been violated in the CPI controversy. Let us go through them one by one and, for the sake of variety, back to front.

Impartial experts. Far be it from me to insinuate that the Commission members were partial in the sense of having pursued a definite political or economic agenda with their work on the price index. However, it has been claimed that only those economists who testified to the effect that the CPI *over*stated consumer price inflation

were invited to join the Commission (Baker 1997). This, together with the fact that there were (and continue to be) dissenting views indicates that the Commission as a whole wasn't neutral on the topic. Perhaps each member gave his or her best judgement about the accuracy of the CPI, independently of any own or third-party interests. On the whole, however, the verdict the Commission reached—that the CPI significantly overstated consumer-price inflation—had been predictable before the Commission started their work.

Democratic experts. As discussed above, making alleged scientific judgements on the basis of utility theory smuggles in specific values that may or may not be the values a society would endorse by an explicit political procedure. Without evidence that such a procedure would have resulted in a different set of values, I of course cannot claim that the goals and values implicit in the Commission's reasoning was necessarily mistaken (though I will give some arguments to that effect in the following Chapter). The point here is once more methodological: the mistake was not to make it evidence that many measurement decisions behind the CPI are in fact normative, to put the implicit set of goals and values on the table and discuss the matter. Perhaps, for practical purposes, we can't do better than the Boskin Commission. But we shouldn't sweep the fact that the CPI is a value-loaden concept under the carpet.

Relevant authorities. All Commission members were economists. That raises the question whether economists, and only economists, are experts in the right field. I do not want to deny that economic expertise can contribute to making the CPI a more accurate measure of consumer-price inflation. After all, consumers, prices and goods are involved here, and what is economics but the systematic study of the relationships among these? Nevertheless, there are at least two problems with appealing *exclusively* to economic expertise. First, one of the purposes of measuring

consumer-price inflation is, as we have seen, to keep social security benefits recipients' standards of living constant. If that is so, one type of relevant expertise concerns the lives of benefits recipients. Thus, expertise from social workers, sociologists or perhaps political scientists, that is, specialists who studies these lives, should have been included.

Second, economists tend to look at the world through economic looking glasses (no more and no less than any specialist looks at the world from the perspective of his or her speciality). They tend to regard economic phenomena as being produced by a giant machine that satisfies the assumptions of general equilibrium theory (GET). And thus they tend to make inferences on the basis of what would be the case were GET true. But we need to know what is the case in this world, not what would be the case had GET been true. This matters in particular in the area of assessing the impact of quality changes on consumers' wellbeings. Quite clearly, these should be studied empirically, and not on the basis of casual observation and GET.

And yet, one of the Commission members, Robert Gordon, suggests that such detailed research is not needed (Gordon 2000: 27, emphasis original): 'For instance, even though we will never precisely measure the value of the invention of the jet airplane, as economists we *know* that consumer surplus triangles have an area that is positive rather than zero'. This is an intriguing statement. Of course, it is *plausible* to assume that substituting jet airplanes for propeller-driven ones bumps up consumer welfare. One may even agree that it would be *objectively rational* for consumers to prefer jet to propeller planes. But neither claim is a substitute for evidence to the effect that consumers value jet planes more highly than propeller planes. The first claim is a piece of guesswork. The second, an instance of what one may call the

'reverse naturalistic fallacy'—an argument from an 'ought' to an 'is': since people ought to prefer jet airplanes, they in fact do so.

A inference the Boskin Commission uses time and again is that from 'the market share of some good x_i in a market X increases' to 'consumers prefer x_i to its alternatives x_j $(i \neq j)$'. But this inference rule is fallacious in the same way that 'post hoc ergo propter hoc' is fallacious. This is because many other factors apart from consumers' preferences may be responsible for the shift in market share. To name but a few, the relevant outlets (or producers, as in the case of airplanes) may just stop selling the older good; consumers' preference are induced to change (by, say, aggressive advertising or peer pressure); a third or environmental variable changes so that the preference between *i* and *j* is reversed; consumers are forced or tricked into buying *i* without having a genuine preference for it. Unless we control for these other factors, the inference may always be misleading.

The advisors in the Boskin Commission are experts in economic theory. But advice based on economic theory, that is, economic theory *alone*, can lead to damaging consequences. I could not state it better than Paul Klemperer (an auction theorist!) who traces this point back to Alfred Marshall (Klemperer 2004: 124):

> Some academics also need to widen the scope of their analyses beyond the confines of their models which, while elegant, are often short on real-world detail. Marshall always emphasized the importance of a deep 'historical knowledge of any area being investigated and referred again and again to the complexity of economic problems and the naivety of simple hypotheses' (Sills 1968). Employing 'know it all' consultants with narrowly focused theories instead of experienced people with a good knowledge of the wider context can sometimes lead to disaster.

