

SECTION 7

INDEXES: PROPERTIES AND PROBLEMS

Following on from [Section 5](#), which introduced using indexes to measure deprivation, and [Section 6](#), which introduced a selection of the better-known indexes, this section examines the construction and properties of indexes, and some of the problems that can arise.

7.1 Introduction

Indexes are one of the basic tools for measuring inequality. They are widely used and new ones continue to be devised. Indexes are available to measure most aspects of health and disease, quality of life and many interpretations of deprivation. They are often employed where there are no simple or direct measures of a phenomenon.

Indexes vary in complexity. They may represent the core phenomenon well, partially or poorly; and a particular index may have properties, including statistical properties, that makes it ideal or quite unsuitable for specific applications.

This section examines the construction and properties of indexes and possible consequences of using them uncritically. The material is relatively technical, but, not excessively statistical. If you require additional statistical details, there are several excellent references which are referred to at various points in the discussion.

The section opens with an example describing the measurement of health with indexes of different levels of complexity. Next there is a more formal summary of the structure and construction of an index, followed by details of the main methods of testing. This section ends with a discussion of the problems that arise if an index is poorly constructed or not suited for its purpose, and what can happen if users forget that an index is only a constructed measure.

7.2 Measuring Disability and Limiting Illness with Indicators and Indexes

This extended example will illustrate the different degrees of complexity of indexes from simple one-component measures, to elaborate multi-level constructions. The example concentrates on how one might measure differences in the incidence of limiting or disabling illness between electoral wards.

Single component indexes

The question on the incidence of disabling illness can be answered with a single item from recent Decennial Censuses. It asks whether each member of the household has a Limiting Long-term Illness (LLTI). This data item is widely used in U.K. health indexes and is often included in both national and local health surveys. When divided by a count of the number of residents in households, it provides an easily computed index of local health, which, in technical terms, is a single component index.⁵

Single domain, multiple component indexes

It can be argued that the Census item on LLTI is too restricted and perhaps too subjective to use on its own. It might be improved if combined with more 'objective' measures, such as the numbers of people claiming state benefits relating to disability (ward level data on disability claimants is available on the ONS neighbourhood statistics website). Such a composite index is more 'objective' than the original single component measurement and also now has a different meaning. However, although it is based on two components or indicators, it still addresses a single domain (incidence of limiting illness).

There are several technical and theoretical issues when constructing domains from more than one component:

- ◆ Why choose these two items, rather than other possible measures of (or proxies for) limiting illness? Is there any theoretical or statistical justification for the choice?
- ◆ How do we combine the different variables, in this case, data on self-report LLTI and rates of benefit claims?
- ◆ Do we transform the variables to have similar values and distributional characteristics?
- ◆ Do we weight their relative contribution to an overall score?

⁵ In practice, one would not usually compare crude rates of LLTI, but would try to adjust for differences in the age–sex composition of the ward populations. Standardising by age and sex is described in [Section 3.1](#).

We can broaden the index further to cover more aspects of (ill)health and its proxies, such as mortality. This combination is similar to the health domain of The Welsh IMD, in which three other items are added to the two we have chosen (Box 22). The five components were selected from a larger group that were suggested by various consultation exercises and literature reviews. The choice was narrowed on practical grounds, such as lack of data at a suitable level, or data that were insufficiently robust (e.g. infant mortality ratio). Other candidate variables, such as the numbers of people using alcohol or drug misuse services, were rejected because their values would be too dependent on the availability of relevant services, and others such as poor dental health amongst children were thought to be too specific to be markers of general health.

Box 22

Components of the health deprivation domain of the Welsh IMD

- ◆ Age- and sex-SMRs for people under 65 for 1995–2000.
- ◆ People receiving attendance allowance or DSS for 1998.
- ◆ People (aged 16–59) receiving incapacity benefit or severe disablement allowance for 1998 and 1999.
- ◆ Age- and sex-standardised ratio of LLTI (1991 Census).
- ◆ Proportion of births of low birth weight (<2500g) ONS for 1993–1997.

Clearly, as its authors intend, we now have a measure that tries to measure most aspects of poor general health, albeit limited by data availability.

Even a casual glance at the list of indicators suggests there may be difficulties combining them; not only do they have different metrics – some are rates, others are standardised ratios, but they also contribute unequally to the overall score.

