Comparative Review and Assessment of Key Health State Measures of the General Population

Patrick Sturgis *London School of Economics*

Roger Thomas *National Centre for Social Research*

Susan Purdon *National Centre for Social Research*

Anne Bridgwood *Office for National Statistics*

Tricia Dodd *Office for National Statistics*

This Research was Funded by the Department of Health, UK.

Table of Contents

1 Introduction

1.1 Purpose of the review

The research project reported here was commissioned by the Department of Health. Its aim was to provide information that would assist in the design of surveys and interpretation of survey results bearing on the *general health* of the population of England, resident in private households. The work was conducted by a team of researchers drawn from the *National Centre for Social Research* and the *Social Survey Division* of the Office for National Statistics. Team members had had experience of conducting the Department's main continuous survey used for monitoring the health of the general population, the Health Survey for England (HSE), and also of carrying out methodological research on the measurement of *general health*.

The project was empirically driven, being based around results from large population surveys in the UK. Several of these had been commissioned by the Department itself and by other governmental bodies, such as the Health Education Authority, over the previous decades. These included the General Household Survey, the Health Survey for England, the Health Education Monitoring Survey and a number of others. Evidence from health surveys conducted outside central government was also considered and special attention was paid to 'calibration' exercises designed to throw light on the way questions about respondents' *general health* are interpreted and answered by members of the public. Several such exercises had been carried out over recent decades by the *National Centre*, by *Social Survey Division* and by academic experts in health measurement.

While the conceptual and technical standards applied in this review are intended to match those of the best academic work, we have been aware throughout, of practical and technical constraints faced by the Department and its research contractors in implementing a health monitoring strategy through surveys of the general population.

1.2 Scope of the review

1.2.1 General and specific health measures

A large number of instruments have been developed to measure health; some of the better-known general measures are the Short Form 36 Health Status Questionnaire or SF-36 (Ware et al. 1992) and its shorter versions, the Short Form 12 (SF-12) and the Health Status Questionnaire 12 (HSQ-12), the Nottingham Health Profile (Hunt et al. 1986) and the Sickness Impact Profile. Examples of dimension-specific measures include the General Health Questionnaire (GHQ) and the Hospital Anxiety and Depression Scale. Among condition-specific measures are those developed to assess specific mental disorders, such as Schedules for Clinical Assessment in Neuropsychiatry (SCAN).

The present project was concerned with measures of health which are generally applicable, cover a range of dimensions of health of importance to the public, and are simple to understand and use in sample surveys of the general population. Condition-specific measures are of limited application for this purpose as it is difficult to use them to compare health across different groups (Brazier et al. 1992), although SCAN, for example, has been developed for use on a general population survey of psychiatric morbidity (Meltzer et al. 1995).

The instruments evaluated for this project are shown in Table 1. Part A of the table shows the measures included in our original list and on which we have carried out further analyses. Part B shows some additional measures that we included as the project progressed. These instruments are discussed in detail in Chapter 3.

1.2.2 Uses of *general health* measures

General health measures can be used for a variety of purposes:

- In a clinical setting, they can be used to measure the impact of disease, the outcome of interventions such as hip replacements, and to assess suitability for preventative or rehabilitative therapy. They are used, for example in the Netherlands, the United States of America and in Spain for assessing eligibility and suitability for residential care.

- They can be used, both at local and national levels, to evaluate health care policy.

- They can be used to construct other measures to produce estimates of Healthy Active Life Expectancy and Quality of Life.

- At a population level, they can be used to produce prevalence estimates and thus provide a method of monitoring the population's health and of assessing the likely demand for health care and services.

When used to produce health measures at a national level, they may be collected either via a *census* or a *sample survey*. The use of the decennial census to collect health information has the advantage of near complete coverage of the whole population and, thus, the ability to provide estimates for small areas of the country. The disadvantage of using a census, apart from the cost, is that space on the census form will be limited, and the forms will be completed by one member of the household, without the presence of an interviewer to probe for a full answer. In addition, there are long periods between censuses, which means that information collected on them will become outdated.

Because of these disadvantages, the monitoring of the nation's health is normally carried out by the inclusion of health measures in continuous or frequently repeated surveys of the general population, using face-to-face interviewing. Therefore the main focus of this report is the measurement of health in the face-to-face sample survey context. However, we have also included in our review the National Censuses for 1991 and 2001, which include general health questions and therefore have an important place in health monitoring strategy.

In addition to reviewing existing results and publications, the project team conducted and reported on some original analyses. These were made possible by the fact that the Department, anticipating this review, had arranged for a number of different measures of general health to be included in the questionnaires for the 1996 round of the Health Survey for England (HSE). This gave scope for the types of cross-analysis needed to compare measures empirically and to explore differences in the way the measures operated when administered to adult members of the general population.

1.3 Content of the Report

There were three main components to the project:

- A review of the conceptual and methodological background, leading to the setting of criteria for assessing measures of general health.
- A systematic review of the literature, and consultation with experts in the field, on the measures under consideration.
- Secondary analysis of existing data sets using the agreed criteria for evaluation as a guideline. The analysis also drew on the results of cognitive question-testing carried out in preparation for the 2001 Census and for the pilot stage of the 1997 Health Education Monitoring Survey (HEMS).

A number of different questions and instruments with some claim to measure *‘general health’* were identified in collaboration with the Department and a list of methodological standards and criteria that such questions should meet were set out. In the chapters that follow our aim, where possible, has been to assess each measure in terms of these criteria. Ideally we would have had available both a ‘gold standard’ criterion for assessing the validity of candidate measures and a comprehensive set of empirical evidence for comparing other aspects of their measurement performance (see section 2.9). In practice this was not always possible and we have often had, therefore, to make the best of rather inadequate data.

1.4 Structure of the report

Chapter 2 of the report reviews some key concepts and methodological issues that arise when seeking to measure general health through questionnaire measures in sample surveys. It also presents the criteria that have been used for the assessments of the existing general health measures made in the subsequent chapters.

The methodological development and existing research literature on the health measures being evaluated is reviewed in *Chapter 3*, while *Chapter 4* presents an empirical comparison of the measures based on data from the Health Survey for England, the General Household Survey and the ONS Omnibus survey. In *Chapter 5* the results of cognitive work carried out on some of the measures are presented. Overall conclusions and recommendations are presented in *Chapter 6.* The data sources used in the study are described in detail in the Appendix.

Table 1 List of Instruments Reviewed

A. ORIGINAL SET OF INSTRUMENTS FOR REVIEW	**UK National surveys on which used**
(Limiting) Long-standing Illness (LLSI) A set of two questions asking the respondent whether he or she suffered from any limiting long-standing illness, disability or infirmity and, if so, from any limitation of activities as a result.	GHS since 1971 HSE since 1993
Activities of Daily Living (ADLs) List of activities routinely performed in daily life presented to respondents to determine whether they can perform each activity and, if so with what degree of difficulty. Developed for studies of the elderly. **Instrumental Activities of Daily Living (IADLs)** Subset of ADLs.	GHS 1996 and other special surveys of the elderly
Health Utilities Index (HUI) Multi-item battery. Never used in major population surveys in the UK, but does offer both a multi-dimensional health state descriptive system and a utility score for each health state it describes.	None
Short Form 36 (SF-36) 36 item battery, with scoring system to provide measures for 8 dimensions of general health. A shortened, 2 dimensional scoring system is also available which gives scores for overall physical and mental health.	HSE 1996
The EuroQol (EQ5D) Five dimensional health state descriptive system and a utility score for each health state. Designed primarily in order to provide a 'Quality of Life' component to health state description	HSE 1996 Omnibus 1996 GHS 1995/6
The General Health Questionnaire (GHQ-12) 12 item battery designed to identify persons in the general population who are likely to be suffering from some form of mental health problem.	ONS Psychiatric Morbidity Survey HSE 1996

B. ADDITIONAL INSTRUMENTS REVIEWED	**UK National surveys on which used**
Modified version of the GHS Limiting Long-standing Illness question (LSI Cen91) Significantly modified version of the LLSI questions used on the GHS, stressing inclusion of conditions "due to old age" and limitations affecting work.	Census 1991
Possible Census 2001 version of the Limiting Long-standing Illness question (LSI Cen91) Question being tested is further modified from 1991 to exclude the term "infirmity" and substitute "health problem". Reference to work limitation remains. Reference to old age deleted.	Census 2001?
Possible Census 2001 version of the Self-Assessed General Health Question (SAGH Cen01) As GHS, with response categories Good/Fairly good/Not good.	Census 2001
Short Form 12 (SF-12) 12 item sub-set of the SF-36, which provides enables scoring of the 2 higher order general health factors (Physical and Mental) only.	None
Self-Assessed General Health (SAGH) A single question asking the respondent to assess his or her overall general health on a fixed response scale. (3 point scale on the GHS, five point scale on HSE)	GHS intermittently since 1977 HSE since 1993
ONS Overall Disability Measure (ONS Dis) Multi-item battery and scoring system developed for the purposes of the OPCS Disability Surveys (1983-84) sponsored by the DSS. Provides cut-off point below which respondents are deemed to be significantly disabled and also a "degree of disability" score. Takes account of multiple disabilities.	ONS Disability Surveys

2. Concepts and Methodological issues

2.1 Why are questions about '*general health*' included in surveys?

Single questions or brief sets of questions about *general health* are frequently included both in specialised health surveys and in general surveys of the population. Three main needs seem to underlie this popularity.

The first need is to control both the burden on respondents and the cost and complexity of surveys by minimising the number of questions on any one topic that has to be included in a questionnaire. A single question providing an indicator of general health is cheap and may appear straightforward to interpret. Simplicity is an important advantage, particularly in the case of large surveys that present and compare results for many sub-samples and over time. If survey respondents are willing and able to answer these simple-seeming questions, why incur more expense and complication by asking more?

The second need, which is implicit in all quantitative survey work, is to derive a simple indicator (or small set of indicators) to subsume the detail which surges to the surface when a person is questioned in depth about something as complex as his or her state of health. Health surveys such as the HSE contain a wealth of measures bearing on specific aspects of health (such as symptoms, diagnosed illnesses, disabilities, accidents, use of health services etc.). These are all valuable to users whose job is to address particular aspects of health and health care. However, there is no obvious rationale for combining detailed indicators for each case, so as to come up with an overall score. Such a score should indicate, in a way which satisfies methodological criteria of validity, reliability etc (see section 2.9), how healthy each respondent is, either in comparison with other respondents, or in relation to some absolute standard (or both).

A third reason for including general health measures in surveys may be as a simple way of estimating the 'burden of ill health' in the population. Here there may be a subtext which defines 'ill health' as 'that which requires input from the health services'. Without such an indicator there is no simple way of using a continuous or repeated health survey to answer to the question 'How well are we (the National Health Service or health services generally) doing in our efforts to improve health in the population?' Of course, the fact that such a measure would be very useful does not, of itself, imply that it is possible to create one that will withstand serious methodological scrutiny.

2.2 The vagueness of the concept of 'general health'

The concept of *'general health'* may at first sight appear rather straightforward and commonsensical. In everyday conversation we often address to a friend or acquaintance such questions as 'How are you these days?'. Sometimes this is mere politeness, but sometimes we actually expect and appear to receive more or less informative answers bearing on the person's *general health*.

However, in asking and answering such questions we seldom give conscious attention to such issues as:

- whether we mean the individual to take account of stable long-term conditions or disabilities, or only of recent or acute episodes of ill health;
- whether they are giving appropriate and consistent weightings to different aspects of health, mental and physical;
- whether or not we expect them to 'discount' for advancing age;

In short, we seldom ask ourselves whether or not the person we are talking to is drawing the same conceptual boundaries around the idea of *'general health'* as we do ourselves, or as other persons to whom we might address the same question. This review tends to expose the implications of such inherent vagueness. Interview respondents probably take formal interview questions more seriously than casual conversational enquiries, but the evidence suggests that terms within the conceptual domain of *'health'* are unlikely to be interpreted very consistently either across different individuals or within the same individual over time.

2.3 General health and disability

Disability and morbidity are related but nonetheless conceptually distinct constructs. This can readily be seen from the facts that some persons with disabilities are otherwise quite healthy and not in need of treatment, while some patients with serious illnesses are not significantly disabled. However, we came increasingly to the view that the attempt to separate consideration of "functional disability" and "general health" is artificial and potentially misleading, for the following reasons.

- The widely used 'Long-standing illness' question specifically use the word "disability" and the follow-up question on limitation of activities is, in effect, a question about functional disability.

- Similarly, many (though of course not all) of the "health" items included in the EuroQol and SF-36 batteries are also about functional disability.

- The measures listed in part A of Table 1 are all implicitly or explicitly measures of *chronic*, rather than acute, ill health and the idea of chronic ill-health is even more closely related to the concept of disability than is the concept of health in general.

- Separating the concepts of ill-health and disability empirically, when eliciting the individual's health experience and health-related quality of life, proves to be very difficult, particularly when dealing with the elderly, who bear by far the heaviest burden of ill-health.

The concept of disability is important in a number of policy areas. It is, for example, of great interest not only to the Department of Health but also to the Department of Social Security, as witness the questions on Activities of Daily Living which have from time to time been inserted in the General Household Survey, even though they tend to have been seen as a specialised tool for studying the health-related problems of the elderly (see Section 3.5).

For the Department of Social Security and local authorities with their responsibility for personal social services, research of prime relevance in this area was the OPCS Surveys of Disability (sponsored by the DSS in the nineteen eighties and since partially updated). Although there is in practice no way in which the full procedures of that type of disability survey could be incorporated into surveys such as the GHS, the HSE or the Census (see section 2.8.8), it might be feasible to include a modified version of the disability screening question that was used. We have, therefore, included in this review measures labelled "disability" as well as measures labelled "general health".

2.4 Health-related quality of life

The measures under review all depend on how people respond to survey questions asking them to *assess their own health* in various ways. Such responses must inherently have a large subjective component, since each respondent applied his or her own standards of judgement to symptoms they experience, health conditions they believe they suffer from, what they believe their health outlook to be and so on (see section 2.1).

Recent debates on the aims of health policy have been much influenced by the concept of *'health-related quality of life'*, which emphasises the subjective view of health and counterbalances clinical concepts and clinical assessment. The focus then shifts from 'Are respondents interpreting questions about *general health* according to the criteria (clinical or other) intended?' to 'How well are we capturing what ordinary people mean by better or worse *general health*?' (see section 2.8.1)

Even if respondents appear to understand consistently what they are being asked to do in providing an assessment of their *general health*, there is no way, without special cognitive studies, that we can assess whether the response given is based on careful and comprehensive thought and the application of 'reasonable' standards of judgement, or not.[1] In other words there can be no ultimate "gold standard" that can be applied to questions inserted in health surveys to distinguish "correct" from "wrong or misleading" responses to questions asking for a self-assessment of health. Ultimately, health-related quality of life is a subjective concept. This again has important implications for the conduct of the review and the drawing of conclusions from it.

[1] Except perhaps in cases of gross discrepancy, such as where persons who on objective evidence appear to be very ill give responses suggesting that they have no health problems.

2.5 Cognitive tasks required by *general health* questions

To arrive at a single, summary answer to a question about his or her *general health* the respondent must, in theory, strike a weighted average of how they feel they stand on different dimensions of health (the weights representing the importance to them, personally, of the different dimensions). In trying to understand what really goes on in respondents' minds as they take in a question about their *general health*, decide what is required by way of an answer and apply standards and judgement to their personal experience in order to produce a response, our main source of information has been to the relatively small amount of "cognitive" question testing work that has been done on the way *general health* questions are answered.

However, recent experimental work in social psychology also offers insights into this issue. There is evidence from the literature that in this kind of situation respondents may resort to "satisficing" (Krosnick, 1999) - that is, giving a response that the interviewer can record while taking short cuts to avoid the demanding intellectual tasks that are really required (Kahhneman and Tversky, 1972). In line with this there is, as we shall see, reason to suspect that respondents often cannot or will not perform the mental gymnastics required to come up with a 'weighted average' of how they stand on all the facets of health. Instead, they take an easier road of deciding whether or not they suffer from any of a limited range of "health problems".

2.6 Capturing the complexity of "health" as a concept

A few moments reflection makes it obvious that 'health' is a complex, many-faceted concept. Indicators of different aspects of health may, perfectly legitimately, move in different directions over time. For example, a person's mobility may improve while their sight (or digestion, or depression, or migraines) get worse; and differential movement of indicators will often be observed at the population as well as at the individual level.

Where the task of reducing everything about their health to a response to a *single question*, while satisfying the measurement criteria to be discussed below, appears to be too much for the respondents, the first or both of two measurement strategies may be tried.

The first strategy (or stage) is to drop the idea of asking respondents to sum up their health status in a *single* response and to substitute a *health descriptive system*. This asks for responses to one or more questions about each of a limited number of *dimensions of health*. The approach, amongst those we have considered that goes most directly for standardisation and simplicity at this stage is the EuroQol instrument. This constructs individual health profiles out of simplified self-assessments made by the respondent on each of five health dimensions. More elaborate profiles are provided by the SF-36 health questionnaire and the first stage of the Canadian Health Utilities Index (HUI).

2.7 Health utility functions and tariffs

The developers of both the EuroQol instrument and the Canadian Health Utilities Index regard this *health descriptive* stage as a prelude to the second stage of constructing a single-number health indicator, as discussed below. However, it is noticeable that users of the EuroQol instrument in the context of health monitoring have tended to use the health descriptive system, but to decide for themselves whether and how to summarise it further.

A *health profile* imposes both simplification and standardisation, but does not in itself satisfy the demand to project an individual's view of his or her health onto just one evaluative dimension, producing a single score on a scale running from "Best health" to "Worst health". To deal with this health economists responsible for developing the health descriptive systems propose a further step, namely, the derivation of a *"health utility function"* (HUF).

The calculation of a HUF essentially requires a model of how human beings evaluate multi-faceted states and that in turn requires special studies to specify and calibrate the function. In these studies a sample drawn from some reference population (say, 'all actual and potential clients of the British NHS') is asked to value a sample of the different health states represented by permutations of scores on the different dimensions of the health state descriptive system. Statistical modelling is then used to derive from the results an algorithm for scoring every possible health state on a *scale of desirability-undesirability* – that is, a HUF.

The developers of the EuroQol profile and of the Canadian Health Utilities Index in this way provide a HUF 'tariff' for each possible health state (as allowed for within the simplified health descriptive systems). The SF-36 does not at this time offer such a tariff, but work is in progress aimed at developing one. If the validity of the method of deriving the tariff is accepted, it is possible to ask each health survey respondent to characterise her or his current health state in terms of the descriptive system and then to read off the valuation attached to that state.[2] This valuation then functions as a one-number measure of that person's *general health*. The developers claim that the health utility function score is equivalent to an interval measurement scale and the theory then leads on to the possibility of valuing changes in health that might be brought about by health interventions.

[2] This, clearly, is not necessarily that particular person's valuation, but the average valuation of that health state for the population as estimated by modelling at the stage of calibrating the model.

2.8 Criteria for assessing measures of general health

Having reviewed the uses to which *general health* measures are put, we found that the list of criteria in terms of which they should ideally be assessed and compared was a rather long and demanding one. However, it is our view that poor performance against any one of the criteria would affect the usability of the measure concerned for monitoring the *general health* of the population.

2.8.1 Does the measure meet the aims of measuring 'general health'?

Preliminary discussions with the Department had raised several different fundamental questions about aims, such as:

- Why are we measuring *'general health'*?
- What are we trying to measure?
- What have we actually been measuring (using existing instruments and questions)?

There is no complete consensus amongst health analysts and researchers regarding the most correct or appropriate answers to these questions. In view of that is was not possible for us to start from a single, agreed, clear and complete definition of *'general health'* and proceed systematically, testing each candidate measure in turn to see how it performed in capturing that defined concept. Instead we had to proceed by first looking at what a candidate measure was said by its originators to measure, or (where such explicit statements were lacking) by inferring intention from the wording of questions and the way the results had been interpreted in practice. Then we looked at the available empirical evidence to see whether the measure seemed to perform as it was intended or assumed to perform, bringing to bear a number of explicit criteria that are itemised and discussed below.

Validity

Validity is the degree to which a measure captures the concepts it is intended to measure and is not *systematically* affected by other, irrelevant variables. Also, the *same* concept needs to be measured in the same way, and using the same standards, for all respondents. There are several different ways of assessing validity.

Face validity appeals to semantic or observational judgements of whether the measure being evaluated appears to capture what it is intended to capture. For example, a question such as 'Do you have any problems in walking about?', with appropriate response categories, might be judged a face-valid measure of 'Personal mobility'. Some measures provide health categories that appear to have an absolute interpretation, while others are interpretable only in relative terms. For example the response category 'Yes' to the question 'Do you suffer from any longstanding illness, disability or infirmity?' has face interpretability. Other types of measure, such as scale scores, are meaningful only by reference to the dimension which they are claimed to measure and the position of an individual on that dimension relative to a population distribution. One might then say that the individual's score fell in the top quintile of scores on a dimension of the SF-36, but one could not specify what responses that individual had given to specific items within the instrument.[3] For some policy uses this lack of 'free-standing interpretability' might be deemed a disadvantage.

Criterion or external validity makes comparisons with other sources. Criterion validity looks for appropriate correlation between the measure being evaluated and some other independent and trusted measure or classification of the same concept. For example, correlation with clinical diagnosis of severe arthritis might be used to validate a questionnaire item about chronic joint pain. Note, however, that it would be very surprising if a measure with face validity as a measure of chronic joint pain did *not* produce *some* degree of positive correlation with diagnosed arthritis. To be convincing as validation of the measure a *very strong positive relationship* would need to be shown.

Construct validity is assessed by testing theory-based predictions of the pattern of statistical relationships between the measure being evaluated and other, conceptually-related measures. For example, a variable said to measure "social isolation" might be predicted to correlate more highly with "living alone" and "mobility problems" than with "digestive problems". Again, it is not enough for the predicted pattern of correlation to be present and statistically significant on large samples. To be convincing the observed differentials need to be quite large.

Predictive validity is assessed by testing theory-based predictions of how health-related outcomes (for example, hospitalisation or death) should vary for cases having different scores on the measure. Once again, almost any measure claiming to detect serious ill-health should be associated with a higher-than-average chance of early death. To provide convincing proof of the validity of the measure the prediction achieved needs to be striking, *even with other variables such as age statistically controlled.*

2.8.3 Freedom from overall bias

The idea of freedom from overall bias is linked to, but not the same as, the idea of 'validity' and also to that of 'sensitivity'. As we shall see, some questionnaire measures give a more 'optimistic' view of the general health of the population than others. Clinical examination of a representative sample of the population would probably show that 'perfect health', like 'very poor health' was relatively rare, though the health defects suffered by many of the population would no doubt be relatively trivial or latent (such as, for example, unfitness and obesity due to lack of exercise or poor diet, which is known to

[3] Similarly, one cannot specify to which particular items in a mathematics test a person with a given overall score gave the correct answer.

be a predictor of serious diseases in middle age). Therefore it could be argued that a questionnaire measure of general health that suggested that as many as half the population had no health defects is either biased in an ‘optimistic’ direction, or, alternatively, that it is insensitive to real differences in health within that part of the population which is free of *major* health problems.