Therefore, there is one reason to believe that economic expertise is not sufficient; and another reason that it might positively impair good judgement.

External evidence. As we have seen above, in most categories the Boskin Commission provided *some* numbers to support their estimates of the bias. But the external evidence the Commission cites often appears to be of little relevance for the problem at hand. In terms of the distinctions made in Chapter 1, we can say that the Commission provided at best *valid* but not *sound* evidence. That is, what they cited may or may not have been good evidence for hypotheses regarding certain categories—but not for the hypotheses at stake.

To mention a few examples, it may be the case that what is true of breakfast cereals is true of food and beverage in general (Hausman 1996, cited in the section on Food and Beverage). Maybe results for anti-depressants are true of drugs in general (Bernt *et al.* 1996, cited in the section on Medical Services), and maybe what is true of heart attack and cataract surgery is true of all physicians' and hospital services (Cutler *et al.* 1996 and Shapiro and Wilcox 1996, respectively, also cited in the section on Medical Services). But without further argument, there is no reason to believe that this should be so.

A curious piece of reasoning appears also in the section on Medical Services. Boskin *et al*. point out that the biases estimated in two studies on heart attack and cataract surgery, respectively, almost agree: 'The closeness in the Cutler et. al. [sic] and Shapiro-Wilcox studies of quite different medical procedures is striking' and conclude that this must mean something about the bias in the whole category.

The argument indeed resembles an argument to the effect that if the results of two different procedures purporting to detect the same phenomenon coincide, the phenomenon is real rather than an artefact of the procedure. However, we cannot ignore the 'purporting to detect the same phenomenon' here. The two procedures considered by the Boskin Commission measure the quality-adjusted price change of heart attack surgery and cataract surgery, respectively—and the two are clearly two different things. A coincidence of the two numbers is not indicative of anything.

Are experts better than mechanical rules? Do we need expert judgement to determine the effect of quality changes on consumers' wellbeing? On the basis of the known evidence, this is hard to decide conclusively but it is important to notice is that this area surely is not inaccessible to alternative evidence-generating methods (*cf*. the Stigler Committee's recommendation that psychologists and other specialists could use surveys to appraise consumers' perceptions of the relative qualities of varieties: Stigler *et al.* 1961: 37). Evidence about the impact of new goods on consumers' wellbeing can be investigated by means of surveys, experiments, field studies and econometrically. Therefore, unless it can be demonstrated that experts systematically beat these methods in terms of either accuracy or efficiency, there is no reason to believe that experts should have been called for this task.

In case of a conflict between a normative theory (in this case given by the five principles) and practice (in this case given by the work of the Boskin Commission), in order to restore consistency, we should either revise the theory or we have to maintain that practice is deficient. How shall we proceed in this case? I think we need a bit of both. The rules clearly need revision; but it is just as clear that something went wrong in the CPI controversy. To reach some conclusions let us, once more, look at each principle in turn.

Subsidiarity. The problem with subsidiarity is that knowledge which of a set of alternative measurement procedures is more accurate is often hard to come by. There exists an extensive literature about expert judgement in many areas (for a recent discussion in the context of economic methodology, see Angner 2006; for a recent book on expert political judgement, Tetlock 2006). Most of this literature, however, presupposes that experts' performance can be assessed relative to some objective standard. For example, some studies measure the 'overconfidence' of an expert as the difference between the stated subjective degree of believe in the correctness of an answer in a questionnaire and the true frequency of correct answers of a given kind in the same questionnaire. Another method is to make experts predict a future event of a certain kind for a definite date and simply to wait till that date and see whether or not the event has occurred (which is assumed to be straightforwardly observable). Thus we know that corn judges systematically misforecast crop yield and misgrade corn because we can wait and observe yield and have accurate methods to grade the product (see Shanteau and Stewart 1992). In the same way we know that expert chicken sexers beat mechanical procedures. Chicken sexing is an art only for a very young chick—its sex is obvious even to the layman when it is five to six weeks old. We therefore only have to wait and see.

The problem we are facing here, and no doubt the methodologically more involved and interesting problem obtains when there is no such objective standard. There is no alternative method to determine whether an arts dealer has correctly judged a painting to be genuine or fake, even with hindsight. Perhaps sometimes new evidence comes in, for example, that a given painter has used a certain technique or material (which was also used for the disputed piece) for works whose origin is uncontested. Often, however, the expert's judgement will be all there is. In these

cases we simply don't know how accurate expert judgement as a whole is at a given task.