In the case of the Welsh IMD, as there was no a priori (theoretical) basis for combining the items, the devisors used factor analysis to examine the patterns of correlations between the indicators. Unlike the conventional use of factor analysis, where one attempts to identify different factors amongst groups of variables, their interest was in “testing a one-common factor model against the possibility of there being more than one”. If other meaningful factors were found, this would suggest that the chosen set of indicators was not measuring a single phenomenon. When the one-factor model was found to be successful, the coefficients of the variables were then used as weights to combine the components into a single domain score.

Multiple domain indexes

So far, we have been regarding health as a single domain (at least for the purpose of measurement) represented by a diverse set of indicators. The multi-dimensional notion of health was discussed in [Section 4.3.9](#) in relation to instruments such as the SF-36 that measure several health domains. In the case of the SF-36 these are: physical functioning; social functioning; energy/vitality; physical impact on social role; emotional impact on social role; mental health; experienced pain; and general health. It is rare to find a health index based on census and administrative data covering the same topics as those based on questionnaires, but such measures could be constructed using, for example, Hospital Episode Statistics (HES).

The domains in multiple deprivation indexes are usually more diverse than different aspects of health. For example, the Welsh IMD has six domains ([Box 23](#)), including the health domain discussed above. This is an example of one of the more structurally complex indexes: a multi-domain, multi-component measure.

Box 23

The six domains of the Welsh IMD (and their weightings)

Income deprivation (25%)
 Employment deprivation (25%)
 Health deprivation and disability (15%)
 Education, skills and training deprivation (15%)
 Housing deprivation (10%)
 Geographical access to services (10%)

By this example, we have tried to illustrate the range of structural complexity found in indexes; these are presented more formally below.

7.3 Key Aspects of the Construction and Structure of Indexes

The basic anatomy of an index

The most elaborate indexes have at least three levels:

- ◆ The lowest level made up of the component, sometimes described as an indicator or variable.
- ◆ The next level, the domain or dimension, is made up of one or more components.
- ◆ The last level is the overall index, comprised of a collection of domains.

Simpler indexes may only have one or two levels. Where there are multiple domains, opinions differ on the value of presenting separate domain scores, or just an overall index score.

Identifying and specifying domains

There are three main methods by which domains are selected:

- ◆ From theory. This is relatively rare and relies on the availability of a theory with sufficient detail for them to be translated in to the domains of an index.
- ◆ By other normative routes. This ranges from simple consensus to a multi-staged process with various levels of public consultation.
- ◆ Through statistical techniques, such as factor analysis. This may involve sifting through large numbers of candidate variables to identify groups that might be interpreted as domains.

Selecting the components of domains

This can be the crucial phase and explains why broadly defined domains are progressively narrowed and redefined by having to work with limited data. The selection of components often involves several stages:

- (1) Deciding what would be appropriate items to include in a domain.
- (2) Checking data availability.
- (3) Checking data reliability and discarding or transforming unreliable sources.
- (4) Examining the correlations between components. Statistical techniques such as Cronbach's alpha and factor analysis may also be used to help decide which variables are central to a domain and those that might be excluded or transferred to other domains.

How the components of domains are combined

This will usually involve two stages:

- (1) Some type of transformation to give all the components similar distributional characteristics (transforming to a Z score is the most common).
- (2) Some type of weighting (often using statistical packages).

How the domain scores are combined

Similar methods are used to combine domain scores into an overall index score (Box 24). At this stage it is worth drawing a distinction between indexes, where:

- ◆ The weights are derived from statistical techniques.
- ◆ The weights are assigned relative importance by index's devisors.

Weights are generally chosen to reflect the relative importance of domains, but other criteria such as data reliability are sometimes used.

Box 24**Welsh IMD: Combination of domain scores into the overall index score**

Because the six domain scores are produced different ways and have, different units of measurements, they need to be transformed into a common metric before being combined into an overall value.

The devisors of the scale rejected two of the more conventional methods of transforming to Z scores or ranks. Essentially one would rank ward (electoral division) scores, but transform the rank (scores) to an exponential distribution, to re-introduce some measure of the distance between observations that is lost when the scores are converted to ranks.

The relative weights of the domains was decided by various consultative exercises and by reference to literature, rather than by any statistical techniques.