2.8.4 Sensitivity

We are concerned that measures should be sufficiently sensitive to differences in health states. A measure needs to able to detect changes over time, or mean differences between groups, in the aspect of health that it is intended to measure. This sensitivity should ideally be standard over the whole range of the underlying health variable, so that there are neither “ceiling effects” (loss of sensitivity in distinguishing “very good” from “good” health), nor “floor effects” (loss of sensitivity in distinguishing “very poor” from “poor” health).

2.8.5 Freedom from bias between sub-groups

The aim is to monitor the health of all sections of the general household population. It is therefore important that the criteria above apply equally to all subgroups of the population, so that the health of *particular subgroups* is not *spuriously* represented as being better or worse than that of other subgroups.[4] In other words, measures must be *equivalent in their meaning and interpretation* for all members of the population. A degree of *random* variability in how individuals interpret and answer survey questions is tolerable, but *systematic relative bias* in the way questions are interpreted and answered (say) by men versus women, or by younger people versus older people, undermines the aims of monitoring (with a view to determining health policy priorities), unless it can be corrected for in some way.

2.8.6 Reliability

The term *‘reliability’* is here used in the technical sense which distinguishes it from *validity*. Ideally it is assessed by special *test-retest* studies[5] but if such studies have not been carried out, as is often the case, then the *internal*

[4] This implies, of course, that there is some independent criterion measure or measures of level of health.

[5] An example would be if a method of measuring individuals’ height gave results which might vary by several centimetres when two measures were taken one day apart. If a measure is reasonably valid, but subject to random variability, the effect on estimates based on that variable is the same as that of cutting the sample size. The level of reliability can be estimated by mounting special experiments. Given such estimates, the effect in reducing effective sample size can be calculated.

consistency of the measure may be assessed for multi-item scales[6]. The statistical idea that links these two approaches is that a reliable measure is one that is not subject to excessive random variability in the results it obtains at the individual level. The presence of such *measurement variance* has the same effect on survey estimates as a reduction in sample size.

2.8.7 Portability

Measurement instruments used in monitoring must not be prone to relative bias[7] in their application. For flexibility in developing a health monitoring strategy it may be desired to mount a measure on different survey vehicles and to vary the precise questioning context or use of proxy responses.

It is therefore very desirable that a measure should be *portable between surveys* in the sense that it will produce the same results, irrespective of whether it is included on a dedicated health survey or a multi-purpose survey *(freedom from context/order bias).* The ideal measure should also be independent of *mode of administration* (whether the measure is administered by telephone or face-to-face, for example) and use of *proxy responses*. Given the fact that health means in the population tend to change rather slowly and that small changes are therefore of interest, lack of portability in measures may have serious consequences in causing statistical artifacts that may be mistaken for true changes or differences in *general health*.

2.8.8 Practicality

While the criteria already discussed are of prime importance in scientific and methodological terms, a criterion which in practice tends to outweigh them is practicality. In survey contexts this concerns:

a) the length of time it takes to administer and complete the survey questionnaire module concerned and hence the associated operational and opportunity costs;
b) whether the results will slot readily into an existing time series and offer scope for useful comparisons.
c) whether survey respondents seem able and willing to provide answers to the questions involved without any untoward reaction;
d) the cost and complication of processing the resulting data;
e) the suitability and tractability of the results for presentation in descriptive survey reports.

[6] For measures based on scales, internal consistency/reliability is usually assessed by a statistic known as Cronbach's alpha.

[7] The term 'relative bias is used because we do not have a 'gold standard' measure of general health in relation to which *absolute* bias of the regular measuring instruments might be assessed.

The heaviest ‘practicality’ weighting is usually allotted to (a). Application of (b) leads to a very conservative attitude on the part of survey sponsors and users, both as regards abandoning measures that are deemed ‘practical’ but have been show to fail on scientific criteria, and as regards adopting measures which seem at all risky or unfamiliar in terms of (b), (c) or (d).

2.8.9 Stability of measurement performance over time

To fulfil the purpose of monitoring health over time, it is important that the format and wording of the measure used and *the way in which they relate* to what we intend to measure be *invariant over time*.[8] Then one can be confident of interpreting a change over time in the measure as indicating a real change in the population’s health, not a change in expectations or in the relative weight of different components of the measure.

The ideal measure would be one that also provided continuity with available time series, thus extending the current monitoring data rather than having to start a new series. However, as we shall see, some doubts arise over whether the health standards applied by respondents do remain stable over time.

2.9 Other factors which can confuse the interpretation of change

Quite apart from the validity, reliability, sensitivity etc. of the measurement process, other factors can complicate the *interpretation* of measured and statistically significant observed changes in mean health status over time. These include, for example, cohort and period effects, which affect means purely because older generations are constantly being replaced by younger generations, but may be mistaken for the effects of health or other policy interventions.

[8] For example, the aims of monitoring over time are undermined if what respondents in one year understand and intend (on average) by the response “My health is good” is different from what respondents understood and intended by the same response five years earlier.

3 The general health measures under review

3.1 Introduction

This chapter describes each of the measures under review and their use in surveys of the general population. It evaluates each in turn by reference to the criteria outlined in section 2.

3.2 Self-assessed general health

Key features

Widely-used on surveys
Short, single-item measure
Easily administered
A good predictor of mortality
International comparisons possible

Difficult to interpret because it implicitly measures several dimensions of health

3.2.1 Origins, Purpose and Scoring

Questions on self-reported general health are commonly included in surveys, often in conjunction with questions on self-reported long-standing illness and disability[9]. The General Household Survey (GHS)[10] has included a single-item question since 1976 and therefore offers twenty years of annual estimates and the Health Survey for England has included a question since its inception in 1991. The GHS question is currently being tested for possible inclusion in the 2001 Census.The Health Survey for England uses the following question, which is recommended by the WHO Regional Office for Europe, as an instrument for collecting internationally comparable data for measuring progress towards achieving WHO-Europe Health for All targets. Use of this question therefore provides a basis for international comparisons of self-assessed health, although respondents' understanding of what

[9] Surveys which have included a question on self-reported general health include The Health Survey for England, the General Household Survey, the Health and Lifestyles Survey, the Health Education Monitoring Survey, the ONS Psychiatric Morbidity Surveys, the National Child Development Survey, the 1970 British Cohort Survey, the Allied Dunbar National Fitness Survey and the Health Education Authority's Today's Young Adults Survey.

[10] A description of the main data sources described in this chapter is given in Appendix B.

constitutes ‘good’ or ‘bad’ health will be influenced by cultural and historical contexts.

[*] *Now I would like to ask you some questions about your health. How is your health in general? Would you say it was..*
RUNNING PROMPT

1 very good
2 good
3 fair
4 bad
5 or very bad?

The General Household Survey uses the following question which, unlike that used by the HSE, specifies a time period[11].

[*] *Over the last 12 months would you say your health has on the whole been...*

1 good
2 fairly good
3 or not good?

Questions on other surveys ask respondents to compare themselves with others; the question used by the Health and Lifestyles Survey, for example, asked respondents to say how good their health was ‘for someone of your age’.

Although self-assessed health is often measured by a single item, there is widespread evidence that this question nevertheless covers several dimensions of health, and that people implicitly go through a process of considering and weighing these dimensions when answering the question. Respondents to the 1984 Health and Lifestyles Survey, for example, were asked what they understood by the term ‘health’: among the aspects which they mentioned were absence of disease, functional ability, and fitness (both physical fitness and psychological well-being). Also identified were a ‘moral’ dimension, whereby health depended on will-power, self-discipline and self-control; health as healthy behaviour (being a non-smoker or non-drinker, taking exercise); and health as a ‘reserve’ which could be diminished by neglect and accumulated by good behaviour (Blaxter, 1990). Cognitive work carried out for the pilot phase of the 1997 Health Education Monitoring Survey (HEMS) identified very similar themes. Respondents interpreted ‘health in general’ as absence of ill-health, the ability (or not) to lead a normal life, a state of mind, and physical fitness (see section 5). Participants in the 2001 Census question-testing programme also referred to frequency

[11] This question has also been included in the ONS Omnibus survey

of doctor consultations, whether or not people were absent from school or work because of ill-health, and whether or not they were taking medication.

3.2.2 Target populations and use in surveys of the general population

Many questions on self-assessed health were specifically designed for inclusion in surveys of the general population. As single items, they take very little time in an interview or when a respondent is self-completing a questionnaire. There is evidence (Calnan, 1987) that those with higher levels of education are able to produce more elaborated definitions of health; there may be therefore systematic differences between social groups in their understanding of questions and hence in the meaning of their answers. Blaxter (1990) believes that this distinction does not hold when people are encouraged to elaborate on their ideas in an in-depth interview, but warns that respondents do not have the time to do this in most surveys. It may be that less well-educated respondents are more likely to draw on narrower concepts of health in the survey setting.

3.2.3 Validity

Self-assessed general health has been shown by studies in several countries to be a good predictor of mortality. In the UK, a follow-up study to the Health and Lifestyles Survey (HALS2) showed that, after the existence of a serious disease, self-reported poor health was one of the most powerful predictors of mortality. Among those who said in their 1984 interview that they had no serious disease, men at all ages who assessed their general health as ‘fair’ or ‘poor’ were twice as likely as those who rated it as ‘excellent’ or ‘good’ to die in the seven years between the initial and the follow-up study. For women, self-assessment was a good predictor only for those aged 55 or over (Blaxter and Prevost, 1993).

Similar studies in Sweden (Sundquist and Johansson, 1997), the USA (Berkham and Syme, 1979; Idler et al. 1990) and France (Grand et al. 1990) have shown similar results. The Swedish study had a very large sample of almost 40,000 respondents. It found that poor self-reported health status was a significant risk for men and women of all ages, when the effects of age, marital status and low socio-economic status (measured by educational level and tenure status) were controlled for.

The validity of questions on self-reported health has also been tested by comparing them with other measures of health. In an analysis of the 1984 Health and Lifestyle Survey results, Blaxter (1990) constructed a health index based on four dimensions: the presence or absence of disease, the presence or absence of illness (as measured by reported symptoms), fitness and unfitness, and a measure of perceived well-being. The presence or absence of disease was partially validated by nurse assessments and by details of medication reported by respondents. The fitness/unfitness dimension was based on physiological measurements such as Body Mass

Index, blood pressure and respiratory function. Blaxter found a high level of agreement between self-reported general health and the index at the two 'extremes'; that is, those whose measured health was best and worst (as measured on the four dimensions) were most likely to give an appropriate self-assessment.

Self-assessed health has also been shown to be associated with doctor consultation rates, with the mean rates of consultation increasing as self-perceptions of health deteriorate. However, Blaxter (1985) found that, once social class was taken into account, self-assessments and consultation rates were clearly associated only for those belonging to the manual social classes; she suggests that not consulting is part of the definition of being in good health for these groups.

Evidence suggests that there is an overall tendency for respondents to give positive rather than negative assessments of their health, but as with other measures discussed in this report, there are systematic variations between the assessments given by people in different social groups. Evidence from a number of surveys suggests that older people have lower expectations of health, and are more likely to make a positive assessment of their health than a younger person with similar illnesses or symptoms might; they consider themselves healthy despite the difficulties associated with ageing. Similarly, people with a disability can give assessments of their health as good, 'despite the disability'. Those in families where the head of household is defined as belonging to the manual social classes are more likely to make a more pessimistic assessment than objective measures suggest is appropriate (Blaxter, 1990).

People in different social groups also emphasise different dimensions in their definitions of health; functional ability is more likely to be mentioned by older people, and fitness by younger people. Psycho-social well-being is stressed more by people in the middle years, by women and by more highly educated respondents.

3.2.4 Reliability

Data from the 1997 HEMS (Bridgwood et al. 1998) indicate that individual changes in self-rated health are associated with objective changes in health. The 1997 survey was a follow-up, in which respondents who were first interviewed in 1996 were interviewed for a second time in 1997. As well as being asked about their health, they were also asked whether they had experienced one or more of a series of events in the last year. Those who reported a serious illness, injury or operation since their first interview were three to four times as likely as other respondents to give an assessment of their health in 1997 which was more than one category 'poorer' than in 1996[12].

[12] This category includes, for example, those who said their general health was 'very good' in 1996, but 'fair' in 1997, or 'good' in 1996 but 'bad' in 1997.

Blaxter (1990) warns that people are often inconsistent in their assessments of their own general health. One of the reasons for this may lie in the answer categories available. The cognitive work carried out for the 2001 Census and the 1997 HEMS pilot explored respondents' understanding of the different answer categories. The 'fairly good' category in the GHS question and the 'fair' category in the HSE question were least easy to define; 'fairly good' was considered to be a vague term, while 'fair' was seen as an average of good and bad days. Those who described their health as 'fair' in the 1996 HEMS were most likely to have changed their assessment; less than half used the same description in 1997. Similarly, about one in six of those who described their health as 'good' and more than a quarter of men and more than two fifths of women who said it was 'bad' in 1996, opted for 'fair' in 1997. If the term 'fair' is difficult to define clearly, than it is perhaps not surprising that some respondents change their assessments over time. Similarly, some respondents had difficulty distinguishing between the 'very good' and 'good' categories in the HSE question; some movement between these two categories is therefore perhaps to be expected.

3.2.5 Ease of interpretation

Responses to questions on self-reported general health offer a simple summary measure with an intuitively comprehensible meaning, which can be used to compare different social and health status groups. They give an overall summary assessment of health, although it is difficult to know whether any differences in reported health for a given population over time are real differences or a difference in the relative weight attached to the component dimensions of health, particularly as these dimensions are implicit rather than explicit. When analysing differences between social groups, it should be borne in mind that there are systematic differences in the dimensions which respondents have in mind when making an assessment of their own health, and in the extent to which these assessments correlate with more objective measures.

3.3 Long-standing and limiting long-standing illness questions

Key features

Widely used on surveys
Short
Easily administered
High levels of agreement with doctors' assessments
International comparisons possible

Difficult to interpret trends in prevalence
Sensitive to changes in question wording
Underestimates prevalence among elderly people, and other social groups

3.3.1 Origins, Purpose and Scoring

With an ageing population, it can be argued that the management of chronic illness and disability is an increasing challenge for the health service. Accurate information on the prevalence of long-standing conditions is therefore of great importance. Questions on self-reported long-standing illness and disability and limiting long-standing illness are commonly included in surveys of the general population, often in conjunction with questions on self-reported general health[13]. Questions on long-standing illness or disability have been included in the General Household Survey since 1971 (with a separate question on limiting long-standing illness since 1974), with a break in 1977 and 1978, which provides time series data spanning a period of 26 years. A question on limiting long-standing illness was included in the Census for the first time in 1991, in part to obtain an improved indicator of the likely need for health services for small areas than could be produced from survey data.

The Health Survey for England, the General Household Survey, and many other surveys, use the following question:

[*] *Do you have any long-standing illness, disability or infirmity? By long-standing I mean anything that has troubled you over a period of time or that is likely to affect you over a period of time?*

[13] Questions on long-standing illness or disability are included in the Health Survey for England, the General Household Survey, the ONS Psychiatric Morbidity Surveys, the Health and Lifestyles Survey, the Health Education Monitoring Survey, the Survey of the Physical Health of Prisoners, The National Survey of Sexual Attitudes and Lifestyles, the 1991 Census Validation Survey the National Child Development Survey and others.

1 Yes
2 No

If yes

[*] What is the matter with you?

The GHS also asks whether the condition is a limiting one:

[*] Does this illness or disability (Do any of these illnesses or disabilities) limit your activities in any way?

1 Yes
2 No

The 1991 Census used the following question:

Do you have any long-term illness, health problem or handicap which limits your daily activities or the work you can do?
Include problems which are due to old age.

These core questions are sometimes supplemented with further questions on Activities of Daily Living (OPCS, 1994) or by a checklist of symptoms (Health Promotion Trust, 1987).

The methodology for recording details of illnesses on the GHS has changed since the start of the survey. From 1971 to 1976 interviewers tried to establish the nature of the illness or injury and the information was coded according to the International Classification of Diseases (ICD). A simpler version was introduced in 1988; informants were asked about their symptoms rather than about underlying causes because they find it easier to report on these and because, on the whole, it is symptoms which affect people’s day-to-day lives (Foster et al. 1990). Blaxter (1990) found that far more people declared a symptom, such as a painful joint, than a named disease such as arthritis. The information obtained is coded to a list of just over 40 codes which correspond in a rough way to chapter headings in the International Classification of Diseases.

The question asking for details of illness is sometimes asked only as a courtesy with no intention of analysing the responses, as in most years of the GHS; at other times, interviewers are asked to probe the nature of the self-reported illness or disability fully. This was done in 1988, 1989, 1994 and 1996 for the GHS, for all years of the Health Survey for England, for the first Health and Lifestyles Survey and for the Survey of the Physical Health of Prisoners.

The dimensions of health covered by the questions are not explicit, but there is some evidence that they measure physical morbidity more successfully than psychiatric morbidity[14].

Answers to these questions are used to produce estimates for the prevalence of self-reported long-standing and limiting long-standing morbidity among people living in private households. Long time series, such as those produced by the GHS, provide a point of comparison for local, ad hoc or irregular surveys. International comparisons are possible, as other countries use similar questions, although prevalence estimates will be influenced by cultural understandings of illness, disability and normal activities. The data have also been used to produce estimates of Healthy Life Expectancy (Bone et al. 1995b) and combined with other measures, including more objective measures such as blood pressure and lung function, to produce a summary scale of health (Blaxter, 1987).

3.3.2 Target populations and use in surveys of the general population

Questions on long-standing illness and disability are short and easy to administer and therefore take little interview time. They can, however, be sensitive to changes in question wording and to mode of administration. For example, the overall prevalence of limiting long-term illness as measured by the 1991 Census among those resident in private households was 12%, significantly lower than the estimate of 18% from the 1991 GHS. The authors of the 1992 GHS report argue that differences in methodology accounted for some of the difference; the census information was collected by self-completion, usually by one member of the household and related to one night in April, while for the GHS all adult members of the household are interviewed individually by a trained interviewer and fieldwork goes on throughout the year (Thomas et al. 1994). The change in wording to include reference to 'the work you can do' may also have contributed to the discrepancy'. A comparison of responses to the Census question and to an identical question on the 1991 Census Validation Survey (CVS) found a 'gross error', that is the proportion of times the answers to the two studies were different, of 4.9% (Heady et al. 1996). Higher estimates of prevalence were obtained in the Census Validation interview than from completed Census forms. The authors of the CVS report point out, however, that the differences may reflect genuine changes in health between the Census and

[14] Respondents to the 1971 GHS were asked whether they had consulted a doctor in the two weeks before interview, and the condition for which the consultation took place. A comparison of these data with that collected from GP records for the National Morbidity Survey showed a lower proportion of both men and women in the GHS reporting mental disorders. The authors of the 1972 GHS report note that this is not surprising; this type of disorder may be ascribed to a category by a doctor without the patient being told; where informants did know of their condition, they may not wish to confide it to an interviewer. General Household Survey 1972 (London: HMSO) 1975

the survey, or lack of knowledge on the Census form-filler's part about the health of other members of the household. The comparison between the Census and GHS questions, together with several other studies, also show that quite small differences in survey design, question wording and possibly in question order also appear to influence response (OPCS, 1994). In this regard some of the most prominent effects are:

- Surveys which attempt to measure both limiting and non-limiting chronic illness with one question tend to produce lower overall estimates of prevalence than those which ask two separate questions.

- Asking whether respondents 'have' a long-standing illness produces higher estimates than asking whether they 'suffer' from an illness; some people may answer 'no' to the latter on the grounds that they are not actually suffering (Goddard, 1990).

- Asking whether an illness limits activities compared with 'people of your age' produces lower estimates than asking whether it limits them 'in any way'; it is believed that elderly people in particular would say no because most of their contemporaries were as limited in their activities as they were (OPCS, 1975).

- Using a checklist of symptoms stimulates reporting (Blaxter, 1987). One advantage of a checklist is that it provides all informants with a common frame of reference; it is possible, however, that it might produce overestimates of prevalence as informants who are not sure whether they have a condition might include themselves (Goddard, 1990). A checklist cannot be used to produce accurate prevalence estimates for more serious diseases as sufferers are more likely than others to be in hospital or unavailable for interview (Blaxter, 1987).

- Analysis of GHS data suggests that asking informants for full details of their illness before they are asked whether the illness limits their activities might result in lower estimates of *limiting* long-standing illness or disability. The authors of the 1988 report suggest that some informants may be reluctant to say that an illness limits them when interviewers know what it is; they also note, however, that unexplained fluctuations in the levels of self-reported limiting illness were a feature of GHS data throughout the 1980s (Foster et al. 1990).

- Asking interviewers to use directed probes, rather than generalised ones, can result in marginally more codeable conditions being reported.

The cognitive question-testing carried out for the 1997 HEMS pilot found that respondents were able to define the terms 'illness' and 'disability' without difficulty, but that some had difficulty in understanding 'infirmity'. For some respondents, infirmity was synonymous with old age.

3.3.3 Validity

Assessments of the validity of questions on long-standing illness or disability have been based on comparisons with standardised mortality ratios (SMRs), the results of clinical examination and doctors' reports. They show a high level of agreement for overall prevalence, although the level of agreement varies for specific conditions and for different social groups. Commentators note that discrepancies do not necessarily indicate that data from self-reported sources is inaccurate; informants may not have brought a condition to the attention of a doctor, medical records could be inaccurate, doctors may not have informed patients of their diagnosis, and lay descriptions may differ from those given by doctors (White, 1995).

A comparison of age-standardised ratios for overall prevalence of self-reported chronic sickness and standardised mortality ratios carried out for the first GHS in 1971, showed that for males, with the exception of Scotland, regions where SMRs were higher than expected also had higher than expected age-standardised ratios of long-term illness. This was also true for limiting long-standing illness and disability. There was less apparent correspondence between the two measures for females (OPCS, 1975). A similar comparison carried out at local authority level on 1987 Census test data showed correlations of 0.80 for men and 0.82 for women between all-cause mortality (as measured by standardised mortality ratios) and limiting long-standing illness (Charlton et al. 1994).

Interview data from the 1984 Health and Lifestyles Survey yielded an estimate of 30% overall prevalence of self-reported long-standing illness; information collected from respondents during a subsequent nurse session, which included recording details of medication, increased this estimate by only two percentage points (Blaxter, 1987).

Evidence from several sources indicates that these questions underestimate the prevalence of long-standing illness and disability among the elderly; for example, a proportion of informants who reported difficulties with Activities of Daily Living nevertheless say they had no chronic illness or disability. Even when there is no reference to 'people of your age', it appears that elderly people regard limitations in their daily activities, particularly difficulties with eyesight and hearing, as a normal part of growing old, not as evidence of illness or disability (Martin et al. 1988). However, when the data from the 1991 Census Validation Survey were analysed, it was found that the proportion of those with a disability who reported a long-standing condition actually increased with age; the overall underestimation of chronic conditions among the elderly arose because the number who are disabled is much greater among the elderly than other age-groups, so that a slightly smaller proportionate under-recording produces a much larger absolute effect (Heady et al. 1996).

Supplementing the questions on long-standing illness with questions on Activities of Daily Living and on eyesight and hearing, as is done periodically on the GHS, would be one way of improving estimates of prevalence for the elderly, as the estimates from the two different measures could be cross-referenced at the case level.