In such cases, does it seem reasonable to require to have a good reason to believe that an expert is better than a mechanical procedure at a given task? I think that the core of truth in this principle is that it should never be presupposed that only experts can be reliable generators of evidence in a given area. Like all evidential methods, expert judgement comes with its own set of virtues and vices attached. Its biggest virtue is also its biggest vice, namely, that experts are humans. As such they are more adaptable to particular circumstances and they can 'judge' as opposed to merely follow an algorithm. And as such they also take bribes, act on interests other than the client's, display overconfidence and get tired. In any domain that has even a mild degree of complexity, we will never get rid of expertise altogether. It is important thus to keep in mind the disadvantages it is necessarily associated with, decide accordingly and try to supplement expert judgement with alternative methods wherever possible and reasonable.

Supporting Evidence. It is a minimum requirement of rationality to take any relevant evidence into account and thus I don't think that there is anything wrong with this principle, even as formulated above. It should be qualified that the evidence be sound in the sense introduced in Chapter 1. This is precisely the qualification needed in the CPI controversy. It is not that the Boskin Commission ignored external evidence; but the evidence they took into account was of a notoriously bad quality. Prima facie evidence is, prima facie, not evidence for anything. And evidence, even if it is good (*i.e.*, valid) evidence for some hypothesis might be entirely irrelevant to the hypothesis under investigation. Only sound evidence should therefore play a role.

Relevant expertise. The discussion of this and the final principle has shown that experts should not be assessed one by one but rather as a group or committee. In many cases it will not be possible to find one person who knows enough to count as an expert in the relevant domain. This is especially so when problems call for interdisciplinary approaches. But this does not mean that the relevant expertise cannot be found in a committee of experts. The more complex the problem, the larger the committee will have to be. To repeat, in this area too the Senate could have done better, namely, by appointing a more heterogeneous Commission.

Democracy. I think, as a rule, it is correct to say that, if goals and values are at stake, it should be the goals and values of the client, not the expert. Consider the following case by Gerd Gigerenzer (Gigerenzer 2002). According to the psychologist, many women 'ascribe almost magical powers to mammography' in reducing the risk of breast cancer, due to the information they receive from the media, physicians and health organisations (p. 74). They thereby neglect that mammography also incurs costs in form of adverse psychological and physical effects of false positives, similar adverse psychological and physical effects of detected non- or slow-progressive cancers, radiation-induced cancers, and, in addition, the financial costs for the health system for carrying out unnecessary tests and treatments *etc*. Gigerenzer argues that women should be informed about these likely costs as well as the benefits of mammography prior to the physician's recommendation so that they can make an informed decision. He concludes (p. 71):

> Before age 50, mammography does not seem to have benefits, only costs. Women at age 50, however, face the question of whether the potential benefits outweigh the costs. Each woman must decide for herself what the answer is. Her physician can help her understand what the benefits and costs are, but not how to weigh them. Her decision depends crucially on her goals, such as peace of mind, keeping her body unscarred, or a

> willingness to take (or not to take) the chance that she is one of the few who benefit from screening.

The same should, in principle, be true of the socio-economic decision-making process. But we have to qualify in a number of ways. First, not everybody has the mental capacity to know what is best for him or herself. This is quite obviously true for young children and the mentally impaired but there can be situations in which an otherwise completely rational person should leave decisions about his or her aims to a benevolent outsider. Second, goals and values, even to the most rational among us, are not always there to be observed straightforwardly and clearly. Rather, they form slowly and gradually in a process of deliberation—in which experts can and should take a part. I still think it is dumb to ask an old catholic priest for advice about family planning—mostly because we know his answer. But what the priest might help the client do is formulate her aims more clearly, see what the moral choices are and simply assist her in finding out what she wants (in the best of all possible worlds, that is). In this area too, it is a great lacuna in the Boskin report not to have made clear that moral choices are involved. Third, when political decisions are concern it might not always be clear who exactly the client is. The Boskin Commission reported to the Senate Finance Committee. Ideally, the SFC should act as a mere representative—but for whom? The relevant group in this case is the (U.S.) consumers. But the consumers are in fact a quite heterogeneous group. In particular we have to expect conflicts of interest between benefits recipients and tax payers. Such conflicts of interest arise whenever there are important choices, and they should be arbitrated (in principle) in the same way as conflicts in other cases. Of course, there cannot be an explicit vote on minute details of the measurement procedure associated with CPI. But there is no reason to suppose that such details cannot be

determined in a way that is as satisfactory as the bringing about of decisions about other moral choices.