7.4 Testing an Index

7.4.1 Introduction

It is important to establish if an index measures what it is intended to measure. We take it for granted that physical measures such as a ruler or set of scales, will have been tested to certain standards, but most indexes of deprivation have had little systematic testing, in contrast to measures of general health, where there is a small industry devoted to the psychometric testing and validation of measures.

The literature on measures of health demonstrates the importance placed on psychometric approaches to testing. In the case of health questionnaires, this is understandable, given their increasing use in clinical trials and other settings to test the effectiveness of interventions. Nevertheless, there is some concern that the balance has shifted too far towards statistical rather than substantive criteria – that too much attention is paid to tests such as Cronbach's alpha, rather than the meaning of a measure and its relevance to intended applications.

Indexes are usually assessed on three criteria: validity, reliability and responsiveness. Unfortunately, there are many different interpretations of these criteria. For example, Hays and Hadorn [208] argue that responsiveness is better regarded as a form of validity. However, there is agreement on most of the basic principles and those with an interest in the statistical detail should refer to Streiner and Norman [195]; McDowell and Newell [142].

7.4.2 Validity

There are many forms of validity testing; three approaches – content, criterion and construct validity, are most relevant here.

Content validity, as its name suggests, is concerned with the selection of components and domains in an index (or the questions in a health status questionnaire). With a multi-dimensional instrument there are three main requirements:

- ◆ The main topic headings (the domains) should be appropriate to the overall aims of the measurement.
- ◆ The components of each domain should relate to the definition and aim of that domain.
- ◆ The components should be a representative sample of those that might have been chosen, and should give uniform coverage of the full range of each domain.

For questionnaires, the wording should be intelligible to respondents, and unlikely to be misunderstood or offensive.

The simplest methods for content validity is review by expert panels or a pilot with samples of representative respondents. Such assessments are rarely repeated when a questionnaire is in regular use, thus, its users may fail to notice if the instrument needs to be adapted to reflect changes in the phenomenon being measured, and, in the case of questionnaires, cultural changes overtime.

One of the more important reasons for testing for content validity is to assess and improve the reliability of the instrument. Measurements will tend to be more reliable if all the components of an index address the same core concept.

Many appropriate statistical techniques are available for content validity. Multi-trait scaling and factor analysis were employed in developing the SF-36 questionnaire both to eliminate items that were least related to the intended constructs and to test whether response patterns reflected the intended conceptual structure.

The principle of *criterion validity* is derived from the testing of physical instrumentation. The aim is to compare a new instrument against a gold standard for which it will then be a substitute. With physical instrumentation, the measurements from two instruments must be very highly correlated if one is to substitute for another – even values of 0.99 may not be acceptable.

The key principles underlying criterion validity in physical measurement – substitution and interchangeability – are hard to apply to indexes and questionnaires, not least because there are no gold standards and no two instruments seem to be designed to measure the same thing. Nevertheless, there is a great deal of this type of testing in relation to measures of health. It is possible to correlate one measure (such as the SF-36) against another (such as the NHP), or to correlate domains of the measure

with single items within the measure itself as shown by Jenkinson et al. [209]. Understandably, the values of the correlations in these comparisons are much lower than those required for one instrument to substitute for another. Consequently, the hypothesis, “is there any evidence that the two instruments are not providing equivalent measurements?” is replaced by the much weaker, “is there any evidence that the measurements are not unrelated?”.

A common problem in criterion validity is when there is no gold standard to use as a benchmark. One well-known exception is where several indexes have been validated against a definition of poverty derived from the Breadline Britain Survey. However, it is far more common for validity to be tested by analysing the correlations between the index score and phenomena which are thought to be related to deprivation. For example, Lee et al. [193] compute correlations between the 10 indexes they review and three factors that are either regarded as part or consequences of deprivation (see Section 5.4 for more details).

Construct validity tests for predicted associations. For example, if there are theoretical grounds for believing that perceived general health should be associated with the number of visits to a GP, or levels of self-medication, then this can be formally tested, and, if confirmed, will add to the evidence of construct validity. Trying to confirm plausible associations and disprove the implausible can be a long process. Often, there will be no definitive answers and, at present, there are no formal ways to weigh the overall evidence. Hence, it is unsurprising that tests of construct validity have been criticised for their failure to set formal hypotheses or specify in advance what will represent significant evidence [142]. The whole process is better described as an art than a science.