When comparing self-reported morbidity among different groups in the population, it must also be remembered that some people are more troubled by a certain kind of symptom than others, and that the need to limit activities will depend on what people usually do (Bennett et al. 1996). Informants may also vary in the amount of information they choose to give or in their knowledge of the extent and nature of their ill-health (Blaxter, 1990).

Comparisons have been made for estimates for specific conditions, as well as for overall prevalence. Blaxter (1990) found an 80% agreement between self-reported data and clinical assessments on the presence or absence of specific chronic conditions. The majority of the serious conditions which were reported were treated (and therefore presumed to be medically diagnosed); conditions which were most likely to be untreated were conditions such as varicose veins, migraine, haemorrhoids and ‘back trouble’. Those belonging to a non-manual social class were more ready to declare a chronic condition, even if it was not functionally troublesome or accompanied by symptoms. Informants in manual social classes, particularly men, were likely to say they had a named disease only if it was actually troublesome; this was particularly true for mental disorders. Very few of those with a severe condition said it did not affect their lives (Blaxter, 1990). Analysis of the 1987 Census Test results showed the highest correlations at Local Authority level between named conditions and standardised mortality ratios were for circulatory diseases (Charlton et al. 1994).

The broad coding of self-reported illnesses used by surveys results in some anomalies for the ICD classification; the Health Survey for England, for example, has identified congenital anomalies and ‘neoplasms’ as particular areas of difficulty (White et al. 1993) . A comparison of the prevalence of self-reported cardiovascular (CVD) conditions and doctors’ reports[15] for 1992 Health Survey for England respondents showed substantial levels of agreement for all CVD conditions combined and for most individual ones. Levels of disagreement varied by condition (heart murmurs and ‘other heart trouble’ had the lowest levels of agreement), and, as in Blaxter’s analysis, by respondents’ social characteristics. Levels of agreement between self-reporting and doctors’ reports were lowest for the youngest and oldest respondents, women, those classified as belonging to a manual social class and the least highly qualified. The Health Survey report warns, however, that American studies show that agreement between self-reporting and doctors’ assessments is higher for CVD and related conditions than for most other chronic conditions. In general, higher rates of agreement can be expected for conditions which require ongoing treatment, have commonly-recognised

[15] Questionnaires were sent to the GPs of a sub-sample of respondents to the 1992 Health Survey for England (with the written consent of respondents). Doctors were asked whether their patient had ever been diagnosed as having one or more of a list of cardiovascular conditions (the same conditions which respondents were asked about during their interview). The results were presented in the report of the 1993 survey. See Bennett et al (1995).

names and are salient to informants because they cause discomfort or worry.

A comparison of data on self-reported long-term morbidity and the use of prescribed medicines from the Health Survey for England (Calhoun et al. 1996) and the Survey of the Physical Health of Prisoners (Bridgwood and Malbon, 1995) suggests a broad association between the type of illness reported and the prescribed medicines being taken. It should be noted, however, that not all medicines recorded by the surveys would have been prescribed for long-standing complaints.

3.3.4 Reliability

There is little or no data on how well these questions perform using a test-retest methodology. There is some evidence on reliability, however, from the 1997 HEMS; respondents who reported a serious illness, injury or operation in the life events section of the interview were twice as likely as others to give an assessment of self-reported morbidity which was poorer[16] in 1997 than in 1996 (Bridgwood et al. 1998).

3.3.5 Ease of interpretation

Data from the GHS enable trends over time to be measured; these show year-to-year fluctuations, but the overall trend for both long-standing and for limiting long-standing illness and disability is upwards. Caution needs to be exercised, however, when interpreting changes in the prevalence of self-reported morbidity as changes over time may reflect changes in people’s expectations of health as well as the prevalence or duration of sickness (Bennett et al. 1996).

[16] Respondents who reported no long-standing illness in 1996, but who reported one in 1997, and those who reported a non-limiting illness in 1996, but a limiting one in 1997 were judged to have given a ‘poorer’ assessment in the second interview.

3.4 The General Health Questionnaire 12 (GHQ12)

Key Features

Widely used on surveys
Easily administered
Well-validated
International comparisons possible

Only measures one dimension of health

3.4.1 Origins, Purpose and Scoring

The General Health Questionnaire 12 (GHQ12)[17] was developed in the UK as a short form version of the General Health Questionnaire (GHQ) which is itself an instrument designed to detect cases as opposed to non-cases of psychiatric disorder in both clinical and non-clinical populations (Goldberg, 1972). It is, therefore, a dimension-specific instrument rather than a measure of general health and, as such, may best be considered as a supplementary measure to accompany other indicators of general health rather than a stand-alone measure in itself. It has been used extensively as a screening test in both general population surveys and in clinical populations.

The rationale behind the development of the GHQ is that although there are many different psychiatric disorders, they nevertheless share a common underlying element. Identifying people on the basis of this common element allows a basic distinction to be made between those who are experiencing some kind of psychiatric disturbance and those who are not. Thus people experiencing psychiatric illness can be identified as a generic class, without making reference to the specific nature of their illness.

The GHQ contains a total of 60 items. These were found to be the items from a pool of 140 which discriminated maximally between psychiatric patients and a control group of ‘normals’. The item selection was carried out to enable the instrument to concentrate on identifying the more common underlying aspects of mental illness such as anxiety and depression rather than mental sub-normality personality disorders or psychotic illness such as schizophrenia[18]. The GHQ was not designed to place individuals on a dimension of severity of disturbance but merely to identify cases as opposed to non-cases[19]. The 12 items included in the GHQ12 are those found to be

[17] © David Goldberg 1978

[18] Although it has been found that such symptoms can be identified, but with less accuracy.

[19] It would be possible, however, to adapt the GHQ scoring system to a dimensional structure of severity.

most highly associated with the main general factor from the overall pool of items.

GHQ items consist of statements about behavioural and psychological functioning; the respondent is asked to say how well the statement applies to them now, in comparison to their ‘usual’ behaviour or state of mind. The response alternatives are: ‘not at all’, the ‘same as usual’, ‘rather more than usual’ and ‘much more than usual’.

Scoring can be done by using the Likert method of summated ratings where:

‘Not at all’	=	0
‘Same as Usual’	=	1
‘Rather More than Usual’	=	2
‘Much more than usual’	=	3

However as little is gained in terms of case identification by discriminating between the severity of symptoms, the scale can be scored:

‘Not at all’	=	0
‘Same as Usual’	=	0
‘Rather More than Usual’	=	1
‘Much More than Usual’	=	1[20]

As the GHQ12 items ask respondents to compare their current state with their normal state, it has been pointed out that this instrument is likely to underestimate the incidence of chronic psychiatric illness. However this effect can be attenuated by including short ‘add-on’ questions about, for example, the use of medication or by employing an alternative scoring system (see Goodchild and Duncan-Jones, 1985).

The threshold (in terms of the number of symptoms identified) for determining ‘caseness’ varies according to the aims of the individual study. Clearly, as the threshold increases, so will the number of false negatives and as the threshold decreases, so will the number of false positives. However, for the GHQ12 the threshold for identification of cases usually varies between four and six (Goldberg and Williams, 1988).

3.4.2 Target Populations and use in surveys of the general population

The GHQ12 is a self-completion questionnaire, taking roughly two minutes to complete. It has been used extensively in both primary health-care and in epidemiological studies of the general population. It was included in the Health Survey for England from 1991 to 1995, and on the ONS Survey of Psychiatric Morbidity among homeless people using day centres and night

[20] This also helps to control for ‘middle category users’ of response scales.

shelters (Gill et al. 1996)[21]. It is reported to be generally well received by respondents and seems to have little problem achieving adequate response rates[22]. Item non-response problems may be encountered amongst elderly populations, but this should be considered a generic attribute of self-completion schedules rather than a specific failure of the GHQ12. The elderly may also experience difficulties in self-completing because of poor eyesight or arthritic fingers (Bowling, 1991). It is not known whether this produces any social desirability bias.

3.4.3 Validity

The GHQ has been extensively tested on both clinical and general populations and in many different countries[23] and is generally considered to have very high diagnostic validity. While the GHQ 60 has demonstrated the best predictive validity, Goldberg and Williams (1988) report only slightly lower levels for shorter versions such as the GHQ12. The construct validity of the instrument derives from the fact that the original item pool was taken from existing, validated psychological dysfunction scales and the content validity has been demonstrated through principal components and factor analysis which consistently reveal a large general factor argued to reflect severity of psychological dysfunction.

Numerous studies have shown GHQ scores to be highly correlated with clinical diagnoses of various kinds (e.g. the Present State Examination) although clearly the strength of these associations will depend on the threshold score adopted and the exact nature of the clinical diagnosis being used as the validation criterion. Goldberg and Williams (1988) report that for studies employing the GHQ12, correlation coefficients varied between 0.71 and 0.91 with a median of 0.86. And although it is explicitly designed as a 'current-state' diagnostic tool, it has nevertheless been shown to demonstrate significant predictive validity in terms of future use of mental-health care (Berwick et al, 1987). People with the highest GHQ scores are reported to also have the highest use of health services (Goldberg and Williams, 1988). Physically ill people tend to score highly without exhibiting any mental dysfunction, and thus to be over-represented among false positives (Bowling, 1991).

3.4.4 Reliability

Reliability has been less frequently examined than validity but those studies which have looked at the issue have generally found more than satisfactory

[21] The GHQ30 was used in the 1984 Health and Lifestyle Survey.

22 The response rate to the Health Survey for England 1994 was 70%

23 Bowling A. (1991) reports that it has been translated into at least 38 different languages

results. Satisfactory internal consistency has been demonstrated by both split-half and Cronbach's alpha analyses (Goldberg and Williams, 1988). Although the problem of test-retest reliability is problematic due to the fact that the dysfunctions captured by the GHQ12 are supposed, by definition, to be transitory a study by Goldberg which considered only the scores of people whose clinical diagnoses had remained stable during the testing lag found that the scores remained very stable.

3.5 Activities of Daily Living and Instrumental Activities of Daily Living

Key features

Good indicator for likely need for health care
Short and easily administered

Of little value for non-elderly populations
More testing required of validity and reliability

3.5.1 Origins, purpose and scoring

Activities of Daily Living (ADLs) have been defined as those tasks which people need to be able to perform to survive without help, while Instrumental Activities of Daily Living (IADLs) are those which are necessary for living a more or less normal life without help (Bone, 1995b). Strictly therefore, questions on ADLs and IADLs measure functional ability rather than general health. Only among the very elderly do a significant proportion report any difficulty with ADLs and IADLs; survey data therefore provide a useful indication of the likely need for health care and services among this group. Blaxter (1990) argues that they are popular for inclusion in surveys as the behavioural manifestations of health are easier to measure than health itself.

The ADL Index was developed in 1959 by Katz and colleagues, to be used by trained observers to describe, for clinical purposes, the state of elderly patients. It covers feeding, continence, transferring (to or from a bed or chair), going to the toilet, dressing and bathing, and can be completed in two to three minutes if the observer knows the patient well (Bowling, 1991).

IADLs were developed by Lawton and Brody (1969), and include using the telephone, shopping, food preparation, housekeeping, laundry, travel, responsibility for own medicine and ability to handle finances. ADLs have been used as a predictor of the course of illness, of needs for care, and of functional/socio-biological outcome of chronic disease (Wilkin et al. 1992).

Questions on both ADLs and IADLs have been included in surveys of the elderly living in the community and occasionally used with younger age-groups. Rather than having trained observers rate patients, interviewers ask informants how well they can cope with the different activities. The results have been used to produce estimates of the proportion of people who have difficulty with tasks (Goddard and Savage, 1994), to construct disability (Martin et al. 1988) or dependency scales (Bone, 1995), and to estimate Healthy Active Life Expectancy (Bone et al, 1995) and Quality Adjusted Life Years (QALYs).

The distinctive feature of the clinical ADL Index is that, among the elderly, the items form a hierarchy, so that those able to feed themselves but who are incontinent are usually unable to manage any of the other tasks; individuals recovering from a disabling illness resume the functions in the same order, beginning with feeding and so on (Bone, 1995). Wilkin et al (1992) note, however, that the hierarchy was achieved only by omitting functions such as walking or climbing the stairs, which do not fit the pattern. A natural hierarchy seems to be less well established for IADLs than for ADLs (Bone, 1995). Survey questions on ADLs usually have some hierarchy built into them so that, for example, those who report no difficulty in preparing a cooked meal are not asked whether they can prepare a snack.

When the ADL Index is used in a clinical setting, patients are graded on three-point ordinal scales by observers according to the difficulty they have in performing the six types of activity; 0 represents ‘no difficulty’, 1 ‘with some difficulty’ and 2 ‘unable to do alone’. Scores on individual scales are totalled, all items being treated as equally important (Bowling, 1991). When used in surveys, ADLs are not usually scored and summed to provide an index. Four answer codes are typically used: informants are asked if they find a task ‘not difficult’, ‘quite difficult’, very difficult’ or ‘impossible’. Those who choose the third or fourth answers are then asked whether they need help to perform the task.

3.5.2 Target populations and use in surveys of the general population

ADLs and IADLs were designed primarily for use with the elderly, and have been adapted for surveys of this age-group in the general population[24]. The 1991 Census Validation Survey addressed questions on ADLs to all age-groups. Hunt (1978), in her study of the elderly, reports that the ‘younger’ elderly, who were experiencing no functional difficulties, sometimes queried the relevance of some of the questions to them, but were happy to answer them when the interviewer explained that their answers were needed for

[24] Examples of surveys using questions on ADLs and IADLs include Social Welfare for the Elderly, The Elderly at Home, the General Household Survey 1980, 1985, 1991 and 1994, the OPCS Disability Surveys, and the 1991 Census Validation Survey. The Census Validation Survey did not restrict the questions to elderly informants.

comparative purposes. When used in non-elderly populations, ADLs are highly prone to ‘ceiling effects’, with most respondents able to perform all the activities.

Some questions may also reflect gender differences in which tasks men and women usually perform. The 1991 GHS report on the elderly, for example, notes that in 1980 men aged 65 and over were twice as likely as women in the same age-group to say they could not prepare a hot meal for themselves even if they had to. This difference had disappeared by 1991, which the authors suggest was probably due to the increased availability of convenience food (Goddard and Savage, 1994). It is unlikely that the 1980 data measure differences in the physical capabilities of men and women; rather, they reflect the likelihood that a proportion of men in that age-group were unaccustomed to preparing their own meals. Bowling (1991) argues that a limitation of the questions in a clinical setting is that they do not take account of the degree of adaptation to the environment. This may be less of a problem in a survey, as the questions ask for respondents’ own assessments of whether they can manage a task, with or without help.

3.5.3 Validity and reliability

There appears to have been little evaluation of the validity and reliability of questions on ADLs and IADLs. Evidence of construct validity is mainly restricted to work which was carried out in developing the Index; the best evidence relates to its ability to predict outcomes in chronic illness, but Wilkin et al (1992) argue that this is usually measured by fairly gross indicators such as death and admission to hospital. Some testing of inter-interviewer reliability in a clinical setting has been carried out; Katz found discrepancies in 5% of ratings (Bowling, 1991).

3.6 The Short Form 36 (SF-36)

Key features

Widely used on surveys
International comparisons available
Associated with objective measures of health and use of services

More work needed on measuring change and reliability
Difficult to interpret the meaning of scale values
Difficult to obtain an overall or global assessment of health

Dimensions of health covered
Physical functioning
Social functioning
Role limitations (physical)
Role limitations (emotional)
Well-being: mental health
Vitality
Pain
Overall evaluations of health: general health perception

3.6.1 Origins, purpose and scoring

The Short Form 36 (SF-36) was developed in the United States in 1988-9, using items from an earlier Long Form version containing 108 items, which was developed as part of the Health Insurance Experiment conducted by the Rand Corporation and the Medical Outcomes Study (MOS). An anglicised version of the SF-36 was developed for use in the United Kingdom in 1990 by Brazier et al (1992).

The aim of the SF-36 is to measure health as a multi-dimensional concept with items covering a range of important health dimensions, including levels of psychological well-being, physical and social functioning and personal evaluations of general health. It is not designed to produce a 'one number' summary of general health and is therefore a profile rather than an index (McHorney et al 1993). It was designed for use in comparing the outcomes of different methods of delivering medical care (Jenkinson & Wright 1993) and for providing epidemiological data at the general population level. McHorney et al (1993) argue that multi-dimensional assessments of health are important because most patients have multiple, co-existing conditions, often both physical and mental (Meltzer et al 1995).

The 36 items on the Short Form version measure eight dimensions of health (shown in the box above); each dimension being measured by a multi-item scale. An additional item, on health change, is not scored. These dimensions were chosen because they were among the most frequently used in health surveys and because the developers of the instrument felt that they were the most important dimensions to include in order to achieve a reasonably comprehensive coverage of the domains of health (Ware & Donald 1992). The SF-36 has been criticised for not including items on sleep (Jenkinson & Wright 1993), although inclusion of such items on measures like the GHQ12 and the Nottingham Health Profile, has led to criticism that they suggest the presence of morbidity which does not exist (Bowling 1991, Garratt et al 1993) The developers of the instrument have stated that other dimensions considered for inclusion but omitted for the sake of parsimony were: health distress; sexual functioning; and family functioning (Ware et al 1993).

The SF-36 scales are scored by using the Likert method of summated ratings (see Section 3.4.1). For each dimension, item scores are coded, summed and transformed onto a scale from 0 (worst) to 100 (best) (Ware & Donald 1992). Brazier et al (1993) argue that the implicit health state valuations underlying the scoring of responses were not explicitly derived and have yet to be tested.

The SF-36 was not initially designed to provide a single summary measure of health status (Long 1993), but factor-analytically-derived summary measures for physical and mental health have been developed, and an overall summary score can be obtained by using weights derived by regressing a criterion measure on the individual domain scores of the instrument. This approach, though, begs the question of which criterion should be used as the dependent variable in such analyses. Furthermore, Hays et al (1993) warn that empirical weighting methods like these require close scrutiny to determine their worthiness in future Quality of Life (QOL) research. The summary scales are described in more detail in Section 3.6.6.

3.6.2 Target populations and use in surveys of the general population

The SF-36 was designed for use both with general populations and with patient groups (Brazier, Jones & Kind 1993, Dixon, Heaton & Long 1994); it was constructed for self-administration by people aged 14 and over, or alternatively for administration by a trained interviewer in person or by telephone (Ware & Donald 1992). It takes about 10-15 minutes (but up to 20 minutes for older respondents) to complete. The instrument has been tested on and used with many patient groups, but has also been used in surveys of the general population conducted by face-to-face interview (e.g. Health Survey for England 1996), telephone interview (McHorney, Kosinsky & Ware 1994) and by post (Garratt et al 1993, Welsh Health Survey 1996). It is perhaps the most widely used generic health measure currently available (Jenkinson et al 1996). Reported response rates vary from 59% (Welsh Health Survey 1996) to 83% (Brazier et al 1992). Several of the early surveys in the UK were undertaken to test the reliability and validity of the measure; Jenkinson et al (1993a) provide normative data for adults of working age, while the Welsh Health Survey provides baseline data against which changes in health may be compared.

There has been very little assessment of the effect of the mode of administration on response distributions. One study in the US found that missing data was more common on a postal than on a telephone survey, and that those who completed by post also had lower mean scores on the different scales (indicating worse health). After adjusting for socio-demographic differences between mail and telephone respondents, mean differences were reduced on the four scales which best measure physical health, but remained the same or increased on the three scales which best measure mental health (McHorney, Kosinsky & Ware 1994). Weinberger et al (1994) compared telephone and face-to-face administration among 31 elderly outpatients who were currently being prescribed at least 5 medications. Telephone administration required significantly less time than face-to-face interviews (10.2 v 14 min). In terms of distributions on the eight dimensions, although systematic differences between modes were generally small, there were significant differences for all 8 scales. These were greatest for emotional role functioning, physical role functioning, social functioning and bodily pain.

As with any self-completion instrument, the suitability and relevance of the SF-36 for different groups in the population varies. The extent of missing data is also positively associated with increasing age for the scales on pain, role limitations due to physical problems and role limitations due to emotional problems (Brazier et al 1992). Questions phrased in terms of work and activities have been criticised as possibly irrelevant or insensitive for elderly people (Hill and Harries, 1993). Brazier et al (1996) found that the SF-36 had the highest rates of missing data in a head to head comparison with the EuroQol and the OPCS disability scale when all three were administered to a sample of elderly women. Floor effects (an insensitivity to high levels of morbidity) were not particularly apparent for the SF-36 in this study apart from on the role functioning dimensions. The authors provide evidence that the SF-36 is the most sensitive of these three instruments for detecting low levels of morbidity.

For people whose first language is not English, successful administration may depend on how well the concepts underlying the items can be translated. Anderson et al (1993) describe how independent item translation was carried out on the Nottingham Health Profile (NHP) by bilingual experts from community medicine, sociology and different medical specialities, who attempted to replicate the underlying concept in the original. Their translations were graded by panels, and were also back-translated. International versions of the SF-36 are being developed. For the Welsh Health Survey, the anglicised version of the questionnaire was translated into Welsh, then back-translated by two independent translators, following procedures recommended by the International Quality of Life (IQOL) Assessment project. Apart from the Welsh Health Survey, this initial review of the literature has not found any reports of attempts to translate the SF-36 into minority languages in the UK.

3.6.3 Validity

In the initial stages of its development, the SF-36 was validated against the original measure from which it was drawn - the full length 108-item Medical Outcomes Study scale. While the SF-36 performed well in this respect, this is an internal form of validation, which does not indicate how well the SF-36 performs against other, more objective measures of health. McHorney et al (1993), using both components analysis and clinical assessments on data from the general population and four chronic condition groups, found that different scales were most valid for detecting different conditions. The physical functioning scale was shown to be most valid in distinguishing between levels of severity for the four chronic medical conditions (hypertension, diabetes, congestive heart failure, and myocardial infarction), while the mental health index, then the role-emotional and social functioning scales proved most valid for the presence or absence of a psychiatric condition (McHorney, Ware & Raczak 1993).

Stansfeld et al (1997) report findings from a longitudinal survey of health which found that the social functioning dimension was associated with social contacts, total satisfaction and total management scores on the Social Adjustment Schedule and with social isolation and emotional reactions on the Nottingham Health Profile. The mental health dimension was associated with marriage, social contacts, leisure scores on the Social Adjustment Scale and emotional reactions, energy level and social isolation on the Nottingham Health Profile. The physical functioning dimension was most strongly associated with physical abilities and pain on the Nottingham Health Profile. They conclude that this pattern of association provides evidence of construct validity for the SF-36. Similar results were found using the summary scales for physical and mental health in the Welsh Health Study, although it should be noted the mental health problems in this survey were self-reported rather than diagnosed by a doctor. Bowling (1997) lists other studies which have shown similar patterns of results: different treatment groups exhibiting distinctive SF-36 profiles and different domains of the profile being most sensitive to particular conditions. Garratt et al (1994) and Ruta (1994) report data showing that the SF-36 is sensitive in detecting changes over time in health state.

Bowling (1997) reports that the Mental Health dimension "has a particularly impressive validity" and cites a study by Davies et al (1988) which found correlations between this dimension and the full Mental Health Inventory in the range of 0.92 - 0.95. McCabe et al (1996) compared the performance of the mental health dimension of the SF-36 and the GHQ12 on a sample of GP patients. They conclude that the mental health dimension performed at least as well as the GHQ12 in terms of its internal consistency and its ability to distinguish between relevant treatment groups in the sample.