Impartiality. The problem about the requirement that experts be impartial is that there won't always be impartial experts. Who outside the Catholic Church will be an expert on matters catholic? If that's not enough, think of Opus Dei or of the Free Masons. More seriously, the nature of the problem at hand will often make likely that all available experts are partial in one or another way. Issues with global consequences are a case in point. No matter whether the expert is an environmentalist, a rural farmer in the Amazon, a representative of the car industry or the head of the George W. Bush administration, he will always have *some* stake in the issue of global warming. Even the most selfless and well meaning physician will be subject to some external forces, be they constituted by national health care systems, patients' direct payments, superiors and administration (say, when employed by a hospital) and so on and so forth. If sacrificing the goals and values of one patient will result in helping many other patients to pursue their goals and values, this physician might, and quite reasonably, not make the recommendation that is in that patient's best interest.

These difficulties notwithstanding there are a number of clear cases of partiality that should be ruled out such as bribery and obvious sectarianism. In more subtle cases, it will help to publicise affiliations. It is one thing to hear an expert reporting on the efficacy of a new drug; it is quite another hearing that expert when it is known that he works for the pharmaceutical industry that will profit in case the drug is licensed. Other cases will be far less obvious. The U.K. public policy regarding sunlight exposure was formulated by dermatologists—which, prima facie, does not seem to be a bad thing as sunlight most directly affects the skin. One consequence of this is

that the policy focuses too much on the potential risks of sunlight exposure (because sunlight is implicated in melanoma and other forms of skin cancer) and not enough on the potential benefits (because the most substantial benefits are not directly related with the skin). The dermatologists are partial because they see only part of the problem, and not necessarily even the most important part (or so argues Oliver Gillie, see Gillie 2004). Here too it could help to make clear that this policy was formulated having in mind the consequences for the skin alone (or better still, to try to formulate a more balanced policy that weighs off risks and benefits for overall health).

VI

What, thus far, are the lessons from the CPI controversy for methodology and expert judgement? When does a report by the Bureau of Labor Statistics that 'The Consumer Price Index for All Urban Consumers (CPI-U) decreased 0.5 percent in October [2006], before seasonal adjustment' provide evidence for the hypothesis that (U.S. urban) consumer prices have decreased by half a percentage point in that period? And what kind of evidence?

To satisfy readers who enjoy labels, let me point out that the lessons have a pragmatist, an empiricist and a political aspect. The major methodological problems in CPI measurement are:

- How do we aggregate prices (that is, what's the right index formula)?
- How do we deal with changes in the market place?

There is no one correct answer to the first question. Whether or not an index number formula is 'correct' depends on the purpose pursued. The point I am making here is thus roughly pragmatist in character: there is no 'true' inflation rate; but there are numerous inflation indices that are more or less useful relative to the given purpose. Once the purpose is given, the numerical determination of the index should be based on empirical evidence. This is the empiricist aspect of the lessons. If, as I have taken for granted in this Chapter, the CPI purpose is to measure a constant-utility index, then the effects of various changes in the market place on consumers' utility should be estimated empirically. This is the second major methodological issue as measuring utility is hard. Perhaps estimating the effect of changes in the market place on utility is practically speaking impossible; perhaps utility can only be measured (by standard empirical methods) under restrictive conditions regarding consumer behaviour, and these conditions are not met, by and large, for typical consumers. But then the Boskin Commission had an answer to this problem: expert judgement. Expert judgement is often called for when standardised or mechanical procedures do not yield satisfactory results. When regarded as a measurement instrument, an expert should be tested for reliability just like any other instrument. To this effect I have described five principles or tests that seemed intuitively appealing. We saw, however, that none of them was met in the CPI controversy. This led us to both revise and qualify the principles as well as to deduce that the Boskin Commission had a number of shortcomings in this respect.

The American pragmatists saw that values pervade science through and through. The CPI is no exception. Values come in immediately when the index number purpose is determined. The Nixon administration chained social security payments to the CPI in order to keep benefits recipients' cost of living constant. But what is the 'cost of living'? We have seen that the economists' answer, namely the 'cost of

achieving a given level of utility' is only one among many, and that different answers have different implications for the measurement of the CPI. Which answer is right? My view is, again, that there is no one correct solution. Rather, answers to questions of this kind have to be legitimised in the same way as other normative decisions in a society are legitimised. The following Chapter investigates whether we cannot go beyond this platitude with the information at hand.