Reaching conclusions on construct validity is further complicated by the problem that evidence tends to be interpreted in two different ways: to use these associations to test whether the instrument is a good measure of the intended constructs; or to look for associations that will confirm and clarify the constructs themselves. There are major differences of statistical principle here, some of which are discussed by Streiner and Norman [195]. It is also very important to be clear about the different questions being answered by different indexes, and therefore what any given test for construct validity is trying to achieve (Table 7.1).

7.4.3 Reliability

Reliability may be defined as the capacity to produce the same result in precisely the same circumstances. This is a most relevant to physical instrumentation that are used repeatedly, such as thermometers or speedometers. With physical instrumentation, the classic test for reliability is to take repeated measurements in the same conditions. With health status questionnaires, reliability is evaluated by administering the instrument on

Table 7.1. *Indexes and the corresponding questions*

<i>Indexes</i>	<i>Questions</i>
Indexes based on death	Who dies?
Excess over 'average'	Excess over 'average'
Relative likelihoods	Relative likelihoods
Mortality rates	
Standardised mortality rates	
Ratio of mortality rates or of Standardised mortality ratios	
Percentage of survivors	Who lives?
Standardised survivorship ratios	
Ratio of survivor rates or of SSRs	
Indexes based on morbidity	
Nottingham health profile	Subjective assessment illness
Symptoms reports	occurrences 'objective' morbidity
Condition and/or disease incidence/prevalence	

two separate occasions, separated by a time interval, anywhere from several days to several months, sufficiently long to minimise memory effects but sufficiently short to reduce the likelihood of health having changed. During this period, there will need to be an independent assessment of changes in health, so respondents whose health has changed can be excluded from the re-test. It is difficult to find suitable controls since most health profiles are unique in what they measure. Measurements used as controls have included single item self-reports of health.

Because, it is hard to apply such procedures to indexes, a variety of indirect tests are used instead. Most of these are based on the assumption that a measure will be reliable if its components are highly inter-correlated. One method, the split-test approach, divides the components of an index or questionnaire in half. A single sample of test data is required, and the correlations between the two halves are computed. This technique is limited to measures that can be split into equivalent halves. This examines the correlations between all the components which are summarised by test statistics such as the KR-20 (Kuder and Richardson formula, 20) and Cronbach's alpha.

Very high reliability is not necessarily good as it may point to redundancy, e.g. several components providing very similar measurements. In such cases, it may be possible to reduce the number of components, making the instrument easier to use. Moreover, high reliability is most likely to be a feature of uni-dimensional measures, so if the instrument is intended to measure a multi-dimensional phenomenon, high reliability may be a sign of a partial measurement.

As with most of these psychometric assessments, the acceptable level of reliability will depend on the application. The most demanding applications

are those measuring changes in individual's health or average health scores for small groups.

For more on statistical tests for reliability we recommend Streiner and Norman [195] and Nunnally [196].

7.4.4 Responsiveness

Responsiveness denotes the capacity of the instrument to measure difference or change. For example, if the aim is to map demographic variations in health then the ability to detect cross-sectional differences between the health status of different age, gender and condition groups may be sufficient evidence of responsiveness. (See [Box 25](#)).

Greater responsiveness is needed for clinical applications where an instrument is required to detect changes in health due to an intervention. Although it may be sufficient to show that the instrument can differentiate between those with different levels of clinical severity for the same condition, more often it will be necessary to demonstrate the ability to detect change due to treatment. In all such tests there is a problem of what should count as significant change. Statistically, significant differences in health status scores may not correspond to clinicians', patients' or carers' views of significant change. Conversely, the changes that these groups regard as significant may not be detected by the instrument.

Responsiveness is normally tested by piloting the instrument in conditions similar to the intended application. However, inspecting the contents of a measure and its distributional properties in the general population may be a guide to its responsiveness. There are two key points to check:

- ◆ Are the end-points suitably defined, e.g. will there be 'floor' or 'ceiling' effects?
- ◆ Is there good coverage of all the intermediate points – the number and spacing of items and response levels?