The SF-36 has been shown to be capable of detecting low levels of ill-health which were not detected by the Nottingham Health Profile (Brazier et al 1992). 'Floor' effects, that is, a high proportion of respondents getting zero scores (very poor health), have been noted for the role functioning scale (Anderson et al 1993). Ware and Donald (1992) suggest that it might be desirable to add supplementary questions, for example about ADLs, to represent the low extreme of the continuum; an attempt to do this as part of a Feasibility Study for the Welsh Health Survey, however, made the interview appear repetitive to respondents, who felt they were being asked the same questions several times (Bridgwood 1993).

Several studies have found associations between scores on the SF-36 and use of health services. Brazier et al (1992) found significant differences (at the 95% level of confidence) in scores on the physical functioning, social functioning and pain scales between respondents who had seen a GP in the last two weeks or visited an outpatients department in the last three months and those who had not. Garratt et al (1993) found that patients in three condition groups (low back pain, suspected peptic ulcer and varicose veins) who had been referred to a specialist had lower scores (poorer health) on all eight scales than those with the same conditions who had not been referred. There were no significant differences between patients with menorrhagia who had been referred and those who had not, or between the condition group and the general population.

3.6.4 Reliability

Ruta et al (1993) say it is difficult to assess whether inconsistent results over a 2-4 week period are due to measurement error of the SF-36 or true changes in the health of the population. Dixon et al (1994) argue that it is important to know that any changes reflect real differences in health and are not due to imprecision of individual measurements; independent assessment of any changes in health are therefore needed so these respondents can be excluded. This review has so far found no studies which have done this.

The SF-36 performs well on internal consistency measures (Garratt et al 1993; Dixon et al 1994; Anderson et al 1993). Ware et al (1993) review studies which have examined the reliability of the SF-36. For the eight dimensions, Cronbach's Alpha varied from 0.62 to 0.94[25] and test-retest coefficients were all between 0.43 and 0.90. McHorney et al (1994) examined the internal reliability of each of the eight scales across socio-demographic and patient groups. The lowest coefficient was 0.65 for the general health scale amongst patients with psychiatric and complicated medical diseases. Jenkinson et al (1996) suggest that this raises concerns about reliability of measurement amongst patient groups with serious illnesses. Brazier et al (1992) found that the mean of differences from measurement at time one and time two after a two-week interval did not exceed one point on the 100-point scale, which they argue is clinically insignificant. The SF-36 Manual includes estimates of sample sizes needed to detect changes at different levels of magnitude. The sample size needed to detect prespecified levels of change varies across dimensions.

3.6.5 Ease of interpretation

Jenkinson (1993b) argues that scales do not lend themselves to easy interpretation, especially when the items are differentially weighted, as in the SF-36. This is borne out by analysis carried out by McHorney (1993). She found the physical functioning and mental health scales to be relatively pure, and their interpretation unequivocal. When observed differences are found on these scales, interpretation attributed to physical or mental causes can be

[25] Nunally (1978) suggests 0.70 as the threshold for acceptability.

made with a high degree of confidence. Observed differences on the role physical scale can be interpreted as role disability associated largely, but not entirely, with physical health effects. Interpretation may be more complicated when psychiatric conditions are present. Differences in role emotional scores can be interpreted as role disability associated with mental health problems. The social functioning scale is most sensitive to social disability associated with mental health problems but is also moderately sensitive to burden of physical health problems too. Therefore interpretation of this scale is complex. The pain scale is most valid in group discriminations involving patients with back pain and arthritis. The general health scale is the most sensitive to physical health problems.

Julious et al (1995) question the common practice of presenting SF-36 data in the form of scale means due to the non-normal distributions of the scales and the impact that outlying cases have on the mean. Furthermore, there is confusion over what statistic should be used to detect significant differences in health across population sub-groups on the SF-36 dimensions. Ruta (1995) criticises Ziebland (1995) for using the effect size statistic for detecting aggregate change in dimension scores, arguing that the standard hypothesis testing of differences between means with confidence intervals is the appropriate method. Ziebland (ibid.) argued that the SF-36 was inappropriate for measuring population health change as, using an hypothetical example, even if social class V improved to the level of health of social class I, the effect size statistic would not register this as a significant change. Jenkinson et al (1996) have proposed an alternative way of presenting SF-36 data. As SF-36 dimension scores are not normally distributed across population sub-groups such as social classes, they suggest comparing the bottom quartile of scores from each sub-group. Using this approach would yield significant effect sizes for the original example provided by Ziebland.

3.6.6 Summary Dimensions: Physical and Mental Component Scores

The developers of the SF-36 have produced a manual and test data sets which provide scoring algorithms for two summary dimensions of the SF-36. The two summary dimensions are argued to represent 'physical' and 'mental' health, referred to as PCS36 and MCS36 respectively. The rationale for the development of these summary scores is that factor analyses of the eight SF-36 dimensions provide a robust two factor solution with the 'physical' dimensions of the SF-36 (physical functioning; role - physical; bodily pain; general health) correlating highest with one factor and the 'mental' dimensions (vitality; social functioning; role-emotional; mental health) correlating highest with the other factor.

The benefits of reducing the SF-36 from eight to just two dimensions lies in the greater ease of interpretation that the two component scores provide and in the consequent reduction in the chances of detecting spurious differences in health state when evaluating health outcomes and differences in population sub-group health (Jenkinson et al 1997). Against this, however, must be balanced the loss of information entailed in moving from an eight to a two dimensional health state description. Ware et al (1994) argue that because the two components together capture between 80 and 85 percent of the variance on the full eight dimensions of the SF-36, little is lost in terms of explaining changes and variations in health.

Ware et al (1994) report that studies examining the internal reliability of the PCS-36 found coefficients ranging between 0.92 and 0.94 and between 0.87 and 0.89 for the MCS-36. This represents a more than adequate level of reliability. Few studies have examined the test-retest reliability of the PCS-36 and MCS-36 although Brazier et al (1992) report correlations for a two week gap between administration at time one and time two of 0.89 for the PCS and 0.80 for the MCS.

In terms of validity, Ware et al (1994) report that "the PCS and MCS performed in the 80-100 percent range relative to the best SF-36 scale in empirical tests of validity". Using Medical Outcomes Study data, they found that differences in health between patient and socio-demographic groups which were detected by one or more of the eight SF-36 dimensions were very rarely missed by either of the two summary scales.

In the UK, Jenkinson et al (1997) have scored the PCS and MCS using UK population norms from the Oxford Healthy Life Survey. They found very similar results to those reported by the developers in the US. Internal reliability of the scales was high, with Cronbach's alpha coefficients ranging from 0.91 to 0.94 across a number of population sub-groups (sex; those reporting a long-standing illness; and those who had seen a doctor in the previous two weeks). Differences in health state across population sub-groups that were detected by the eight dimensions of the SF-36 were also detected by the two summary scales although the magnitude of change registered by the summary scales was generally smaller. They conclude that their results may indicate that:

"the PCS and MCS are less sensitive to change than individual dimensions of the SF-36. However, because the PCS and MCS have substantially more levels than the domains, and consequently greater precision, it is likely that meaningful change may be reflected in smaller effect sizes".

While evidence supporting the validity and reliability of the PCS-36 and MCS-36 is growing, the literature on these two summary scales nevertheless remains sparse. Much of what has been published has been produced by the developers of the instrument. However, these two summary measures certainly represent a potentially powerful addition to the utility of the SF-36.

3.6.7 Development of a utility score for the SF-36

The SF-36 currently has no utility based score by which different health states defined by the eight dimensional profile can be valued against one another. Such scores are derived through separate ‘valuation surveys’ which assign each health state described by the instrument a utility score on the basis of preference judgments made by the survey respondents (see sections 3.7 and 3.8 for descriptions of the utility scores of the Health Utilities Index and the EuroQol). However, although no such score is currently available for the SF-36, researchers at the School of Health and Related Research (ScHARR) at the University of Sheffield have obtained funding to undertake a valuation exercise on a representative sample of the general population. The National Centre for Social Research has been contracted to undertake the fieldwork and interviewing for the main stage of the survey is scheduled to begin in June 1998. The final report containing utility values for the SF-36 is scheduled for publication in April 1999. The utility values produced from this valuation exercise can be attached retrospectively to SF-36 health state descriptions (i.e. utility values may be attached to SF-36 data which was collected prior to the development of the utility values).

3.6.8 The Short Form - 12 (SF-12)

A short, 12-item version of the SF-36 called the SF-12 has been produced by the developers of the SF-36. It is suitable for self-completion and for interviewer administration and takes only 2-3 minutes to complete (Ware et al 1994), thereby significantly reducing the burden on respondents relative to the full SF-36. Another 12 item version of the SF-36, the Health States Questionnaire (HSQ-12) has also been developed by a group of researchers at the Health Outcomes Institute in Bloomington, USA. Both contain very similar items and perform similarly in terms of validity and reliability. The SF-12, would appear to be the more commonly used of the two and is the more frequently cited in the literature. Discussion here is therefore limited to the SF-12. The UK version of the SF-12 is slightly different to the US version. The wording for several items has been slightly altered for the UK context and for the social role item, there are six as opposed to five response alternatives in the UK version.

All 12 items in the SF-12 come from the SF-36. They were derived by regressing PCS36 and MCS36 scores on to the 36 individual SF-36 items. The SF-12 was then created by selecting those 12 items which best captured the variance in the two summary measures (see Ware et al (1995) for a detailed description of the scoring of the SF-12). The SF-12, therefore, only produces scores for the PCS and the MCS and not for the eight dimensions of the SF-36[26]. The two summary scales produced by the SF-12 are called the PCS12 and the MCS12 to distinguish them from the scales produced from the SF-36.

[26] Although, as the SF-12 contains at least one item from each of the eight dimensions of the SF-36 so that single item scores can be used as a measure of health status on each dimension, the developers of the SF-12 advise against this.

Ware et al (1995) report that the SF-12 summary scales reproduce the PCS36 and MCS36 with at least 90% accuracy. Comparisons of descriptive statistics broken down by population sub-group also reveal the near equivalence of the PCS and MCS when derived from the SF-36 and SF-12. It should be noted that, while the SF-12 captures around 90% of the information supplied by the PCS-36 and MCS-36, as these latter dimensions capture around 80-85% of the information provided by the full eight dimensions of the SF-36, the PCS-12 and MCS-12 should be regarded as being able to capture only around 70% of the information of the full eight dimensional SF-36. Nevertheless, Ware et al (1996) report that in 12 separate studies which compared the criterion validity of the full SF-36, the PCS36, MCS36, PCS12 and MCS12, the SF-12 component scores rarely failed to detect differences between treatment groups where a difference was detected by the SF-36, the PCS-36 or the MCS-36.

The fact that the majority of empirical comparisons of the SF-12 with the SF-36 physical and mental component scores come from studies in which the SF-12 scores were derived from the full SF-36 raises concerns that the SF-12 may produce different distributions when administered as a 'stand-alone' instrument than when it is embedded within the SF-36 (i.e. that it may be prone to context effects). However, Jenkinson et al (1997) have shown, using a split-half sample design, that the scores from the stand-alone version of the SF-12 are near identical to the scores produced when the SF-12 is administered embedded within the SF-36. They also found that the SF-12 summary scales were able to discriminate accurately between different treatment groups. This study used UK data to derive the PCS-12 and MCS-12.

While the SF-12 summary scales will always be a less precise measure of health status than those produced by the SF-36 because of the fact that they are comprised of single rather than multi-item measures of each dimension of health, the SF-12 nevertheless represents a useful compromise between minimising respondent burden and achieving comprehensiveness and precision of health measurement.

3.7 Health Utilities Index

Key Features

Easily administered
Produces a single utility score
Evidence of reliability and validity

No UK data
Complex and controversial method of deriving utility score
Not able to detect change due to ‘ceiling effect’

Dimensions (8)
Vision
Hearing
Speech
Ambulation
Dexterity
Emotion
Cognition
Pain

3.7.1 Origins Purpose and Scoring

The Health Utilities Index (HUI) was developed by a Canadian team of researchers based at McMaster University. It contains both descriptive and evaluative components of health status. The first version - **HUI Mark I** was based on an early multi-attribute general health measure developed by Bush et al (1972) called the Quality of Well-Being Scale. It consists of a four-attribute descriptive system, with each health attribute (physical function, role function, social-emotional function and health problems) containing four to eight levels of severity which together describe 960 different possible health states. It was developed to assess the effect on later health of low birthweight. Following work by Cadman et al (1986) on what a sample of parents and their children considered to be the most important aspects of health, a further three attributes were added to form the **HUI Mark II** which was applied in several clinical studies (Feeny et al 1994). The seven dimensions of the HUI Mark II are: sensation; mobility; emotion; cognition; self-care; pain; fertility. This system describes 24 000 unique health states.

In the latest stage of development, the HUI Mark II has been modified to make it more suitable for administration in general population surveys. This has involved removing the fertility dimension, replacing self-care with dexterity and sub-dividing sensation into vision, hearing and speech. This gives a total of eight attributes, each containing either five or six levels of severity and producing a total of 972 000 different possible health states (see box above). Each of these health states is theoretically possible as the attributes are structurally independent. That is, it is possible to be on any given level of one attribute irrespective of the scores on the other attributes. In practice, however, many of the possible health states will have zero or near zero frequencies. This is the most recent version of the instrument - the **HUI Mark III**. The dimensions of health covered by the HUI III are deliberately limited to what the developers refer to as 'within the skin' dimensions of health. This means that there are no dimensions covering social role limitations. The instrument explicitly focuses on functional *capacity* as opposed to *performance* in an attempt to cancel out the effects of differential proclivities to indulge in the activities described by each attribute. The levels within each health attribute are not interval scales but have ordinal properties.

Scoring of the HUI III is done in two stages. Firstly the respondent is placed on the level of each attribute which he or she believes best describes their health for that attribute. This provides a descriptive profile of health across eight dimensions which can be studied at both the individual and the aggregate level. Secondly, by deriving preference-based scoring functions from separate studies for each of the possible health states, a utility index can be incorporated. This gives a single summary measure of quality of life (ranging from 0 = 'worst possible health' to 1 = 'best possible health') for the health state of each individual in the sample. The HUI III is similar, then, to the EuroQol (see Section 3.8) in that it provides both a descriptive profile and a preference based index of quality of life for each profile described. In the same way as the EuroQol, the HUI III index score is argued to have interval properties and is thus suitable as an outcome measure in clinical trials, as the quality component in quality adjusted life years (QALYs) and for assessing the costs and benefits of different medical interventions.

The technique used to obtain a single index preference score from the multi-attribute profile is based on von Neumann-Morgenstern's multi-attribute utility theory (MAUT) (See Torrance et al 1995). This is a technique which provides utility scores for the complete range of health states without having to get respondents to value full descriptions of multi-dimensional health profiles. Instead, due to the hierarchical nature of the levels within each attribute, utility scores can be derived by modelling values provided for individual levels within attributes relative to other levels within the same attribute and of the best and worst states within each attribute relative to the best and worst states within other attributes. This is a great benefit as, due to the fact that people are only able to value health states comprised of limited numbers of attributes at the same time, the number of attributes that can be used in the descriptive part of the instrument can be massively increased. For example, the EuroQol technique which requires respondents to value 'whole' health state descriptions, describes 243 different possible health states, while the HUI III which uses MAUT can describe 972 000.

The basic approach used is that a separate utility is calculated per person for each of the eight attributes. This is done by getting respondents to place each level of each attribute on a visual analogue scale of 'preference/dispreference'. By this method each respondent's opinions of which is the worst level within each attribute is elicited, giving what is termed the 'corner' state within each attribute for all respondents. Next, respondents are required to provide 'time trade-off' preferences for each of the 'corner' states, assuming perfect health on all other attributes. When this is complete you then have, per person, the utility of single attribute health states *relative to the best and worst health states in the list* (from the visual analogue results) and the *relative* utilities of the worst health states for each of the eight attributes (from the time trade-off). To get utilities for *combinations* of health states these utilities are combined into a single expression (or model) per individual. From here various modelling procedures are used to determine utility scores for multi-attribute health profiles (see Feeny et al (1996) for a complete description of the models used to derive utility scores). At present utility scores are not available for the HUI III and the developers urge users to collect data which allows scoring of both the HUI II and HUI III. When utility scores for the HUI III have been adequately modelled and tested, these can be applied retrospectively. Until this point, HUI II can be used if utility scores are required. The developers report that work is currently under way to provide utilities for the HUI III descriptive system.

3.7.2 Target populations and use in surveys of the general population

The Health Utilities Index Mark III is suitable for use in both clinical and general populations (but not for the production of utility scores) and has been used in self-completion and interviewer-administered formats in both child and adult versions. There is no 'one' questionnaire format and several different versions have been used in different contexts (Furlong et al, 1996). Both 'current' and 'usual' health of respondents can be asked about. The mean time taken to complete the HUI III in the 1994/5 National Population Health Survey (NPHS)[27] was 2.01 minutes. However, it is possible that in certain formats the questionnaire could take longer to complete for those in worse health. This is because of the hierarchical nature of the levels within each attribute. Respondents at level one on every attribute would have to answer only 8 questions, while respondents at the lowest level of health on each attribute would have to answer 44.[28]

Reported response rates have generally been high (more than 75%) (Feeny et al 1994), with the interviewer-administered versions achieving slightly higher response, as would be expected. Reported item nonresponse rates are also low, with only 0.5% of respondents to the '94/95 NPHS failing to complete items on the HUI III. It has been used in several large-scale general population studies, including the 1990 Ontario Health Survey, the 1991 Canadian General Social Survey and the 1994/5 NPHS mentioned above. It is currently available in English and French Canadian and the developers report that translations have been made into: French; Dutch; Spanish American; Swedish; Norwegian; and Japanese.

3.7.3 Validity

There is a question over the degree of longitudinal sensitivity that the HUI III is capable of attaining in surveys of the general population. In the 1990 Ontario Health Survey, four of the eight attributes had over 95% of respondents at Level 1 (representing best possible health on the attribute) and over 85% of respondents achieving utility index scores of higher than 0.8 (0 = worst health, 1 = best health) (Berthelot, Roberge & Wolfson 1993). Such scores clearly raise the issue of 'ceiling effects' - that is to say that because baseline population scores are so high on a number of attributes in

[27] Multi-stage stratified cluster sampling design with an issued sample of 23 200 households using face-to-face Computer Assisted Interviewing.

[28] This would only apply to questionnaire formats in which the interviewer starts at level one on each attribute and asks the respondent if that category applies to them, moving on to the next level if it doesn't and moving on to the next attribute if it does.

this instrument, there is little scope for detecting population, or indeed individual improvement in health

The content validity of the HUI III has, like other general health measures, relied on the careful selection of the appropriate attributes of health and of the levels within attributes to ensure content validity. However, as the primary aim of the HUI III was to derive utility scores, the number of different dimensions of health that it was possible to include was in part limited by the need for the system to produce brief, understandable health state descriptions. In addition to the lack of any dimension covering social functioning mentioned earlier, the developers also note the absence of mental health and of general health dimensions (although it might be considered that the utility score for each health state is effectively a measure of general health even if it is not a judgement made by the respondent about their own health). Furthermore, while the developers give detailed accounts of the development of the content domain for the first two incarnations of the instrument, they are less explicit about the methodology used to decide how or why to make the changes from version II to version III. Given that the first two instruments were designed more specifically for use with clinical populations (survivors of neo-natal intensive care) and were derived from interviews with mothers and children from that target group and that fairly significant changes were made from HUI II to HUI III this may raise concerns over the coverage and relevance of the dimensions selected to the general population.

In terms of criterion validity, there appears to be little published work in this area. However, Torrance et al (1995) report that the validity of the modelled utility scores was validated against a limited number of health states that had been directly valued (i.e. whole health states, rather than single level descriptions were presented to respondents to be valued) and that the results were "encouraging" - the standard deviation of the prediction error being equal to 0.058. The error of the prediction was also random, meaning that the modelling neither consistently undervalued nor overvalued the direct valuations.

It has been shown that HUI III scores vary amongst normal population sub-groups in expected ways. For example utility index scores have been found to vary with sex, income level, educational level, socio-economic status and geographic region in the expected directions (Feeny et al 1994). However, while the utility scores did vary in the expected directions in these studies, the actual magnitude of the differences between groups was often small. For example, Pluschauskas (1992-4) showed in the 1990 Ontario Health Survey that, though there was the familiar correlation between income and health, it is not quite of the gradient we might expect and raises questions over the sensitivity of the instrument. As Fenney et al (1994) comment:

“Additional analyses are... required to establish both the discriminative and evaluative validity of the... system and its preference scoring function. In addition, further analyses are required to develop an understanding of the

magnitude of difference in average utility scores that is meaningful for policy purposes."

Further studies have shown that the HUI III can discriminate between different clinical groups and between clinical groups and the normal population (Torrance et al 1995). Franks et al (1996) have shown that the utility scores of the HUI III used as a measure of current health is a good predictor of future health outcomes.

3.7.4 Reliability

A test-retest reliability study on data from the 1991 Canadian General Social Survey (Boyle et al 1995) found consistently high reliability coefficients for each of the eight attributes, although the exact level varied across attributes. For the index scores, the intra-class correlation coefficient was 0.77. This represents substantial reliability in the context of clinical trials but is only in the mid range when compared to other functional health status instruments (McDowell & Newell 1987). Accurate estimation of reliability of this instrument is hindered by the fact that on certain attributes, the vast majority of respondents place themselves on Level 1 (no problems on the attribute). This test-retest study, however pertains only to the reliability of the descriptive part of the instrument. In terms of the valuation procedure, Torrance et al (1995) report that the valuation survey used to derive the utility scores for HUI II was repeated on a group of parents of childhood cancer patients. Virtually identical valuations were obtained, indicating that the valuation data can be considered of adequate reliability. In terms of mode of administration, evidence reported in the same study indicates that whether the questionnaire is self-completion or interviewer-administered has little effect on the distribution of attribute and index scores.

3.7.5 Ease of Interpretation

With eight attributes, each containing five or six levels, the HUI III profile scores are not easy to interpret. In addition, because the attribute scores have only ordinal properties, it is not particularly meaningful to report average scores on the dimensions. However, in contrast to the SF-36 which has been criticised because of the lack of meaning of the scores on each of the scaled dimensions (see Jenkinson 1996), the HUI III has the advantage that a full semantic description of a respondent's health state can be obtained from their profile score. Two people scoring 3 on the pain dimension of the HUI III will have answered the pain dimension questions identically while two people scoring 30 on the pain dimension of the SF-36 may have answered the pain dimension questions of the SF-36 very differently. However, in practice interest is usually focused on relative differences in health state which favours the scaling approach of instruments like the SF-36 over the less psychometrically sophisticated, Guttman type scales of the HUI III.

The obvious advantage of the HUI III in terms of interpretability is the utility score. This provides a single summary measure of health ranging from 0 (worst health) to 1 (normal health) and is argued to have interval properties. Like all instruments that provide single summary measures of health, however, the validity of the index measure is far from universally accepted. Furthermore, the method used for determining the utility value for a particular health state is very computationally intense and involves a number of difficult transformations and the particular adaptation of multi-attribute utility theory that the authors describe relies on a number of controversial and largely untested assumptions**.**

3.7.6 Vontinuity with currently available UK results

The HUI III is an instrument that has been developed and used almost entirely in the North American context, particularly in Canada. The literature search undertaken for this review failed to identify a single study which reported use of the HUI III in the UK. Even if there are some studies soon to be reported on or studies that were missed in the literature search (MedLine and PsychLit were used) it is nevertheless safe to conclude that on the grounds of continuity, the HUI III does not score highly.