Box 25

Responsiveness – a sensitive indicator

An indicator should be sensitive. If someone's socio-economic classification according to an index changes, it is obviously important that this change can be related to a change in that person's position in the social hierarchy, to which the classification refers. Goldthorpe [210] correctly deploys this argument against the use of women's own occupation as the basis for their social class because married women might change their job for life-cycle reasons unrelated to any change in their social position. It is unrealistic to be rigorous here, but it is reasonable to demand that the index be relatively sensitive to 'real' change and relatively stable when there is no 'real' change.

7.5 The Purpose of Testing

One problem with tests such as Cronbach's alpha is the focus on the internal characteristics of an instrument rather than its relationship to the outside world. The statistical testing of indicators should not distract us from asking basic questions about the meaning and performance of indicators or their suitability for specific applications.

For example, depending on the intended use an index of health might:

- ◆ Reflect the socio-economic dimension to inequalities in health.
- ◆ Reflect the experience of the entire population.
- ◆ Be sensitive to changes in the distribution of the population across socio-economic categories.

7.6 Matching the Index to the Application – Example of an Index for Policy Use

Depending on the application, each index requires different properties. For example, what is required of an instrument for policy applications?

An instrument intended for purely academic use may be complex and opaque, however, one intended for policy purposes, such as targeting interventions, should be explicable and defensible to a wider constituency. It should also have:

- ◆ *Technical robustness.* It should be based on established analytical techniques and evidence.
- ◆ *Transparency.* In general, the index should be simple to understand.
- ◆ *Objectivity.* The index should be objective and capable of application to all areas.
- ◆ *Plausibility.* It should be capable of reasoned and unambiguous explanation.
- ◆ *Freedom from perverse incentives.* It should not create financial incentives that appear to conflict with sensible interventions.
- ◆ *Reliability of calculation.* Indexes should use data whose quality is sound, consistent between areas; and available for all areas.
- ◆ *Comprehensibility to non-specialists.* The index, should be capable of commonsense justification to non-specialists.
- ◆ *Durability.* It should not become quickly outdated.
- ◆ *Practicality.* It should be derived — updated in a manageable manner, within the time constraints of the annual financial cycle.

The following characteristics are also highly desirable:

- ◆ *Clarity of contribution of constituent indicators and domains.* It is desirable that the relative significance of individual indicators can be quantified.

- ◆ *Flexibility*. It should be possible for the index to take account of future changes of responsibilities or structure (e.g. reorganisation or boundary changes).
- ◆ *Stability*. Fluctuations in the index arising from fluctuations in data for component indicators should be well founded, rather than a side-effect of limitations in the quality of those data.

It is customary for indexes in policy applications to be frequently reviewed for their capacity to generate policy-relevant results. The criterion of *materiality* comes into play when one is considering changes to existing indexes.

Materiality refers to the question of when is it worth introducing an index, or changing an existing index. The extra technical complexity of the proposed change to the index must be set against the impact of the change on the populations, or resources affected by the indexes.

7.7 Pitfalls and Problems of Using Indexes

7.7.1 Difficulties in Using an Indicator may be a Pointer to Design Problems

If there are difficulties using an instrument it may be that the instrument does not meaningfully reflect the reality it is intended to represent. This can happen in the case of health status questionnaires whose questions seem irrelevant to respondents. It can also happen when classifications oversimplify or misrepresent a phenomenon. In both cases, the questionnaire or index would be invalid for the intended purpose. The rules provided by the OPCS assigning social class to women demonstrate some of these problems ([Box 26](#)).

Box 26

Example – how to assign social class to women

The set of rules used by the OPCS for assigning a woman to a category in the Registrar General's Social Class Scale (SCS) varies according to her formal marital status. Thus:

- ◆ When married and living with a spouse the woman is classified on the basis of her husband.
- ◆ When not living with a spouse and in employment the woman is classified on the basis of her own occupation.
- ◆ When not living with a spouse and unemployed, a variety of solutions are adopted.

The convolutions become impenetrable when attempting to classify the single, never-employed woman, who lives on her own and who does not remember her father's occupation.

The problem is that the SCS has violated two technical requirements for a classification: that it should have a *uniform basis*, and that there should be a unique assignment for each case.

In order to provide a unique assignment for each case, the SCS abandons any pretence of having a uniform basis for classification. A woman may be classified either by her own occupation or by that of any near male relative. This is absurd as well as sexist. The SCS may provide an exhaustive as well as a unique classification, but this is only achieved at the expense of the sole purpose for classification.