3.8 The EuroQol (EQ5D)

Key features

Relatively easy to interpret
Overall summary index score
Construct validity supported empirically
Discriminant validity and sensitivity poor compared with e.g. SF-36

Dimensions (5)
Mobility
Self-care
Usual activities
Pain/discomfort
Anxiety/depression

3.8.1 Origins purpose and Scoring

The EuroQol (‘European Quality of Life’) health status questionnaire has been explicitly developed as a health related quality of life measure to provide descriptive information about the health state of populations and individuals across a range of important dimensions of health. Primary amongst the aims of the international group of researchers who form the 'EuroQol group' is that the instrument should provide a cardinal index of health for use in economic evaluation of health care interventions. The group first met in 1987 and the instrument, as described in this chapter, has undergone several changes and revisions during the ten years or so since it was first conceived.

EuroQol is a questionnaire-based measurement approach specifically designed to provide descriptions and valuations of a universe of possible health states and enable them to be compared in a systematic, quantitative way.

The EuroQol descriptive system is subjectively-based, in the sense that it relies on judgements made by individuals in response to questions about their competence to carry out physical and mental functions of daily life and also about whether they experience pain, depression or anxiety. The judgements (responses) are constrained within highly simplified and schematic descriptive structures built into the standardised questionnaire.

Over and above systematic description, a major aim of the EuroQol group has been to value health states relative to one another by assigning utility scores[29] to them. This involves placing the universe of health states defined by the descriptive system on to a single scale of 'utility' (or relative desirability). The utility score can be used as a measure of the quality of life associated with each health state that the descriptive system defines. The scale produced by the range of utility values is argued to have interval measurement properties and can therefore provide the 'quality' measure in 'Quality Adjusted Life Years' (QALY) and other such applications. The aim of providing a utility tariff applicable to all health states and to base it on valuation methods accepted in the health economic literature has been central to the development of the EuroQol instrument.

3.8.2 The health state descriptive system

Description is achieved through a system which identifies five different attributes / dimensions of health, listed above.

For each of these areas, three levels of functioning are identified, namely (in broad terms):

1 No problems
2 Some problems
3 Severe problems

Different health states can then be characterised as sets of five digits - e.g. 11111 ('No problems in any of the five areas'), 21213 ('Some problems with mobility, no problem with self-care, some problems with carrying on normal activities, no pain or discomfort, severe anxiety or depression') and so on.

Even though it is so highly simplified, this system yields 3x3x3x3x3 or 243 different health state descriptions or profiles. In practice some of these occur very rarely if at all (e.g. any state involving severe problems with self-care combined with no mobility problems). On the other hand, states involving no problem on any dimension, or some problems on only one dimension, are common.

In addition to rating aspects of their health in this way, respondents are asked to rate their health *overall* using a 'thermometer' or Visual Analogue Scale (VAS). This is explicitly calibrated with scores running from 0 'Worst imaginable health state' to 100 'Best imaginable health state'.

The VAS measure depends on the assumptions that respondents can grasp the idea of an equal-interval scale running from 0 (Worst imaginable health state) to 100 (Best imaginable health state) and can assess and locate their own overall health in terms of that scale.

[29] For a detailed discussion of utility values as applied to health states, see Torrance et al (1994)

3.8.3 Derivation of a single generic health utility index

The VAS score provides a single measure of assessed personal health on a continuous scale which can be considered equivalent to the self-reported general health question (see section 3.2) in terms of the dimensions of health it covers. However, the originators of EuroQol do not consider that its metric properties are adequate for a measure of the 'utility' (or 'relative desirability') of the person's current health state.

They therefore prefer to derive such a measure, where required, from the responses to the five questions on aspects of health. Combinations of responses are converted into scores using the results of modelling carried out on data from separate studies. This, it is claimed, yields a utility measure with interval scale properties.

When it is intended to derive a single health state utility score using EuroQol, it is necessary also to obtain respondents' ratings of 'being dead' as a health state, since scores corresponding to 'being dead' and 'being in perfect health' are needed as anchor and calibration points. There are several ways of obtaining such ratings and this has been a major technical issue in the development and use of a EuroQol utility tariff for health states.

The key interval scale properties claimed for the EuroQol health state utility score are: that the distance, in terms of utility, between any two health states can be read off; and that improvements or deteriorations in health, as reflected in movement between health states, can be compared. Thus, for example, the mean improvement resulting from a specific psychiatric intervention designed to ameliorate anxiety or depression can be compared with the mean improvement resulting from some specific surgical intervention such as hip replacement (presumably having its direct effect on mobility and pain).

The studies through which utility tariffs are derived involve obtaining a sample of subjects whose valuations are deemed, in the aggregate, to be definitive. In the case of the calibration survey mounted in the UK (Dolan, Gudex, Kind & Williams 1995) the view taken was that the subjects should be an unweighted random sample of the adult population of Great Britain, on the grounds that all members of that population were actual or potential patients of the NHS and should therefore have an equal voice in determining the relative utility values attributed to health states.[30]

Using members of the representative sample as respondents, judgements are elicited about samples of the 243 possible health states which exist within the EuroQol descriptive system. This is achieved by having the respondents participate in trade-off tasks, such as 'Standard Gamble' or

[30] It is, however, possible to argue an alternative case that the calibration population should, for example, consist of professionals knowledgeable (albeit through observation rather than subjective experience) about a wide range of different health states (which members of the general population generally are not).

'Time Trade-off', in which they value particular health states relative to the two standard states of 'having no health problems' and of 'being dead'. Each respondent can typically manage only 10-15 health state valuations before becoming bored or fatigued, but different samples of health states can be valued by different (representative) subsamples of respondents. At the analysis stage values for the remainder of the 243 health states are then estimated using statistical models.

The data analysis depends heavily on modelling, because preferences were directly elicited for only 45 health states. Modeling is required partly in order to counteract the effects of statistical 'noise' (random fluctuations) in the data, arising from the fact that respondents do not always make consistent judgements, perhaps because they find it hard to understand or accept the conventions of the trade-off 'games'. Also, different respondents do not in general agree exactly as regards the valuations which they assign to particular health states; sampling variation about the mean valuation estimated for each health state therefore has to be taken into account.

Once a tariff giving acceptable reference scores for each state is available, studies which do not themselves involve the complex modelling process can interpolate a utility value associated with each state reported by their respondents. Use of utility scores derived in this way involves accepting that the reference population from whose responses the utility scores have been derived is appropriate, and also taking on trust not only the adequacy of the EuroQol health state descriptive system, but also the validity and reliability of the responses upon which it is based, and the results of the calibration and modelling procedures, which remain somewhat controversial.

The results of the GB calibration study tended to confirm the prediction that the utility measure derived using model-based scores differs in some of its metrical properties from the crude 'valuation' derived directly from the VAS scores. The idea of deriving an estimated utility score from VAS scores using a mathematical function has not so far provided any convincing results.

3.8.4 Target populations and use in surveys of the general population

There is a standard two-page EuroQol questionnaire, designed for self-completion or interviewer administration, which embodies the descriptive and measurement system. The brevity, simplicity and robustness needed for successful routine administration of the questionnaire as a postal survey instrument have had high priority in its development. There is evidence that high response rates can regularly be achieved in interview surveys where individuals are asked to assess their own health (see Brazier, J, Jones, N, and Kind, P 1993), but there may be more difficulties, both in terms of rate of response and in terms of data quality, where the instrument is used in self-administration (e.g. postal) mode.

It is clearly practicable to administer the EuroQol questionnaire as an element in a large scale health survey. In 1996, for example, it was included in the Health Survey for England and as a module in the ONS Omnibus Survey. It is more concise in terms of the number of separate response items required than something like the SF-36, for example.

3.8.5 Validity

As a descriptive health profile measure EuroQol appears crude when compared with, for example, the eight-dimensional, more finely calibrated profile which the SF-36 and the Health Utilities Index (HUI) provide. This is largely a consequence of the small number of independent health states that the EuroQol instrument defines in order to keep the task of deriving the utility index within manageable proportions. Thus when EuroQol was run on the 1995 ONS Omnibus survey using an equal-probability random sample of the British population, 60% of respondents were classified to the '11111' (best health) group and 90% were classified to the 12 most common health states (Brazier, Jones & Kind 1993). In another general population study already cited (Long 1993) 95% of respondents placed themselves in one of the ten most common health states. This suggests what other parts of the latter study confirmed, namely, that the EuroQol descriptive system is relatively coarse and is subject to important 'ceiling effects' (i.e. it has problems in detecting low levels of ill-health).

On the other hand, we should not assume that greater technical elaboration of profiles automatically confers greater criterion, discriminative or predictive validity or greater measurement reliability. For example, in a random sample of nearly 2000 patients from GP lists in Sheffield Brazier, Jones & Kind (1993) found that EuroQol scores differed between population sub-groups in predicted ways and also that they were significantly associated with scores on the SF-36. However, significant ceiling effects were again found for the five EuroQol dimensions. Essink-Bot et al (1997) also provide support for the construct validity and sensitivity of the EuroQol, though they too found that it was not as sensitive at discriminating between patient groups as the SF-36. Brazier et al (1996) also provide evidence that the EuroQol is not as sensitive to low levels of morbidity as the SF-36. However, set against this was the higher completion rate on the EuroQol, leading the authors to conclude that the EuroQol may be more suitable than the SF-36 for administration in elderly populations due to the reduced response burden.

3.8.6 Reliability

As part of the GB EuroQol calibration exercise carried out in 1993 and referred to above, a test-retest check was carried out using a subsample of 221 of the original respondents. All parts of the original interview were repeated exactly, including the five-dimensional and VAS assessments of own health, as well as the special exercise to derive utility values for a sample of other health states. The levels of test-retest reliability with respect to the Time Trade Off exercise found were deemed to be satisfactory by the EuroQol development team (Kind 1995). However, test-retest values for the 'own health' items seem not yet to have been reported. Brazier et al (1996) also report adequate levels of test-retest reliability when the EuroQol was administered to a sample of elderly female population. As the EuroQol is not a scaled measure of health like the SF-36, tests of internal reliability are not appropriate.

3.8.7 Ease of Interpretation

Compared to other multi-dimensional health measures the EuroQol is, on the face of it, fairly easy to interpret. Because there are only three levels within each dimension of health and only five dimensions in total, presentation of EuroQol health profile data is fairly unproblematic. Frequency distributions of the twenty or so most common health states give an easily understandable breakdown at the population sub-group level. In addition, the three levels on each dimension can be broken down into binary variables indicating the presence or absence of 'problems' on each dimension. The VAS thermometer scores can be presented as means or medians depending on the skewness of the distributions, as can the TTO utility values. As with other single number indices of health related quality of life, however, it is not entirely clear what being at a particular point on the scale 'means' in terms of how healthy / unhealthy an individual is or what constitutes a significant (in the psychological rather than the statistical sense) difference between two scores.

4 Empirical comparisons

4.1 Introduction

Together, ONS and the National Centre hold a number of large datasets containing the health measures covered by this review. The surveys on which the various instruments have recently appeared and which form the basis of this chapter are described in detail in Appendix B. This data presents us with an opportunity to empirically examine the validity and reliability of the estimates produced by these instruments. The Health Utilities Index has not been included on any of the surveys on which data is held and therefore does not feature in this chapter.

It should be noted that the data provided by these surveys, although extensive, still allows only a limited and somewhat piecemeal assessment of the relative validities of the health measures covered by this report. This is primarily because, due to the limitations on space in these surveys, it is rare for more than one or two measures to be included on any one survey at the same time. This means that opportunities for assessing the performance of the instruments relative to one another are rather limited. For example, only one dataset contains both the SF-36 and the EuroQol[31] while none contain both the SF-36 and the GHQ12. It has not been possible, therefore, to examine the inter-relationships between all the instruments in as systematic a manner as would be desirable.

Furthermore, because we are dealing with surveys of the general population, we have little or no objective data on the health status of individuals in the sample. Thus, we must rely on self-reported measures of health status to evaluate other self-reported measures of health status, a circularity which it is hard to avoid when using general population survey data. The Health Survey for England is unusual in that it does contain objective (i.e. not self-reported) measurements of health[32], but these measures by themselves cannot be viewed as very good indices of *general* health. In fact, it is exactly because of the inadequacy of these objective and often domain specific measures for obtaining a more general measure of health that there is a perceived need to include generic health measures in these surveys in the first place.

Despite these problems, the data at our disposal does permit some interesting analyses which shed light on the validity and usefulness of these instruments for use in general population surveys. In Section 4.2 the issue of context effects is examined. This pertains to the portability of instruments between surveys and the extent to which they are affected by changes in factors unconnected to the variables they aim to measure. Section 4.3 looks at associations between the instruments and self-reported use of health services (such as G.P consultations). In Section 4.4 correlations between

[31] And even in this case, the VAS part of the EuroQol is not included.

[32] Nurses are used to take measurements of height, weight, lung function, blood pressure and blood samples.

each of the instruments and age and sex are examined. As increments in age are strongly associated with decrements in health, we would expect all the instruments to show a strong relationship with increasing age. In Section 4.5 we examine the patterns of association between instruments that have been used on the same survey to see whether they are correlated in directions that we would expect theoretically. This is a form of construct validation. Section 4.6 investigates the extent to which the instruments are affected by 'floor' and 'ceiling' effects. This is the extent to which they are (in)sensitive to improvements / decrements in low levels of ill health and high levels of ill health respectively. Finally, in Section 4.7 the issue of criterion validity is explored by assessing the extent to which the subjective health measures are able to predict scores on more objective measures of health such as lung function and blood pressure. All Tables referred to in this section are contained in Appendix A which begins at page 89.

4.2 Context effects

There are a number of reasons why questions which aim to measure the same concept produce different estimates for the same population; even a relatively small difference in the wording of the question or of the response categories, as on the self-assessed health questions on the HSE and GHS, can have a significant effect. Consistency of results across surveys cannot, however, be guaranteed, even if identical questions are used, because of the context in which the questions are asked. There is a substantial body of methodological and survey literature demonstrating such context effects for a wide range of different types of questions. Secondary analysis of the HSE, GHS and Omnibus surveys provides evidence of the scale of context effects for three of the general health measures under consideration: self-assessed general health, long-standing illness and limiting long-standing illness.

Identical questions on self-assessed general health (see section 3.2) were asked in 1996 on both the ONS Omnibus survey and on the GHS. Although there was no difference between the two surveys in the proportions of men and women reporting good health, there were slightly higher proportions of men and women reporting fairly good health on the Omnibus survey; 33% of men on the Omnibus reported fairly good health, compared with 30% on the GHS; the comparable figures for women were 38% and 35%. Such differences may appear small, but with large samples they are statistically significant and are of an order that is certainly important to policy and other users of the results. The between-survey difference was not consistent across the age groups for men, although for women a small (non-significant) difference was apparent in each of the age groups over 25 (Table 4.1). It is difficult to identify the reasons why this difference may have occurred; it may be that the content of the rest of the questionnaire has had an influence on the answers; this is discussed further below.

Self-reported long-standing illness (see section 3.3)was asked using the same question on the HSE, the GHS and the Omnibus. Compared with the 1995 HSE and the 1996 GHS, the 1996 Omnibus recorded lower proportions of both men and women with a chronic condition (Table 4.2). The HSE recorded the highest proportions, although among women the difference between the HSE and the GHS was not statistically significant. Authors of the HSE reports have suggested that respondents to a health survey may be more likely than those participating in a general survey to report an illness (White et al 1993), but this does not explain the lower proportions recorded on the Omnibus survey compared with the GHS. In 1996, on both the Omnibus and the GHS, the questions on long-standing illness were immediately after the EuroQol questions, so the immediate context of the questions was the same for both surveys.

Given the differences between the 1996 Omnibus and GHS in the proportions reporting a long-standing illness, it is not surprising to see that the GHS records a higher proportion of adults with a limiting long-standing illness than the Omnibus (Table 4.3). This was the case for both men and women and, in general, in each of the age bands. The question on limiting long-standing illness was not asked on the HSE.

Thus, it can be seen that identical questions do not produce identical estimates – although any differences tend to be small. Differences could emerge for a number of reasons; if, for example, the surveys had differing approaches to the taking of proxy information, or if they were affected by different types of non-response bias. On the three surveys analysed, however, questions on health would not be answered by proxy as they are opinion questions, and, in general, all three have similar characteristics of non-response (younger adults tend to be under-represented). It is quite possible, therefore, that the observed variation may occur because of the context in which the questions are asked. It might be expected that there would be a difference between answers to questions asked on a general survey, and those asked on a specific health survey, but there was also a difference between the two general surveys, the GHS and the Omnibus. Despite both of these surveys covering several different substantive topics, they are quite different in their actual content. The GHS carries relatively long question modules on major aspects of a person’s life, such as housing tenure, education and employment, while the Omnibus carries a selection of much shorter modules that could be on a wide variety of topics. It may be that the latter survey does not encourage as much consideration of health issues before the answer is given, but this can only be speculation.

4.3 Service use

One way of validating health measures is to examine how they relate to use of health services. In this section results from the 1996 GHS are used to show the relationship between the health measures included in that survey, and whether or not a doctor had been consulted in the two weeks prior to interview. There is, of course, no reason to expect that all those reporting a health problem will have consulted a doctor recently, particularly if the health problem is of a long-standing nature, but the proportions of those who have consulted do give some indication of the validity of the measure.

Table 4.4 shows that 22% of men and 30% of women with a long-standing illness or disability had consulted a doctor in the two weeks before interview, while slightly higher proportions of those with a limiting long-standing illness had done so. A better predictor of doctor's consultation (though not necessarily ill health) appears to be the question on self-assessed general health. Around a third (35%) of men and two-fifths (42%) of women who said that their health in the last 12 months had not been good, had consulted a doctor in the previous two weeks. A fifth (19%) of men and a quarter (24%) of women who said that their health had been fairly good had consulted a doctor, while only 9% of men and 14% of women who reported good health had done so.

Table 4.4 also shows the five EuroQol dimensions and it can be seen that, for each of them, those who reported some or severe problems were more likely to have consulted a doctor in the previous two weeks than those who reported no problems. For men, around a quarter to a third (26-35%) of those who reported problems on at least one of the dimensions had consulted a doctor (compared with 10-14% of those reporting no problems). For women, slightly higher proportions had consulted a doctor, both among those with problems and those without.

Thus, for these three general health measures (all asked on the GHS in 1996), the expected relationship between poor health and doctor consultations was observed. However, of all three measures, the presence of a long-standing illness or disability showed the weakest association.

The ability of health measures to predict use of health services is clearly important from a policy perspective. An instrument which discriminates well between those likely and those unlikely to use health services would clearly be of benefit for planning for future demand. It should be borne in mind, however, that the associations discussed here are not really predictive relationships in this sense. The use of services reported here refers to GP consultations *prior* to completion of the general health measures. It is equally likely that consulting a doctor affects how one subsequently rates one's general health rather than causality running in the opposite direction. In order to assess the ability of general health measures to *predict* future service use, a longitudinal design would be necessary.

4.4 Distributions by age and sex

This section examines the relationship between the measures under review and age and sex. Age is very strongly associated with health. The prevalence of virtually all forms of morbidity increases with age. The only major departure from this trend is in the area of mental health, where evidence suggests that there is no major decline in health with increasing age. There is some evidence to suggest that there is, in fact a minor effect in the opposite direction - mental health improving slightly as people get older. We would therefore expect measures which are sensitive to differences in general health to exhibit distributional gradients across age groups.

4.4.1 Self-reported general health

A question on self-reported general health is asked on all five surveys. The Health Survey for England uses the five-point scale (very good, good, fair, bad, very bad) recommended by the World Health Organization[33], while the GHS and Omnibus surveys use a three-point scale (good, fairly good, not good).

A direct comparison between the prevalence of self-reported good health, as measured by the HSE on the one hand, and the GHS and the Omnibus on the other, is not possible because of these differences in response scale format. 'Good' health is normally derived on the HSE by combining the categories 'very good' and 'good'; this category almost certainly includes some of those who would rate their health as 'fairly good' in response to the GHS or Omnibus question. More than three-quarters of respondents to the 1993 and 1996 HSE, for example, rated their health as 'very good' or 'good', compared with between half and three-fifths of GHS and Omnibus respondents who chose the 'good' option. This in itself shows that how people rate their health depends crucially on how the question is framed. In addition, the GHS and Omnibus question specifies a time period, 'in the last 12 months', while the HSE does not.

All surveys, however, show a similar pattern of association between self-reported general health, age and sex. Men were consistently more likely than women to say that their health was good, although the differences were not significant on the two Health Surveys for England. Similarly, all the surveys showed a strong relationship between self-reported health and age, with the proportion of respondents who reported being in good health declining with age (Table 4.5 – 4.7).

Differences between the proportions of men and women who said they had 'bad' or 'very bad' health on the HSE, or 'not good' health on the Omnibus or GHS, were small and not always statistically significant. Not surprisingly, however, the likelihood of reporting poor health increased with age on all five surveys.

4.4.2 Self-reported long-standing illness, disability or infirmity

All five surveys use the same question to measure self-reported long-standing illness or disability (although the 1991 Census used a significantly different version – see section 3.3). Respondents are asked if they have any long-

[33] See *Second consultation to develop common methods and instruments for health interview surveys*: *report on a WHO meeting, 18-20 Sep 1990*. (World Health Organisation Regional Office for Europe and Netherlands Central Bureau for Statistics: 1990)

standing illness, disability or infirmity which has troubled them for some time. Respondents to the GHS and Omnibus who report such illness are also asked whether it limits their activities in any way.

There were no significant differences in the proportions of men and women reporting a long-standing illness, disability or infirmity on any of the surveys. All surveys showed a clear association between the prevalence of long-standing illness, disability or infirmity and age. Below the age of 55, between a fifth and two-fifths of respondents reported a chronic condition; among those aged 55 and over, between a half and two-thirds said that they had such an illness, disability or infirmity (Table 4.8). The prevalence of long-standing illness as estimated by the HSE was higher than for the GHS or the Omnibus; authors of previous HSE reports have suggested that respondents to a health survey may be more likely than those participating in a general survey to report an illness due to the subject matter of the questionnaire stimulating them to think more closely about all aspects of their health. On those surveys which included a question on limiting long-standing illness, between a tenth and a half of respondents said they had such a condition, the proportion increasing quite steeply with age. On the 1994 GHS, for example, 10% of men and women aged 16-24 said they had a limiting illness, compared with 44% of men and 48% of women aged 75 and over (Table 4.9).

4.4.3 GHQ12

The GHQ12 instrument is a measure of psycho-social well-being consisting of 12 questions, it was included on two surveys: the HSE 1993 and HSE 1995. Answers to each of the 12 questions were converted into a score of 0 or 1, distinguishing those individuals thought to have some form of psychiatric disturbance from those who do not. Individual item scores were then aggregated to form three categories: respondents scoring 0, those with a score of 1-3, and those scoring four or more. A score of four or more is used as a standard threshold for identifying individuals with a psychiatric illness. In 1993, 16% and in 1995, 17% of respondents had a total score of four or more. Women were significantly more likely than men to be above the four point threshold. The lowest proportions scoring four or more were in the 55-74 age-group in both survey years, although the proportions reaching this threshold were quite low in all age-groups (Table 4.10).

4.4.4 The EuroQol Tariff

The EuroQol instrument was included in three of the surveys covered in this report: the Omnibus 1995, the Omnibus 1996 and the HSE 1996. There are five EuroQol dimensions; mobility, self-care, usual activities, pain/discomfort and anxiety/depression. Each dimension is tapped by a single question, using a three-point scale equivalent to; 'no difficulty', 'some difficulty' and 'severe difficulty' (see section 3.8.2). A total or ‘general health’ score can be produced for the EuroQol by taking the associated utility or tariff for each of the individual health state descriptions. The minimum score for the EuroQol utility tariff is -0.072, indicating the utility for the worst health state possible, and the maximum score is 1 - indicating the utility for the best possible health state, with zero being the utility score for death.

Mean scores for the EuroQol tariff decline quite substantially with age. The mean EuroQol tariff score for the youngest age-group (those aged 16-24) on the Omnibus 1996, for example was 0.96 for men and 0.95 for women, compared with 0.72 and 0.71 respectively for those aged 75 and over (Table 4.11).

4.4.5 SF-36

For the total aggregate sample, mean scores on the eight dimensions of the SF-36 for the HSE 1996 were (maximum score = 100):

- physical functioning:	80.9
- role limitations (physical)	80.2
- bodily pain	76.7
- general health	69.2
- vitality	62.7
- social functioning	85.1
- role limitations (emotional)	84.3
- mental health	75.4

With the exception of the mental health scale, scores tend to decline with age. This was particularly noticeable on the physical functioning and role limitations (physical) scales, where the gap between mean scores for the youngest and oldest age-groups was 39 and 31 points respectively. On the mental health scale, those aged 75 and over had the highest scores (indicating better health), but scores on this dimension were fairly similar across all age-groups (Table 4.12).

4.4.6 Activities of Daily Living and Instrumental Activities of Daily Living

Questions on Activities of Daily Living (ADLs) and Instrumental Activities of Daily Living (IADLs) were included on the 1994 GHS and the 1996 Omnibus survey. On the 1994 GHS, they were addressed only to respondents aged 65 and over. For ease of analysis, responses to these questions were used to derive the dependency scale developed by Bone[34] for her work on trends in

[34] Bone, M. (1995b) *Trend in dependency among older people in England*. (London: HMSO)

dependency. Not all of the questions necessary for this scale were included in the 1996 Omnibus survey, so data presented here for the dependency scale are from the 1994 GHS only. Data from the 1996 Omnibus survey was used to calculate the proportion of respondents who could not manage one or more ADL without help.

The dependency scale runs from 1 (independent) to 6 (most dependent). The results from Bone's report are reproduced, and an equivalent for the 1994 data is shown (Table 4.13).

The 1994 data show that a higher proportion of men (94%) than of women (90%) living in private households were classified as 'independent' or 'least dependent' (i.e. a score of 1 or 2)[35]. Analysis by five-year age-groups shows a steady decline by age in the proportion of relatively independent respondents at this level of dependency in all four GHS years. Thus, in 1994 for example, 95% of those aged 65-69 had a score of 1 or 2, compared with 78% of those aged 85 or over (Table 4.14).

Data from the 1994 GHS and the 1996 Omnibus show that the proportion of respondents aged 65 and over who cannot manage one or more ADLs alone increases with age; those aged 75 and over were at least twice as likely as the 64-75 age-group to be unable to manage one or more ADLs alone.

4.5 Associations between instruments

Given that all the instruments under consideration are, in some way or other, measures of ‘general health’, we should expect them to show positive associations with one another. That is to say, if individuals and groups get high scores on one measure of health, they should also get high scores on the others (assuming, of course, that they are scored in the same direction). Furthermore, not only should we expect positive associations between instruments but we should also expect *a priori* to see certain patterns of association between different instruments and between different dimensions on different instruments. For example, it would be surprising if the GHQ12 was most strongly associated with the physical functioning dimension of the SF-36 and only weakly associated with the mental health dimension. If, on the other hand, instruments (or attributes of instruments) intended to measure the same dimension of health were found to be strongly associated with each other but not so strongly with other dimensions of health, this would provide support for the validity of the instruments[36]. This type of validation is known as convergent/discriminant validity (Campbell and Fiske 1959). Where it is unclear exactly which dimensions of health an instrument is covering, examining the patterns of correlation with instruments containing explicitly defined dimensions can provide an indication of what dimensions the former

[35] Data are not broken down by sex in Bone's report.

[36] Of course, it is still possible that neither instrument is actually measuring the underlying construct they are intended to measure.

instrument is actually covering. This is of particular interest for single item, 'global' measures of health such as the self-rated general health item.

4.5.1 The SF-36 and EuroQol

As the SF-36 and EuroQol both take a multi-dimensional approach to health measurement, we should expect to see not only an overall positive relationship between them but also a distinct pattern of correlation between corresponding dimensions on each measure. Table 4.15 below shows Eta coefficients for the comparison of each dimension of the SF-36 with each dimension of the EuroQol. Eta is a measure of association that is used when comparing an independent variable with a limited number of categories and an interval scale dependent variable. Eta can be roughly interpreted as the proportion of variance in one variable explained by differences among groups in another variable (range = 0 - 1).

As can be seen from table 4.15, the overall pattern of Eta coefficients shows that the two instruments are strongly associated with one another. A check on the mean scores shows the pattern of association to be ordinal (i.e. mean values on the SF-36 increase with increasing values on the EuroQol dimensions[37]).

Furthermore, the pattern of correlation between individual dimensions lends support to the construct validity of these two instruments. For example, the explicitly 'physical' dimensions of the EuroQol (mobility, pain/discomfort and self-care) are most strongly associated with the physical dimensions of the SF-36 and least strongly associated with the social functioning and mental health dimensions. The pattern is reversed for the 'mental' EuroQol dimension (anxiety/depression) which shows the highest Eta scores for the 'social/mental' SF-36 dimensions and the lowest values for the more 'physical' dimensions. Each EuroQol dimension has the highest Eta coefficient with the dimension to which it intuitively most closely corresponds (Mobility - Physical function; Pain/discomfort - Bodily pain; Usual activities - Role (physical); Self-care - Physical functioning; Anxiety/depression - Mental health).

4.5.2 Self-reported health and long-standing illness

All five surveys under consideration included questions on self-reported general health and self-reported long-standing illness although, as noted earlier, the wording of the question on general health and the response categories used varied across surveys. All surveys show an association between the two measures, with respondents reporting good health much less likely than those whose health is not good to report a long-standing illness or disability. Thus, for example, only 19% of respondents to the 1994

[37] The scoring of the dimensions on the SF-36 and EuroQol dimensions are actually opposite; higher scores on the SF-36 indicating better health and higher scores on the EuroQol indicating worse health. The positive association referred to takes this into account.

GHS with 'good' health said they had a chronic illness or disability, compared with 86% of those with 'not good' health. Similarly, 97% of respondents to the 1996 HSE with 'bad' or 'very bad' health reported a long-standing complaint, compared with 28% of those whose health was 'very good' or good' (Table 4.16).

While this represents a high degree of congruity between these two instruments, it should be noted that a significant minority of respondents whose self-reported health was ‘good’, nevertheless said they had a chronic illness or disability, suggesting that the two questions are measuring somewhat different aspects of health. The key to this difference probably lies in the fact that the self-rated general health question contains an implicit valuation component while the long standing illness question does not. Therefore, while someone may report having a long standing illness the same person may nevertheless report their general health as being very good, because they may see the long standing illness as minor or unproblematic (e.g. minor skin complaints or correctable visual problems). Some support for this hypothesis is provided in section 5.1.1.

4.5.3 Self-reported general health, long standing illness and ADLs

Data from the 1994 GHS show that, among those aged 65 and over, the likelihood of reporting good general health declines as dependency score increases; almost half (48%) of respondents classified as 'independent' (level 1) said their health had been 'good' in the last 12 months, compared with 12% of those with a dependency score of 3.[38] Conversely, more than half (57%) of those with a score of 3 said their health was 'not good', compared with over a tenth (13%) of 'independent' respondents (Table 4.17).

Among respondents aged 65 and over, 41% of those who could manage all ADLs reported good general health, compared with 12% of those who could not manage one (Table 4.18). This supports the point made in section 3.5.2 suggesting that ADLs capture only very high levels of ill health (i.e. there is a strong ceiling effect). Even among those who could perform all the activities in the ADL checklist, over half reported that their health was less than good.

The likelihood of reporting a long-standing illness was also associated with dependency levels; just under half (48%) of respondents to the 1994 GHS aged 65 and over who were classified at Level 1 said they had a chronic condition, compared with 87% of those at Level 3 (Tables 4.19 and 4.20). This again supports the contention that ADLs are sensitive only to very high levels of physical dysfunction. However, 13% of those at level 3 said they had no long-standing illness, suggesting either that respondents were allowing for age when making their assessment on the long standing illness question or that the ADLs are measuring a dimension of health which is missed by the

[38] The number of respondents reporting difficulty with more than one ADL was too small to comment on.

long standing illness question. The former is probably the more likely explanation given the tendency for people to use just their peers rather than the entire population as a social comparison reference group.

4.5.4 Self-reported general health, long standing illness and the SF-36

Table 4.21 shows Eta coefficients for both self reported general health and long standing illness against the eight dimensions of the SF-36. While both are strongly and positively associated with all eight dimensions, the self-rated general health question has higher Eta scores than the long standing illness question on every dimension. This provides evidence for the superiority of the former item as a measure of general health. This is likely to be due to the fact that (a) there are more response alternatives on the self-reported general health measure and (b) the long-standing illness question does not require respondents to take the impact of the illness/disability on their functional *performance* into account. This means that many people report having a chronic illness which does not actually hinder their performance on the major dimensions of healthy functioning nor detract from their overall health related quality of life.

Another point to consider from table 4.21 is that the highest coefficients for both self-rated general health and long standing illness are the ones against the general health dimension of the SF-36. For the self rated general health measure, this in part reflects the fact that one of the scaled items of the SF-36 is a near identical version of the HSE self-rated general health item. Also, both items are least associated with the two 'mental health' dimensions of the SF-36 (emotional role and mental health). This gives further weight to the findings reported in section 3.2.3 that these single item 'global' measures of general health tend to be weighted primarily toward the physical dimensions of health, underplaying the influence of psycho-social factors on overall health.

4.5.5Self-reported general health, long-standing illness and EuroQol

On all five EuroQol dimensions, the likelihood of reporting good health was higher among those who said they had 'no problems' than among respondents who reported problems. On the 1996 HSE, for example, 79% or more of those reporting no problems on the different dimensions said their general health was 'very good' or 'good', while between 16% and 40% of those with moderate or severe problems described their health in this way across the five dimensions (Table 4.22). While these differences are large and in the expected directions, it is nonetheless surprising that significant proportions of respondents who say they have moderate or severe problems on important dimensions of health still rate their overall health so positively. This is likely due to both the insensitivity of the EuroQol descriptive system and the tendency of people to ‘discount’ problems in their global assessments of health. Virtually everyone who has any sort of problem on any of the

EuroQol dimensions is offered only two categories to describe the level of severity. This results in a situation in which people with very differing levels of severity of problem are placed in the same category. Likewise, people who have genuine health problems on particular dimensions may nonetheless report their health as ‘very good’, on the grounds that such problems are quite normal for their age, and/or is not life-threatening.

A similar association between self-reported long standing illness and the likelihood of problems on the EuroQol dimensions was also evident, with a higher proportion of those with severe problems than of respondents with no problems reporting a long-standing condition (Table 4.23). However, between 7% and 33% of those with 'some' or 'severe' problems on the EuroQol dimensions said they had no chronic condition, suggesting that the long standing illness question is missing levels of morbidity that are being detected by the EuroQol. Alternatively, however, because the EuroQol asks about health *today* and the long standing illness question asks only about *long standing* morbidity, it could be that the difference in estimates of ill health are at least partially the result of the EuroQol picking up acute illnesses which the LSI question explicitly asks respondents to ignore.

Mean scores on the EuroQol tariff were substantially higher amongst those reporting good health than among those whose general health was not good. On the 1996 HSE, for example, mean scores ranged from 0.93 for those with 'very good' health to 0.33 among respondents whose health was 'very bad' (Table 4.24). A similar pattern was found for the LSI question, with a mean EuroQol tariff score of .76 for those reporting a long standing illness and .93 for those reporting no long standing illness.

Almost 90% of those with 'good' health on self-rated general health placed themselves at 80 or above on the EuroQol VAS, compared with around half of those with 'fairly good' health and less than a fifth of those whose health was 'not very good'. This suggests that, despite these two questions using very different wordings and formats, they are still measuring essentially the same construct; a 'global' self-assessment of health (Table 4.25).

4.6 Floor and Ceiling Effects

Floor and ceiling effects concern both the sensitivity of instruments to differences in health state between population sub-groups and also their ability to detect longitudinal changes in health state at the population level. Instruments which have been designed primarily for use on clinical groups often place the majority of the general population in the 'best possible health' category. This means that the instrument is unable to detect small but real differences in health state between population sub-groups and may also fail to detect any improvement in health over time. Floor effects, as the name suggests, are the opposite of this; they occur when an instrument does not differentiate well between individuals and groups with poorer health. As with ceiling effects, any over-time decrement in health will go undetected by the instrument. Table 4.26 shows the proportion of respondents at the ‘ceiling’ and the ‘floor’ on each of the instruments under review (and for which we hold data). What is perhaps most apparent from table 4.26 is that the SF-36 has significantly fewer respondents at the ceiling than any other instrument. All four of the other instruments place somewhere near half the sample in the 'best possible health' category. In contrast to this, less than one percent of respondents are thus categorised on the SF-36. For the individual dimensions of the SF-36 significant proportions of the sample are at the ceiling, although these proportions are considerably lower than those found for the individual dimensions of the EuroQol.

Whether a ceiling effect is a serious problem on a health measurement instrument is not an absolute judgement but is determined by the level of sensitivity to difference/change in health state that is required in a particular context. For example, of the instruments under review, the EuroQol has a pronounced ceiling effect when compared to the SF36 - with 52% of respondents at the ceiling on the EuroQol and only slightly more than zero on the SF-36. If it were decided that the ceiling state of the EuroQol defines a level of health above which it would be unrealistic to aim to improve the health of the nation (or put another way if the ceiling health state of the EuroQol were used as a ‘Health of the Nation’ target) then the ceiling effect of the EuroQol would be unproblematic. However, if it were felt necessary to be able to make distinctions between population sub-groups in terms of low levels of ill health, then the quite pronounced ceiling effect of the EuroQol would make it an inappropriate instrument to use as a measure of the health of general population.

4.7 Criterion Validation

From previous sections it is apparent that any "valid" generic health measure would ideally permit us to:

(a) distinguish between those in ill-health and those not in ill-health
(b) discriminate between degrees of severity of ill health.

To establish the extent to which any of the measures we are considering meet these two criteria we, in principle, require an objective "gold standard" measurement per survey informant against which the operation of each generic health measure can be assessed. However such a gold standard does not exist, and it is in fact very hard to think of how one might be constructed.

In the absence of a "gold standard" criterion validation is not possible. However some approximation may be possible. The approach we have taken is necessarily piecemeal and incomplete. The results should be seen more as circumstantial evidence than as direct validation. Two main approaches have been used:

(i) validation against named long-standing illnesses

(ii) validation against "objective" health measures collected by the survey nurse (blood pressure, lung function).

We discuss each of these in turn below.

4.7.1 Validation against named long-standing illnesses

The question on long-standing illness on the Health Survey is followed, for all those with a long-standing illness, with questions about the nature of those illnesses. The responses are subsequently recoded using the ICD disease classification system.

The long-standing illness question is itself one of the generic health measures being assessed and the illnesses reported by respondents cannot be assumed to be a definitive list of the illnesses of respondents. Nor can we assume in advance that every illness listed is of sufficient severity to merit being considered as "ill-health" and hence to be included under an ideal generic health measure. Nevertheless, a considerable number of the inherently serious illnesses reported ought, it would appear, to be picked up as "ill-health" by a generic health measure. Furthermore, the data can be used to establish how different generic health measures rank different types of illness and hence how different generic health measures lead to ranking of illnesses in terms of severity. For the SF36 dimensions in particular, the data offers an opportunity to determine to what extent the eight dimensions differ in terms of the way in which they reflect the presence of these specific diseases or conditions.

The types of long-standing illnesses classified cover a broad range of conditions and include illnesses such as cancer, heart disease, back problems, migraine, hayfever, and mental illness.

The analysis clearly has its limitations. Although it allows us to make statements about those who report a long-standing illness, we have no means of validating the responses to the generic health measures made by those informants who claim to have no long-standing illnesses. Nor can we assess whether the generic health measures are adequately discriminating between degrees of health and ill-health among those claiming to have no long-standing illnesses. In particular there is some evidence from cognitive question testing work, that mental health problems are often not perceived as a long-standing illness and so the analysis gives rather little opportunity for validation of the mental health components of the generic health measures (although what little data there is on mental illness in the long-standing illness classification does reveal some interesting differences between the generic health measures - see below).An additional problem is that respondents frequently have more than one long-standing illness and this may affect the comparisons between conditions.

Tables 4.26a, 4.27, 4.28 show the 'position' of reported (and classified) long-standing illnesses as reflected by the self-rated general health question (percentage reporting bad or very bad health), the EuroQol tariff score, and the eight SF36 dimensions respectively. The actual number of long-standing illnesses classified was considerably greater than the seventeen shown in these tables: the seventeen were selected so as to be 'representative' of the range of conditions.

The tables are designed to demonstrate two things: firstly they show the ranking of illnesses in terms of the generic health measures (a position higher up the table always implying worse health), and secondly they show the relative distance between illnesses in terms of these measurements. For example, Table 4.26 demonstrates that 41% of those reporting having bronchitis or emphysema rated their health as bad or very bad. This percentage is higher than for any other condition shown. The next highest is the percentage for stroke/cerebral haemorrhage/cerebral thrombosis at 32%. The 'distance' between this condition and bronchitis/emphysema is shown on the table as a distance of nine percentage points. At the bottom of the table, just 3% of those reporting having hayfever as a long-standing illness rated their health as bad or very bad.

The patterns that emerge from the tables are fairly complex. Some of the most important things to note are as follows:

- Comparing across the eight SF36 dimensions (Table 4.28) the change in the ranking of long-standing illnesses is broadly as might be expected: the illnesses which score the lowest on physical functioning and role-physical being stroke, bronchitis/emphysema, heart attack, and arthritis/rheumatism. Interestingly, for the six dimensions from bodily pain through to mental health, illnesses classified as mental illness or anxiety score the lowest. Those classified as having epilepsy, fits or convulsions also score relatively low on these six dimensions.

- Moving across the eight SF36 dimensions, the relative position for some illnesses changes quite considerably. For instance, as might be expected from the above comments, mental illness/anxiety and epilepsy move from a low ranking on physical functioning to a high ranking on the last six dimensions. The pattern for migraine and headaches is similar. In contrast diabetes moves from a high ranking to a low ranking. Comparing rank positions across the dimensions in this way suggests that conditions such as diabetes cause physical problems without necessarily causing mental health problems. Other conditions, such as bronchitis, which have a high relative rank on all dimensions appear to cause physical problems and mental health problems. An external assessment of whether or not these patterns are reasonable would give a means of assessing the validity of the individual SF36 dimensions.

- The ranking of long-standing illnesses by the mean EuroQol tariff score is shown in Table 4.27. The ranking is very similar to that for the SF36 physical functioning dimension (with mental illness being placed close to the bottom of the table). This suggests that the two measures (i.e. EuroQol and SF36 physical functioning) may be operating in broadly similar ways. Perhaps surprisingly, one obvious difference between the two sets of rankings is that back problems and slipped disks have a similar mean on the EuroQol scale to cancer, whereas on the SF36 physical functioning scale scores for those with cancer are considerably lower than the scores for those with back problems.

- Table 4.26 shows long-standing illnesses ranked by the percentage of respondents with a particular illness rating their health as bad or very bad. It is possible to interpret placing on the table as a measure of perceived severity by respondents. In this instance the ranking of illnesses is broadly similar to the ranking on the SF36 role-physical scale. Comparing the two single dimension measures (i.e. self-rated general health and EuroQol) the rankings for the two are similar with the exception that mental illness has a much higher ranking on the self-rated general health scale.

In conclusion, all of the three measures of general health being assessed show rankings of long-standing illnesses that appear 'sensible'. For instance bronchitis, stroke etc. always appear toward the top of the tables and hayfever always appears at the bottom. Broadly speaking, the illnesses identified by the SF36 physical health dimensions as being most severe are also the ones identified as being the most severe by EuroQol and the general health question. Mental health problems are, however, rated lower on EuroQol than they are on either the general health question or on any of the SF36 dimensions.

4.7.2 Discriminating between good and ill-health

The analysis of the previous section addressed the question of whether the self-rated general health question, the EuroQol tariff, and the SF36 dimension scores discriminate well between illnesses which have been inferred from the way in which they are described to be of different severity. In this section the question of which (if any) of the three generic health measures discriminates best between good and ill-health is addressed.

There does not seem to be any way to assess this directly from the data because we have no means of independently identifying those in good health. However, again if we are willing to make use of imperfect indicators, then those with no long-standing illness might be taken as a sample approximating to a sample of those in good health. In practice the sample will include some people not in perfect health. Nevertheless if we contrast this (imperfect) sample with people who from the analysis of the last section we can assume are, on average, in fairly poor health then we might anticipate that a valid generic health measure would be able to discriminate well between these two groups. To test this four logistic regression models were fitted for each of the following illnesses: cancer, diabetes, mental illness/anxiety, epilepsy/fits/convulsions, stroke/cerebral haemorrhage, heart attack/angina, bronchitis/emphysema, arthritis/rheumatism/fibrosis. The first logistic regression model in each case uses age and self-rated general health as the predictors of ill-health; the second model uses age and the five EuroQol dimensions as the predictors; and the third model uses age and the eight SF36 dimensions as predictors. The fourth model used age alone as a predictor of ill-health. The difference between this final model and the previous three models gives a measure of the extent to which the generic health measures increase the ability to discriminate between good and ill-health once differences in health by age have been accounted for. In each case the comparison group was the set of respondents with no long-standing illnesses.

To compare the models, a chi-squared goodness of fit measure was calculated for each model. A summary of the results of the model fitting are given in Table 4.29. The smaller the model chi-squared value is, the better the model fits the data.

It can be seen from Table 4.29 that with the exception of stroke/cerebral haemorrhage, the eight SF36 dimensions give the best discrimination between "good health" and "ill-health". (For stroke/cerebral haemorrhage the five EuroQol dimensions do very slightly better.) The five EuroQol questions are better predictors of ill-health than the self-reported general health question for cancer, mental illness, stroke, and arthritis/rheumatism, but are poorer predictors for diabetes, epilepsy, heart attack/angina, and bronchitis.

4.7.3 Validation against blood pressure and lung function readings

The Health Survey does include some more obviously objective measures of health than the illnesses recorded after the long-standing illness question. In particular in the 1996 Health Survey (the year in which the SF36 questions and EuroQol was included) measurements were taken by nurses of blood pressure and lung function.