7.7.2 The Indicator Becomes the Reality

A common problem is to confuse the index with the phenomenon it purports to measure and, as a result, forget that it is only a proxy or partial measure. If the index is widely accepted to the point where there is little questioning of the content and construction, a number of problems can arise, in particular: reification, circularity and impurity. Most of the examples relate to the Registrar General's SCS but the principles apply widely.

Reification

Reification is a common problem. Although a proliferation of indexes may cause confusion, the domination by a single index may be equally undesirable as its operational definition may start to substitute for the meaning of the concept of which it measures. It is especially prevalent in the measurement of self-reported health status, where questions derived from the SF-36 and its shorter versions are now virtually accepted as the lay conception of health.

The same tendency is experienced with measures of deprivation where it is more common to use phrases such as "the ten most deprived local authorities", rather than "the authorities with the top ten scores on the IMD". Reification is likely to lead to forgetfulness – that what is being measured is not deprivation, but a very elaborate combination of factors chosen by a combination of political, theoretical and pragmatic criteria.

Circularity

Forgetting about the properties of the measure and its origins, can lead to circular forms of thinking ([Box 27](#)).

Purity

The basis of the classification of the indicator should be 'pure'. For example, the use of tenure as an *index of 'social class'* could not be validated by showing that tenure discriminates access to amenities. A household's tenure directly affects its access to amenities, so we cannot tell if there is any effect upon access to amenities due to 'social class' rather than to

Box 27**The circularity of social class**

The indicator proposed should not be *circular*: that is, the evidence produced to justify a causal link should not itself rely on the *explicandum*. For example, a group of people are poor because they are in social class V; they are in social class V because they are poor. Again, the tabulation of educational attainment by social class of head of household shows a typical gradient, but the juxtaposition of observations that: “you are in social class V occupation because you are poorly educated” and “your lack of education is typical of those in social class V,” does not constitute the basis of an explanation. This is not trivial. Stevenson [24] (the originator of the social class scheme) fell into this trap when attempting to validate his proposed scheme. The evidence he gives for the validity of SCS as a reflection of social distinctions is that it discriminates mortality. He then claims that, because SCS is a valid mirror of society, we can go on to examine the social process bringing about differential mortality.

tenure. Stevenson [24] was greatly concerned with this problem and accordingly rejected several other possible alternatives to occupation as the basis for his social class classification. He recognised that it was also a problem when using his own occupation-based classification to discriminate mortality, because some jobs were in themselves dangerous, so that the observed differential mortality on the basis of his occupationally based scheme was a combination of a ‘social class’ effect and occupational risk. This led to his attempt to isolate the ‘pure’ social class effect by looking at the mortality of wives by the occupational group of their husbands. Whilst the problem of circularity is usually relatively easy to avoid and detect, the problem of ‘impurity’, like that of ‘reification’, is much more difficult. On the one hand, detailed and complex measures may extend and confuse the meaning of the core phenomenon, while on the other, over-simple measures, may reduce the intended phenomenon to something trivial and ultimately meaningless.

Testing for purity involves ensuring that indexes have a clear relation to the phenomenon being indicated. Townsend made a similar point,

“It is, we believe, mistaken to treat being a member of an ethnic minority as part of the definition of deprivation. Even if many among this minority are deprived, some are not and the point is to find out how many are deprived rather than operate as if all were in that condition. It is the form their deprivation takes and not their status which has to be measured” [211].

The measure becomes the explanation

One consequence of reification, especially when linked to circularity, is a confusion between the index and an explanation of the phenomenon being explored.

For example, in many reports, SCS is used not only as an *index* to portray differences but that portrayal is also assumed to constitute an *explanation* of those differences [212]. There are, of course, many reports that do not make that assumption. The Black Report on Inequalities in health [213] is exemplary in this respect, offering four possible explanations for an observed SCS distribution. But, they have difficulty in maintaining the distinction between those four explanations and ‘social classes’ and the political consequence was that the then hostile government found it easy to ignore their findings (these difficulties were cited by Patrick Jenkin, Secretary of State for Social Services, 1980, as one of the reasons for not endorsing their recommendations), which is a high price to pay for methodological sophistication.

7.7.3 What does a ‘quality of life’ index mean?

Reification may occur because many multiple domain indicators are so complex that it is near impossible to tell what they are actually measuring, and it is tediously repetitive to list all the components every time an index is referred to. The most complex indicators often purport to measure deceptively simple, ‘common-sense’ notions, such as deprivation, class, health and, not least, quality of life. If these indexes are used for resource allocation reification can be dangerous. For example, we need to know precisely what interpretation of quality of life has been used in a cost-effectiveness assessment if that assessment concludes that an intervention is not recommended to someone over a certain age. Equally, we need to know what measure of deprivation directs money to one area than another.

The meaning of quality of life measures

The meaning of indexes that claim to represent our own experiences should be subject to special scrutiny, especially if they have policy implications. This may require an investigation of the process of construction of the index, asking how and why critical decisions were made.

Unpacking the meaning of an index is rarely easy – consider the case of a ‘quality of life’ index (see [Section 4.3.10](#)). There are many difficulties with devising an overall quality of life index. In fact, there are two distinct sets of problems: establishing a coherent set of component indicators; and interpreting combinations.

Choice of components

There is no consensus over the components or the weighting procedures to be employed in 'composite' quality of life indexes.

Whilst everyone wants a certain minimum of conditions, few can agree on the optimum level or combination required.

Whilst nodding in the direction of consumer sovereignty for choosing and combining components, few have actually attempted to take that position seriously.

There is the counter argument that each of the components is the product of a gradual process out of which some degree of consensus has emerged. But, that argument also is the foundation for the objection that it is an historical consensus. Whether or not such components or weighting are relevant to different populations is important. There is no consensus as to how relevance ought to be measured, nor differences reconciled. If public perceptions are to be an eventual component of their experienced quality of life, then the relative importance of different aspects of their situation must also be essential.

Problems of interpretation – trade-offs are obscured

There is a very close correlation between life expectancy and per capita income at a macro (national) level. The relationship between income and health is not so simple, e.g. isolation of elderly and certain forms of child abuse prevail more in high income nations.

Etzioni and Lehman [212], argued against 'formalistic-aggregative measurement of collective attributes', as with the U.S. Crime Index, which aggregates a broad range of crimes, thus giving the same weight to a murder and a \$50 theft.

For many applications it may be important to ensure that although the use of indexes has advantages, we should not lose interest in individuals, and a concern with extremes as well as averages; and should not lump everything together which will tend to produce bland results.

The point is that not only is well-being multi-dimensional, its aspects are incommensurable in that although they are inter-related, they are not substitutable for each other. For example, a sufficient income to ensure good nutrition increases life expectancy, but you cannot compensate early deaths with high income. Although an index, through continued use can be presented as being simple, such as GNP, the underlying presumptions are often quite complex and obscured.

Lack of disaggregation

Few quality of life indexes address distributional aspects of the different components of the 'quality of life' or 'well-being' of particular population groups.

This is principally because of the difficulty of collecting sufficient nationally comparable data to yield meaningful estimates at the community level or for small groups; such indexes can usually only be calculated for highly aggregated and often inappropriate geographic units of analysis.⁶

7.8 Conclusions

Although this section has concentrated on what may seem to be relatively obscure and technical aspects of indexes, the questions covered are increasingly important when indexes are being used to inform decisions on resource allocation, targeting and rationing at every level of government.

Although it is unlikely that there will ever be a perfect index for every (or any) application, it is at least worth trying to ensure that the chosen instrument:

- ◆ Is not circular or prone to misinterpretation or reification.
- ◆ Provides a uniform basis for a unique assignment for each case.
- ◆ Can be derived from easily collectable data in a form which corresponds to the underlying phenomenon.
- ◆ Is relatively sensitive to changes or stability in the underlying phenomenon.

It is also worth remembering that there is no such thing as a universally valid or reliable index; that different applications will require different properties, which may be substantive – such as using a transparent index in resource allocation; or technical – such as using an index with sufficient responsiveness to detect the suspected inequality.

Finally, each index comes complete with assumptions and premises, resulting from its methods of construction, the choice of components and any attendant theory. These may not always be obvious, but they will influence the instrument's performance and results. Hence, after any statistical testing, one should always ask "what does an index really mean?"

⁶ This is also because of confidentiality, where data at small area levels or for small groups are 'Barnardised', that is, -1 , 0 , or $+1$ are randomly added to the counts.