The main problem with these two objective measures, of course, is that they give only a very partial indicator of health. Whereas having high blood pressure or low lung function may be an indicator of poor health, having neither high blood pressure nor low lung function is not an indicator of good health. So the same problems of analysis arise as arose with the analysis of particular long-standing illnesses.

At a minimum however, we would expect average health, as measured by the various generic health measures to be poorer for those with high blood pressure or those with low lung function (for their height and age). Table 4.30 gives the means scores for the general health question, the EuroQol tariff, the SF36 dimensions, and the proportions with a long-standing illness(after standardising for age).

The figures of Table 4.30 demonstrate that, as might be expected, all of the generic health measures show those with either high blood pressure or low lung function to be in poorer health than others. Furthermore the SF36 dimensions are associated with high blood pressure and low lung function in ways we might expect, with the largest associations being found for dimensions relating to physical rather than mental health.

Using regression techniques (analogous to those described in the previous section) it appears that the best discriminator between high and normal blood pressure (after controlling for age) is long-standing illness. The best discriminators between normal and low lung function are the eight SF36 dimensions. (Table 4.31)

5 Cognitive work on the performance of health measures

This chapter summarises the results of two pieces of cognitive work carried out to assess health measures used on the Health Education Monitoring Survey (HEMS) and those proposed for possible inclusion in the 2001 Census.

Cognitive interviewing techniques are used to explore respondents' comprehension of a question, the strategies they use to retrieve relevant information from memory, and the decision process they follow when giving an answer. Thus, these cognitive methods explore the processes respondents use to arrive at an answer, rather than focusing merely on the actual answer. In the context of a survey sample, the number of respondents participating in cognitive question testing is small. The aim, however, is not to select a large representative sample, but to explore the meanings and processes which a sample of respondents – and by inference respondents in general - use when answering questions.

5.1 Health questions on HEMS[39]

The Health Education Monitoring Survey (HEMS) has been carried out annually since 1995 on behalf of the Health Education Authority (HEA). As part of the pilot for the 1997 survey, the HEA commissioned the Qualitative Methods Unit of SSD to carry out some cognitive testing on a number of questions, including general health and long-standing illness. Fourteen men and 16 women aged between 18 and 75 were interviewed as part of this work.

5.1.1 Health in general

Ask all respondents

Now I would like to ask you some questions about your health. How is your health in general? Would you say it was..

1 very good
2 good
3 fair
4 bad
5 or very bad?

When asked to define the term 'health in general' most respondents were able to say clearly what aspects of health they thought the term was referring to and mentioned a number of dimensions. 'Health in general' was seen as the absence of ill health. Respondents talked about not being sick, never getting colds, rarely needing to see a doctor or take time off work. 'Health in general' was also seen as related to the extent they felt their heath restricted their ability to live a normal life. To have good health

[39] This is a summary of a report by Linda Mortimer, SSD, published in ' Title to be decided (HEMS 1997)' TSO 1998

in general meant to some respondents that their health did not prevent them from doing anything they wished to do. Respondents also referred to it as their state of mind, how they felt in themselves. 'Health in general' was also seen by respondents in terms of how 'fit' they thought they were, and was related to the amount of exercise taken and the type of diet followed.

Although the question does not ask about a specific time period, respondents were asked what time period they were considering when answering this question. Some respondents reported thinking about how their health had been over the last 12 months, either because some of the HEMS questions mention 'twelve months' or because they had just been asked to recall significant events over the previous 12 months[40]. Those who reported thinking of a different period of time considered their health over the last one to two months, while others reported thinking back across the last 'few' years and even the whole of their life. Some respondents even reported thinking of how their health was going to be in the coming months.

Although the question did not ask respondents to compare themselves with others, some did, usually comparing themselves to friends or family of a similar age but whose health was worse. Their response therefore included some relative assessment of health.

Respondents were asked to explain what they understood by each answer category in this question. Someone with 'very good' health in general was seen as having no restrictions on the things they could do. Such a person was thought by many respondents to have no health problems at all, and to be in exceptional condition for their age, again a relative assessment. Some respondents talked about such a person even being an Olympic athlete. Very good health in general was thought to be not only physical health but also mental health, both maintained by lots of exercise. Some saw no difference between 'very good' and 'good'.

The phrase 'good' health in general was understood by respondents to be referring to 'normal health', where they may have slight health problems such as colds or flu, or mental health problems that prevented them from being in very good health. Someone with good health was thought by respondents to be fit, but not an athlete.

The phrase 'fair' health in general was thought by respondents to be referring to health that was worse than good health, and that could be the average of good and bad 'health' days. Someone with fair general health would not be as fit as someone with good health, and might have a sedentary job. Respondents reported that there would be some restrictions to the things that someone with fair health could do, such a person would be more sickly and perhaps even need a short stay in hospital for some

[40] The questionnaire included a section asking about important life events in the 12 months between the two interviews. In the pilot, these questions were asked *before* the general health section of the interview, but were asked *afterwards* in the main fieldwork.

reason. However respondents thought that someone with ‘fair health in general’ could be worse.

The phrase ‘bad’ health in general was seen as applying to the negative aspects of health. Someone who had bad health was thought by respondents to have been ill for some time which would have made them unfit for work, again introducing a time dimension, and would restrict their ability to perform their ‘daily tasks’. It was also seen as referring to smokers and drinkers, and people who do not know about fitness and diet. ‘Very bad’ health was thought to be severe and a condition that could affect all aspects of a person’s life.

5.1.2 Long-standing Illness, disability and infirmity

Ask all respondents

> *Do you have any long-standing illness, disability or infirmity? By long-standing I mean anything that has troubled you over a period of time or that is likely to affect you over a period of time?*
>
> *1 Yes*
> *2 No*

Respondents’ understanding of the terms ‘illness’, ‘disability’ and ‘infirmity’ was explored during the cognitive interview. ‘Illness’ was seen by respondents as the extent to which normal life was restricted by minor illness, such as colds and coughs, or more serious ailments such as asthma. An illness was seen as either treatable, or something respondents thought they would recover from. Unlike the question on self-reported general health, this question did include a time dimension. The period of time that respondents thought someone needed to have had an illness for it to be long-standing varied greatly. Respondents did not express a strong opinion, mentioning periods ranging from a matter of months to the whole of someone’s life. Some talked in terms of ‘from the moment of diagnosis’ or ‘for rest of life’.

‘Disability’ was also thought of as a restriction on normal life, and respondents mentioned the term as referring to both mental and physical disabilities. They thought that someone could be born with a disability, or could acquire it through an accident. A disability was seen as permanent, and not thought by respondents to be curable.

Respondents had difficulty in explaining the term ‘infirmity’. Whilst some associated it with old age, and used phrases such as the ‘old and infirm’, most either could not give a definition or saw it as interchangeable with illness and disability.

5.2 Health questions for the 2001 Census[41]

It is likely that the 2001 Census will contain questions on health to be answered about each individual in the household. To aid in the design of these questions, Census Division of ONS commissioned SSD to carry out cognitive work in connection with the 1997 Census test, which included questions on self-assessed health and on limiting long-standing illness. Interviews were carried out with 147 respondents.

5.2.1 Limiting long-standing illness

Ask all respondents

> *Do you have any long-term illness, health problem or disability which limits your daily activities or work you can do?*

The question was answered in two quite distinct ways. Some respondents answered 'yes' if they were *diagnosed* as having a long-standing illness, health problem or disability, while for others, the key factor was whether their ability to work, or take part in other activities, was affected. For this latter group, those who had learnt to live with an ailment, even if it curtailed their activities, often answered 'no' to the question. Those whose ailments were perceived as due to their age (e.g. asthma in the young, or illnesses connected with increased age) also often answered 'no' to the question.

The categories 'long-term illness' and 'disability' were generally understood, but were not always seen as distinct from each other. Some respondents felt that the difference was that the former was possibly treatable, while the latter was not.

A number of respondents found the term 'health problem' difficult to define and differentiate from the other terms. Some felt it was a very general term that might cover a range of both short and long-term ailments, while some felt it would cover only mild ailments.

5.2.2 Self-assessed general health

Ask all respondents

> *Over the last twelve months would you say your health on the whole has been:*
>
> *Good*
> *Fairly good*
> *Not good*

[41] This is a summary of work carried out by the Qualitative Unit of SSD on behalf of Census Division of ONS. Further details can be obtained from Jack Eldridge at ONS.

This question was generally understood by respondents as asking them to categorise their level of health during the past year. Most respondents thought of physical health when answering this question although a few mentioned psychological health.

The answer categories 'good' and 'not good' were generally understood. Fairly good was often thought to be vague and was not as consistently defined by respondents as the other two categories. For some respondents, the difference between 'good', "fairly good' and 'not good' was whether or not the illness or ailment was one that a person could cope with, or had learnt to live with, regardless of it's severity. That is to say that they would not require extensive medical treatment or have to take very long periods off from work or school. In some cases the severity of the illness was used as a guide to answer. Life threatening, chronic or terminal illness were often thought to warrant a classification of 'not good'. Other respondents used their rate of recovery from serious ill health as the key to making a judgement.

In other cases it was not the severity of an illness, but the number of illnesses that were used as a guide when answering this question. Some respondents used an activity based assessment when making a judgement. The difference between the categories was often decided on the basis of whether or not they had been hindered or obstructed from carrying out daily living activities or going to work or school.

Many respondents thought about their health over the preceding year. However, some respondents did not include major ailments during that time if they felt they were not representative of their year as a whole.

Some respondents did not simply think about health over the previous year, but compared it to earlier periods. Longer periods of time were sometimes used if the respondent felt the last year was not representative; perhaps because of an operation or a long period in hospital. Some respondents suffering from chronic health conditions identified their health as being good because they had compared it to other years when their condition had been worse.

In other cases respondents used a shorter time frame of less than a year, back to the beginning of the calendar year or to Christmas. Some respondents included life-cycle ill-health issues in their judgement such as pregnancy related ill-health or ill-health commonly associated with old age. Some respondents compared their health to others when making a judgement. Their reference point for comparison purposes was often a person in very poor health or someone with a disability. Given this spectrum their own health was then *nearly* always identified as good. Others used a preconception of what the average person's health is like, or how often the average person is off sick or visits the doctor.

6 Summary of main findings of the review

The terms of reference of the study specified a literature review plus further analysis to evaluate various measures of “general health” which had been included, or might in future be included, in general population sample surveys in the UK. In this chapter we summarise the results presented in the main body of the report. These have resulted from our attempts to apply the criteria at section 2.8 to the measures listed at Table 1A (and in certain cases Table 1B).

6.1 Single Item Measures

The two question modules examined were:

- A question on “Long-standing illness, disability or infirmity” (LSI or LLSI), with follow-up questions on whether any illness, disability or infirmity limits the person in their normal activities and on what the nature of the illness etc is; and

- A single question on “Self-assessed general health over the past twelve months” (SAGH).

Responses to (L)LSI and SAGH have traditionally been interpreted as direct measures of *general health*. The pre-specified response alternatives classify the person’s state of health in terms of 2-5 ordered categories, with some categories signifying better and others worse health.

In terms of overt conceptual content the two questions differ in that LLSI stresses the *disabling* effects of ill-health, whereas the SAGH does not. In contrast with, say, the EuroQol or SF36 instruments, they appear to have been devised and included in surveys mainly on criteria of face validity and practicality. Neither has any pedigree of theoretical derivation or methodological development and we are not aware of any clear published statements of what each of these questions is *intended* to measure (e.g. what they are intended to include and exclude). This is probably because face validity has been assumed to speak for itself.

The questions are simple to ask, are usually readily answered and take little interview time. For each measure there are time-series based on GHS and HSE results and the questions have also been included in many other health surveys. Practicality, familiarity and long periods of usage have lent these questions canonical status, with new users assuming that their validity, reliability and utility has already been established.

6.1.1 Long-standing illness, disability or infirmity

LSI and LLSI have been included year by year in the GHS since the early nineteen seventies. In certain years the responses to the follow-up question ‘What is the matter with you?’ have been coded to a set of groups which roughly correspond with the broad headings of the International Classification of Diseases (ICD). However, this is not true ICD classification, as ICD is designed to classify according to disease diagnosis and system of the body affected, whereas the nature of the responses compels the GHS to code largely according to reported symptoms[42]. It should also be remembered that a classification such as the ICD, though entirely appropriate to the aim of charting the incidence and prevalence of specific diseases, does not readily yield a measure of health-related quality of life.

Since 1991 the LSI and SAGH questions have also been included in the Health Survey for England. A question inspired by the GHS/HSE limiting long-standing illness question was included in the 1991 Census, but changes in the wording convert the census version into effectively a different question from that used in the sample surveys. This is mainly because of attempts to focus the question on disability affecting employment. The effect can be seen in the different shape of the curves relating (limiting) long-standing illness to age (Thomas and Purdon 1994). A version of the LSI question, which may be further modified, is under consideration for the 2001 Census.

It appears that the LSI question is interpreted, as intended, in terms of the presence or absence of chronic illness or disability as perceived by the respondent. The rather outmoded word ‘infirmity’ seems to add little to the meaning, apart from suggesting something to do with old age. Evidence from cognitive testing studies also suggests that different individuals interpret key words in the LSI and LLSI questions differently, so that they effectively measure rather different things for different people.

Probably the strongest and most systematic bias is one deeply routed in our culture and discourse. It is a tendency for older respondents to discount conditions and disabilities that they regard as inevitable concomitants of old age. This has the effect of understating how much of the burden of long term illness and disability is borne by the elderly.

Further evidence from cognitive question testing studies suggests that in answering what are intended to be all-embracing health questions, some respondents in practice take little or no account of mental ill-health. This may be either because of the stigma involved in identifying oneself as mentally ill, or because some people interpret the terms ‘illness’ and ‘disability’ primarily in terms of physical problems. Chronic impairments, conditions or disabilities to which the sufferer has adapted and which he or she does not see as constituting “ill-health” are also liable to be discounted.

The proportion of the population reporting (limiting) long-standing illness on the GHS in successive annual surveys has tended to drift upward quite substantially over time (with the age distribution of the population controlled). It is generally agreed that an objective trend towards more limiting long-standing illness as recognised clinically is implausible when set beside other, more objective evidence and it has been suggested that responses to the LSI question are affected by rising health expectations. This implies inconsistency between the results for different time

[42] See GHS 1989 for a fuller description

periods, considered as objective indicators of health. There is also suggestive evidence that LLSI may have been reported disproportionately by older working age people who have become long-term unemployed and are claiming disability benefits (SAGH, the wording of which does not mention "disability", does not show similar effects). Such relative bias and instability over time would undermines the aims of general health monitoring using this question.

Response distributions for LSI vary substantially between the GHS and the HSE (the latter showing significantly higher rates) in ways that cannot be explained in terms of sample differences. We believe that this is probably a survey context effect, indicating lack of portability of the measures. A likely cause of this contextual variation in response distribution is the substantive content of the other questions in each survey – with a high content of health questions stimulating respondents to think more about different aspects of their health.

6.1.2 Self-assessed general health

SAGH has been used with two different response formats, one with three categories on the GHS and, and one with five on the HSE. This again has effectively created two different questions. The five point scale, which is an international standard, is in general preferable to the three point scale, though the latter has for many years been used as an introduction to the main health section on the GHS. In contrast to (L)LSI the results have shown considerable stability over time, but of course that could indicate either robustness to disturbing influences or lack of sensitivity to real change. We understand that a version of SAGH is also under consideration for the 2001 Census. There is some evidence from a number of longitudinal studies that persons who assess their own general health as "poor" in response to SAGH have a significantly shorter life expectancy, even controlling for a broad range of important background characteristics.

Both LLSI and SAGH are insensitive in discriminating between the broadly "good" levels of health enjoyed by the majority of the population. Cognitive testing suggests that they tend to miss diseases and conditions which are undiagnosed and do not produce very noticeable symptoms, or are controlled by drugs so that they do not seriously affect quality of life. However, it can be argued that identifying undiagnosed conditions that do not yet produce noticeable symptoms or affect quality of life is more properly a function of targeted health screening than of general population health monitoring.

In view of these problems we conclude that these single item measures are of limited use on surveys as indicators of health levels and trends in the general population. However, they do provide an easily administered measure which can be used for predicting variation in mortality and the demand for services among different groups in the population

6.2 Multi Item Health Profiles and Utility Scores

6.2.1 Background

Different aspects of general health (e.g. physical and mental and different sub-components of these) may well follow different trends and for health monitoring purposes it seems essential to us to be able to demonstrate such trends through what we describe as a ***health profile***. However, that still leaves the problems of how to specify and measure health profiles at a level of detail which is neither too summary, so that important differences and trends are concealed, nor too detailed.

One example of a health profile measure, the EuroQol health description system, has already been discussed. A five-score profile such as EuroQol is arguably too summary, but on the other hand the interpretation of an overly detailed profile can become bogged down in a mass of technical and presentational detail, and might not be succinct and robust enough to be included in surveys administered by interviewers (see section 2.8.8).

There were several other health profile instruments with a *prima facie* claim to be considered, the two main ones being the Nottingham Health Profile (NHP) and the SF-36, together with its reduced version the SF-12. All of these are item batteries which are designed for self-completion, but can also be administered by interviewers.

We consider special health state descriptive systems and health utility scores together because, in practice, the classification systems have been developed as a step on the way to scoring health states and hence individuals reporting those states in terms of a health utility function.

It is generally agreed that, after combating life-threatening disease, a second key aim of health services should be to maximise health related quality of life across the population. The classification systems referred to here focus on health-related quality of life and are to be sharply distinguished from systems which are intended to enable quasi-medical diagnosis of diseases and conditions, such as the International Classification of Diseases (ICD).

6.2.2 The EuroQol Instrument

The first approach that we examined under this heading was the EuroQol health state descriptive system. The Department of Health arranged for the EuroQol instrument[43] to be administered to adult respondents as a self-completion instrument in the 1996 HSE, thus providing data which enabled it to be compared with other measures, including the SF-36 and the single question general health measures, that were also included in that round of the HSE. EuroQol was also included on the

[43] Without its visual analogue health rating scale component.

ONS omnibus to check for context effects as between the two surveys and there were some differences in the score profiles obtained.

The EuroQol health descriptive system may be compared with other descriptive systems developed for similar purposes in the context of the SF-36 and the Canadian HUI.

The scaling to derive utility scores for health states in the EuroQol was done using a special survey of the general public, carried out by the *National Centre for Social Research* for the Centre for Health Economics at the University of York. In the calibration study respondents were asked to “value” selected health states using Contingent Valuation methods known as “Time Trade-off” and “Standard Gamble”. The results were then modelled so as to allow scores to be interpolated for health states not directly valued in the calibration exercise.

In non-technical terms the aim was to score each health state recognised by the descriptive system in terms of relative desirability / undesirability (or ‘utility’). The theory underlying the method suggests that the result should be an ‘interval scale’ of generalised health utility, which can be interpreted as a scale measuring *general health* from a HRQOL viewpoint. That in turn implies that the score differences between health states correspond to standard calibrations of desirability, or units of health-related quality of life.

Unlike the underlying health descriptive system, the utility function allows very disparate conditions to be assigned similar scores – which is, indeed, what *any* attempt to measure “general health” as a single value must do. Unlike the descriptive categories, therefore, the scores lack ‘face interpretability’. On the other hand it allows the utility values of any two health states to be assessed and compared and the positive or negative utility (as evaluated by an average member of the public) of movement between health states (such as might be produced by medical intervention) to be estimated.

6.2.3 Canadian Health Utilities Index

At an early stage of the review we took note of the Canadian Health Utilities Index (HUI), which is constructed on lines that are similar in broad conceptual terms, though different in technical detail, to the health utility function derived from EuroQol. However, it appeared never to have been used in the UK and we found it difficult to obtain sufficient technical information from the developers, so decided the effort of subjecting it to a fully detailed examination would not be justified. However, we feel that, if a health utilities approach is to be seriously pursued, the Canadian HUI should certainly be revisited.

6.2.4 OPCS disability scoring system

Another implementation of a single ‘utility’ scale applicable to heterogeneous ‘health’ states is the method developed by OPCS (ONS) in the early nineteen eighties for placing different and multiple disabilities on a single scale of severity. The disability surveys were sponsored by the (then) DHSS and the object was to calibrate degrees of disability, for purposes of setting disability benefit levels. Methods were devised for equating disabilities affecting different bodily and mental systems and for combining and calibrating multiple disabilities suffered by the same individual. Panels of judges were used, broadly analogous to the sample of ‘judges’ drawn randomly from the general population used in the EuroQol calibration exercise.

Disability may lend itself to this treatment more readily than the wider concept of health, but it seemed important that reviews of the kind we have conducted should take account of technical approaches developed in the area of disability, since as we have pointed out above the concepts and measurement problems of “health” and “disability”, as measured by general population surveys, are closely intertwined.

6.2.5 The Nottingham Health Profile

The NHP, as its name suggests, originated in this country and has been used mainly in the UK. It has been included in a number of studies, though not, so far as we are aware, in any large scale national government surveys, and has generated some methodological research literature.

The instrument is a “health profile” only in the sense that it contains items designed to pick up different symptoms of ill-health, but at the output stage these are not typically presented as a profile, but instead are combined to provide a single overall health score. The NHP has, in fact, been designed to identify individuals with health problems and to provide a single-number index of health which can be used to compare groups, rather than to provide a “profile” of health across the general population in the sense intended here. In that respect it is analogous to the General Health Questionnaire.

Being particularly designed to identify those with significant health problems, the NHP appears to suffer from “ceiling” effects. There is rather little evidence about the internal statistical structure of the instrument and how this structure relates to the measurement of different aspects of health. It does not offer measures of a number of separate and well defined dimensions of health which can be shown to be valid and reliable, taken separately. We therefore concluded that it was unlikely to offer advantages over the SF-36 and decided not to include the NHP in our main review.

6.2.6 The General Health Questionnaire (GHQ)

In spite of its ‘camouflaged’ title the GHQ is actually an instrument designed to identify persons in the community who appear to suffer from significant mental health problems (mainly anxiety and depression, as opposed to psychotic illness). Like the NHP it has a number of items and provides an overall score with a cut-off point that identifies those with mental health problems. We took account of it as a possible substitute for the mental health scales of the SF-36 or as an accompanying measure to go alongside measures without an explicit mental health dimension such as SAGH or LSI but it does not provide a general or overall health profile score.

6.2.7 The Short Form-36

The SF-36 originated in the USA but now has a UK version adjusted for differences between American and British English. The SF-36 health descriptive and measurement system has a developmental history stretching back to the nineteen eighties. It has been used in numerous clinical and general population studies both in this country and in others and has generated a considerable literature, including a detailed account of its derivation and empirical validation. The development process was based on systematic definition of health domains, combined with item and factor analysis, to identify and measure eight dimensions which together covered the health domain. The eight dimensional sub-domains of the SF-36 have face validity and appear to have encouraging levels of concept, criterion and convergent/discriminant validity and sensitivity.

The Department of Health arranged for the SF-36 to be included as a self-completion instrument in the 1996 HSE, thus providing additional empirical data and enabling it to be compared with other measures in the context of a survey vehicle which is central to the Department's health monitoring strategy.

The SF-36 seems to us to be aimed at the most useful level of generality for a policy-relevant health profile. Using the HSE data we have checked as far as possible the construct validity of the eight dimension scores and we are confident that they genuinely reflect different facets of health, including mental health. To get from the responses to the 36 questions of the SF-36 to the 8 dimension scores, the responses are coded, summed and transformed onto scales from 0 (worst) to 100 (best). This transformation procedure inevitably reduces the transparency and immediate interpretability of the numbers thus generated. However, we feel that *as a descriptive system*, its eight-dimensional structure is superior to the EoroQol system both in terms of health domain coverage (eight health dimensions being more adequate then five) and in terms of sensitivity (finer gradations of health are validly distinguished on each dimension).[44] It appears not to suffer from 'ceiling' or 'floor effects, but calibrates the whole range from excellent to very poor health. Therefore, unlike the EuroQol system, it does not invite the interpretation that about half the population has 'perfect health' and that the majority of the remainder can be described in less than twenty different health states. Our analyses provided evidence that the extra sensitivity of the SF-36 is genuine, marking real variations in health.

[44] The HUI also offers eight dimension scores, which, to judge from item content and the factor labels assigned, are broadly similar to but not identical in detail with the eight SF-36 dimensions.

Currently experimental and developmental work is being done with the aim of deriving from the SF-36 health state classification system a method of allocating utility scores to health states which will parallel the EuroQol and HUI products. However, no conclusions can yet be drawn as to whether an SF-36-based health utility measure will perform better or worse against our criteria than the EuroQol measure.

Since its introduction, further research and development have added to the SF-36 health measurement system two scores for higher level factors corresponding to "Physical" and "Mental" dimensions of health. These still require all 36 items to be answered. We think these, while gaining in presentational and interpretable succinctness, clearly offer less conceptual coverage than the eight-dimensional profile.

6.2.8 The Short Form-12 (SF-12)

Another recent development of the SF-36 approach is the introduction of a reduced item battery, the SF-12. This has been created through item analysis of the SF-36 with a view to measuring the underlying dimensions using only 12 items, while minimising the loss of measurement precision. The SF-12 has an obvious practical advantage over the SF-36 in that it requires less time for completion whether interviewer-administered or self-completed. However, only the broad "Physical Health " and "Mental Health" dimensions are derivable from this shortened version. Countering this however, is the time take to complete the instrument. While the SF-36 takes on average 10-15 minutes to complete, the SF-12 takes only 2-3 minutes.

We have examined the limited number of reports in the literature on the performance of the SF-12. The consensus so far appears to be that, as claimed by the developers, relatively little is lost from the "Physical Health" and "Mental Health" factor measures if they are based on the SF-12, rather than the SF-36. The data available to us for analysis resulted from administration of the full SF-36 as part of the HSE and further checking is required to determine whether the SF-12 which we constructed by extracting the relevant data items from the full SF-36 can be assumed to perform in exactly the same way as a "free-standing" SF-12 (i.e. whether the SF-12 items perform in a context-invariant way). However, given the concerns over context effects which apply to all these instruments, more research is required on the behaviour of these items when administered within the full SF-36 and as a stand-alone instrument before real confidence can be placed in the ability of this shortened measure to replicate, in a simpler format, the results of the full 36 item version. If context invariance were established and the requirement is for a simple two-element profile, then the SF-12 would probably be a strong candidate.

However, we remain of the opinion that just two scores will not do such a good job in capturing the complex concept of "general health" in a way that has the transparency and interpretability ideally required for health policy monitoring purposes. On surveys that are carried out regularly, a compromise position might be to administer the full SF-36 periodically, and to use the SF-12 in the intervening rounds. This would ensure a full time-series of Physical Component Scores and Mental Component Scores, with full SF-36 profiles administered intermittently.

6.2.9 SF-36 and EuroQol compared

The EuroQol system aims to describe all possible health states in terms of five dimensions, with a three point ordinal “good to bad” scale for each. This gives in theory 243 possible health states. In practice, however, almost all responses from cross-sectional samples of the general public place them in one of about 12 health states, with the “no health problems” state accounting for around a half of all cases. We are not convinced that this system as it stands has been shown to have the required levels of concept coverage and sensitivity, being subject to strong “ceiling” effects (inability to detect differences between persons with better levels of health). Amongst elderly people, however, where ill-health is primarily concentrated, EuroQol surveys find a majority of respondents to have some form of health problem.

EuroQol may give a smaller range of scores than SF-36. On the other hand, the interval properties of EuroQol’s utility index do indicate the magnitude of health problems as seen by the general population and such information could have an important place in setting health policy priorities.

Another consideration is the possibility of linking in with cost-effectiveness analysis (CEA). Evidence-based practice in health care is vitally important and CEA is a key element of this. EuroQol is well suited for cost-effectiveness analysis (whereas SF-36 does not currently have the interval scale utility score necessary for this type of analysis). If population surveys find health problems described in EuroQol terms, there may be relevant CEA studies that show the effectiveness and costs of remedies in the same terms.

6.3 Mental health

Cognitive work has shown that members of the public do not always perceive common neurotic ailments, mental distress and mental disabilities within the same frame of reference as physical health, and hence as relevant to the concept of “general health”. Measures which do not include items explicitly devoted to identifying symptoms of mental illness therefore tend to have poor coverage of this dimension of health. An advantage of the SF-36 is that it does have such coverage.

As noted, we also considered the General Health Questionnaire (GHQ) which was designed as a screening tool to identify persons with *prima facie* mental health problems. We concluded, however, that it appears to offer little more than the SF-36 Mental Health dimension score.

6.4 Measurement of disability and ability to perform activities of daily living

“Disability” and “(chronic ill-)health” can be clearly distinguished at the conceptual level. “Health” is the broader concept. However, at the level of practical survey measurement they are impossible to distinguish clearly, particularly at the poorer levels of health and amongst the elderly. Disability is related particularly closely to quality of life.

Succinct measures of specific disabilities, such as, for the elderly, (Instrumental) Activities of Daily Living and, for the general population instruments such as the OPCS Disability Survey sift questionnaire have high face, construct and criterion validity and good transparency and interpretability. However, they tend for some types of disability to suffer from “ceiling effects” – that is, from insensitivity in discriminating between milder levels of health-related disability. Also, of course, they do not on their own cover the whole of the conceptual domain of “health”.[45]

We believe that functionally referenced measures are highly relevant to health-related quality of life, as viewed by the individual, and that responses to questions of this kind are more likely than others to be resistant to extraneous biases, precisely because they invoke specific behavioural criteria and are rooted in everyday life experience. Items have been devised to measure (in)ability to cope with emotional and intellectual demands as well as the physical and mental deterioration and frailty most typical of elderly people. Questions used in the GHS and the Family Resources Survey to measure individuals’ degree of dependency upon household carers have carried this line of development forward in a way which is directly relevant to the interests of many users of population information on general health. It is arguable that this is the type of measure that would be most helpful if included in the 2001 Census.

However, most current functional approaches are designed to detect the grosser forms of functional disability that are typically suffered by the elderly. The corresponding functional concept for younger people is perhaps “fitness”, in the sense of being able to perform more demanding physical (and also mental) activities.[46] There is evidence that people who are unfit and overweight in their youth and middle age are more likely to become seriously disabled at younger ages than those who keep fit. More work needs to be done to devise and validate measures of fitness (measures such as lung function and physical stamina, for example), with a view to discriminating the health status of predominantly younger people who do not, as yet, show obvious symptoms of degenerative disease, and this would be a valuable area for future research.

[45] See section 6.1.2 above.

[46] There is evidence from cognitive question testing studies that some, predominantly younger, people interpret “global” general health questions partly in terms of “fitness”.

Appendix A

Table 4.1 A comparison of self-reported general health using GHS question, on GHS 1996 and Omnibus 1996

Adults aged 16 and over ***England***

Age		**16-24**		**25-34**		**35-44**		**45-54**		**55-64**		**65-74**		**75 and over**		**Total**	
Sex	**Self-reported general health**	%		%		%		%		%		%		%		%	
		GHS	Omnibus	GHS	Omnibus	GHS	Omnibus	GHS	Omnibus	GHS	Omnibus	GHS	Omnibus	GHS	Omnibus	GHS	Omnibus
Men	Not good	2	1	4	4	6	7	12	11	17	16	18	12	18	15	10	9
	Fairly good	26	23	26	24	24	30	29	31	34	35	39	47	48	56	30	33
	Good	72	75	70	72	70	63	60	57	49	49	43	40	34	29	59	58
Women	Not good	5	5	9	6	10	10	13	8	17	16	18	18	25	18	13	11
	Fairly good	34	32	30	32	31	34	33	36	35	40	42	46	45	52	35	38
	Good	61	63	61	61	59	56	53	56	48	44	40	36	29	29	52	52
Bases																	
	Men	*744*	*343*	*1162*	*416*	*1104*	*384*	*1100*	*423*	*803*	*350*	*728*	*273*	*514*	*178*	*6155*	*2365*
	Women	*833*	*285*	*1393*	*531*	*1312*	*488*	*1220*	*457*	*926*	*318*	*856*	*330*	*694*	*215*	*7234*	*2624*

Table 4.2 Percentage reporting a long-standing illness by survey, age and sex

Adults aged 16 and over ***England***

Age		**16-24**	**25-34**	**35-44**	**45-54**	**55-64**	**65-74**	**75 and over**	**Total**
Survey	**Sex**	Proportion reporting a long-standing illness							
Omnibus 1996	Men	14	21	26	36	47	52	58	34
	Women	20	20	31	38	43	55	55	35
GHS 1996	Men	22	26	32	39	55	60	63	39
	Women	24	26	30	41	55	58	68	40
HSE 1995	Men	21	26	34	44	57	65	64	42
	Women	26	25	33	43	54	60	67	41
Bases = 100%									
Omnibus 1996	*Men*	*343*	*416*	*384*	*423*	*350*	*273*	*177*	*2365*
	Women	*285*	*531*	*487*	*457*	*318*	*329*	*215*	*2622*
GHS 1996	*Men*	*893*	*1365*	*1280*	*1242*	*881*	*758*	*537*	*6956*
	Women	*919*	*1468*	*1358*	*1279*	*953*	*877*	*736*	*7590*
HSE 1995	*Men*	*931*	*1395*	*1386*	*1183*	*1000*	*921*	*519*	*7335*
	Women	*1084*	*1738*	*1502*	*1380*	*1119*	*1060*	*836*	*8719*

Table 4.3 A comparison of limiting long-standing illness using GHS question, on GHS 1996 and Omnibus 1996

Adults aged 16 and over ***England***

Age	**16-24**		**25-34**		**35-44**		**45-54**		**55-64**		**65-74**		**75 and over**		**Total**	
Sex	%		%		%		%		%		%		%		%	
	GHS	Omnibus	GHS	Omnibus	GHS	Omnibus	GHS	Omnibus	GHS	Omnibus	GHS	Omnibus	GHS	Omnibus	GHS	Omnibus
	Proportion with a limiting long-standing illness															
Men	10	6	14	13	18	15	24	24	39	34	41	33	49	47	25	22
Women	12	12	15	12	20	20	28	24	37	29	39	37	53	45	27	23
Bases																
Men	*893*	*343*	*1365*	*416*	*1279*	*384*	*1240*	*423*	*881*	*350*	*758*	*273*	*537*	*177*	*6953*	*2365*
Women	*919*	*285*	*1467*	*531*	*1358*	*487*	*1277*	*457*	*953*	*318*	*877*	*329*	*735*	*215*	*7586*	*2622*

Table 4.4 Health measures by doctor consultations in the last 2 weeks; GHS 1996

England

		Men		Women		*Base*	
		Consulted a doctor	Had not a doctor	Consulted a doctor	Had not a doctor	*Men*	*Women*
Longstanding illness							
Yes	%	22	78	30	70	*2723*	*3074*
No	%	9	91	15	85	*4228*	*4513*
Limiting							
Longstanding illness							
Yes	%	27	73	34	66	*1714*	*2034*
No	%	10	90	17	83	*5234*	*5549*
General Health							
Not good	%	35	65	42	58	*619*	*940*
Fairly good	%	19	81	24	76	*1878*	*2526*
Good	%	9	91	14	86	*3657*	*3767*
Pain or discomfort							
Some or severe problems	%	26	74	33	67	*1933*	*2455*
No problems	%	10	90	15	85	*4226*	*4784*
Self-care							
Some or severe problems	%	35	65	38	62	*275*	*399*
No problems	%	14	86	20	80	*5884*	*6840*
Usual activities							
Some or severe problems	%	35	65	38	62	*708*	*1065*
No problems	%	12	88	18	82	*5451*	*6174*
Mobility							
Some or severe problems	%	30	70	34	66	*1083*	*1388*
No problems	%	11	89	18	82	*5076*	*5849*
Anxiety and depression							
Some or severe problems	%	27	73	33	67	*961*	*1570*
No problems	%	12	88	18	82	*5196*	*5663*

Table 4.5 Self-reported general health using GHS question by survey, age and sex

Adults aged 16 and over ***England***

Age			**16-24**	**25-34**	**35-44**	**45-54**	**55-64**	**65-74**	**75 and over**	**Total**
Survey	**Sex**	**Self-reported general health**	%	%	%	%	%	%	%	%
Omnibus 1995	Men	Good	70	73	70	60	51	45	36	61
		Fairly good	29	23	24	29	34	42	47	30
		Not good	2	4	6	12	15	14	17	9
	Women	Good	60	68	61	55	46	35	33	54
		Fairly good	37	26	30	35	40	46	42	35
		Not good	3	6	9	10	14	19	25	11
Omnibus 1996	Men	Good	75	72	63	57	49	40	29	58
		Fairly good	23	24	30	31	35	47	56	33
		Not good	1	4	7	11	16	12	15	9
	Women	Good	63	61	56	56	44	36	29	52
		Fairly good	32	32	34	36	40	46	52	38
		Not good	5	6	10	8	16	18	18	11
Omnibus 1997	Men	Good	74	77	66	64	55	49	42	64
		Fairly good	22	21	28	23	22	30	30	24
		Not good	3	2	6	13	23	21	29	12
	Women	Good	58	69	59	62	59	42	34	58
		Fairly good	32	25	29	25	32	37	39	30
		Not good	11	7	12	13	9	21	27	12
Bases = 100%										
Omnibus 1995		*Men*	*350*	*426*	*418*	*393*	*306*	*302*	*157*	*2352*
		Women	*397*	*529*	*494*	*443*	*327*	*331*	*221*	*2743*
Omnibus 1996		*Men*	*343*	*416*	*384*	*423*	*350*	*273*	*178*	*2365*
		Women	*285*	*531*	*488*	*457*	*318*	*330*	*215*	*2624*
Omnibus 1997		*Men*	*106*	*145*	*130*	*117*	*103*	*85*	*47*	*733*
		Women	*94*	*154*	*160*	*148*	*103*	*76*	*54*	*789*

Source: OMN95, OMN96, OMN97
Checked:

Table 4.6 Self-reported general health using GHS question by survey, age and sex

Adults aged 16 and over ***England***

Age			**16-24**	**25-34**	**35-44**	**45-54**	**55-64**	**65-74**	**75 and over**	**Total**
Survey	**Sex**	**Self-reported general health**	%	%	%	%	%	%	%	%
GHS 1994	Men	Good	78	75	70	64	51	44	36	63
		Fairly good	18	18	23	24	28	36	40	25
		Not good	4	6	6	12	20	20	25	12
	Women	Good	70	72	65	60	53	44	29	58
		Fairly good	25	22	24	26	31	36	44	28
		Not good	5	6	11	14	16	20	27	13
GHS 1996	Men	Good	72	70	70	60	49	43	34	59
		Fairly good	26	26	24	29	34	39	48	30
		Not good	2	4	6	12	17	18	18	10
	Women	Good	61	61	59	53	48	40	29	52
		Fairly good	34	30	31	33	35	42	45	35
		Not good	5	9	10	13	17	18	25	13
Bases = 100%										
GHS 1994		*Men*	*821*	*1217*	*1251*	*1155*	*888*	*845*	*440*	*6617*
		Women	*950*	*1562*	*1379*	*1222*	*974*	*989*	*775*	*7851*
GHS 1996		*Men*	*744*	*1162*	*1104*	*1100*	*803*	*728*	*514*	*6155*
		Women	*833*	*1393*	*1312*	*1220*	*926*	*856*	*694*	*7234*

Source:Ghs94, GHS96,
Checked:

Table 4.7 Self-reported general health by age and sex

Adults aged 16 and over ***England***

Age			**16-24**	**25-34**	**35-44**	**45-54**	**55-64**	**65-74**	**75 and over**	**Total**
Survey	**Sex**	**Self-reported general health**	%	%	%	%	%	%	%	%
HSE 1993	Men	Very good	40	45	42	36	28	23	20	36
		Good	44	42	43	41	39	37	36	41
		Fair	14	12	12	18	25	29	34	18
		Bad	1	1	2	4	6	7	7	4
		Very bad	0	0	1	1	2	4	3	1
	Women	Very good	35	42	39	34	26	24	20	33
		Good								
		Fair	50	44	44	42	43	40	30	43
		Bad	13	12	14	20	25	28	39	20
		Very bad	1	1	2	4	5	6	7	3
		Not good	0	0	0	1	1	2	3	1
HSE 1996	Men	Very good	41	44	45	39	32	25	21	37
		Good	46	43	40	39	35	37	34	40
		Fair	11	10	12	15	21	28	33	17
		Bad	2	2	2	5	9	8	9	1
		Very bad	3	2	3	6	125	10	12	6
	Women	Very good	36	41	41	35	27	21	18	33
		Good	50	45	42	41	40	41	33	42
		Fair	13	12	13	19	24	29	38	20
		Bad	1	2	3	4	7	6	7	4
		Very bad	0	1	1	1	2	2	3	1
Omnibus 1997	Men	Very good	41	46	42	41	33	27	24	38
		Good	43	46	43	38	29	33	33	39
		Fair	15	8	13	15	24	29	25	17
		Bad	2	0	1	4	10	5	15	4
		Very bad	-	-	1	2	4	6	2	2
	Women	Very good	21	40	40	39	29	16	20	32
		Good	55	44	40	37	45	42	24	42
		Fair	21	13	13	20	19	32	36	19
		Bad	2	3	6	2	4	5	16	4
		Very bad	1	-	1	3	2	5	4	2
Bases = 100%										
HSE 1993		*Men*	*1043*	*1513*	*1366*	*1314*	*1076*	*893*	*475*	*7680*
		Women	*1125*	*1745*	*1560*	*1393*	*1126*	*1086*	*821*	*8856*
HSE 1996		*Men*								
		Women								
Omnibus 1997		*Men*	*106*	*145*	*130*	*117*	*103*	*85*	*47*	*733*
		Women	*94*	*154*	*160*	*148*	*103*	*76*	*54*	*789*

Source:Ghs94.in, GHS96.in, 1993 HSE report,
Checked:

Table 4.8 Percentage reporting a long-standing illness by survey, age and sex

Adults aged 16 and over ***England***

Age		**16-24**	**25-34**	**35-44**	**45-54**	**55-64**	**65-74**	**75 and over**	**Total**
Survey	**Sex**	Percentage reporting a long-standing illness							
Omnibus 1995	Men	18	19	31	35	53	53	58	35
	Women	22	21	30	39	52	56	59	37
Omnibus 1996	Men	14	21	26	36	47	52	58	34
	Women	20	20	31	38	43	55	55	35
GHS 1994	Men	21	24	28	37	50	55	61	36
	Women	22	21	27	35	48	56	64	36
GHS 1996	Men	22	26	32	39	55	60	63	39
	Women	24	26	30	41	55	58	68	40
HSE 1993	Men	23	26	30	43	52	64	63	40
	Women	22	28	31	41	56	60	67	41
Bases = 100%									
Omnibus 1995	*Men*	*352*	*424*	*418*	*393*	*304*	*302*	*157*	*2351*
	Women	*397*	*528*	*494*	*443*	*327*	*331*	*222*	*2742*
Omnibus 1996	*Men*	*343*	*416*	*384*	*423*	*350*	*273*	*177*	*2365*
	Women	*285*	*531*	*487*	*457*	*318*	*329*	*215*	*2622*
GHS 1994	*Men*	*971*	*1387*	*1405*	*1254*	*940*	*864*	*452*	*7273*
	Women	*1026*	*1615*	*1422*	*1269*	*995*	*1015*	*816*	*8158*
GHS 1996	*Men*	*893*	*1365*	*1280*	*1242*	*881*	*758*	*537*	*6956*
	Women	*919*	*1468*	*1358*	*1279*	*953*	*877*	*736*	*7590*
HSE 1993	*Men*	*1042*	*1512*	*1366*	*1316*	*1077*	*896*	*474*	*7683*
	Women	*1126*	*1745*	*1559*	*1393*	*1130*	*1091*	*828*	*8872*

Source: GHS94, GHS96, OMN95, OMN96, HSE 93 report

Table 4.9 Percentage reporting a limiting long-standing illness by survey, age and sex

Adults aged 16 and over — ***England***

Age		**16-24**	**25-34**	**35-44**	**45-54**	**55-64**	**65-74**	**75 and over**	**Total**
Survey	**Sex**	Percentage reporting a limiting long-standing illness							
Omnibus 1995	Men	8	10	17	22	36	31	43	21
	Women	11	9	16	24	33	42	47	23
Omnibus 1996	Men	6	13	15	24	34	33	47	22
	Women	12	12	20	24	29	37	45	23
Omnibus 1997	Men	4	15	10	26	41	43	50	24
	Women	9	12	14	26	31	50	56	24
GHS 1994	Men	10	12	15	20	33	38	44	22
	Women	10	12	15	23	29	38	48	23
GHS 1996	Men	10	14	18	24	39	41	49	25
	Women	12	15	20	28	37	39	53	27
Bases = 100%									
Omnibus 1995	*Men*	*352*	*422*	*417*	*393*	*304*	*302*	*157*	*2347*
	Women	*397*	*528*	*494*	*443*	*326*	*331*	*222*	*2741*
Omnibus 1996	*Men*	*343*	*416*	*384*	*423*	*350*	*273*	*177*	*2365*
	Women	*285*	*531*	*487*	*457*	*318*	*329*	*215*	*2622*
Omnibus 1997	*Men*	*106*	*145*	*130*	*117*	*103*	*85*	*47*	*733*
	Women	*94*	*154*	*160*	*148*	*103*	*76*	*54*	*789*
GHS 1994	*Men*	*971*	*1387*	*1404*	*1253*	*939*	*864*	*452*	*7270*
	Women	*1026*	*1615*	*1422*	*1268*	*995*	*1015*	*816*	*8157*
GHS 1996	*Men*	*893*	*1365*	*1279*	*1240*	*881*	*758*	*537*	*6953*
	Women	*919*	*1467*	*1358*	*1277*	*953*	*877*	*735*	*7586*

Source: GHS94, GHS96, Omn95, Omn96,Omn97

Table 4.10 Percentage with a GHQ12 score of four or more by survey, age and sex

Adults aged 16 and over ***England***

Survey	**HSE 1993**			**HSE 1995**		
Sex	**Men**	**Women**	**all**	**Men**	**Women**	**all**
	Percentage with a score of four or more					
Age						
16-24	13	19	16	12	21	17
25-34	13	20	17	12	21	17
35-44	16	19	18	16	21	19
45-54	16	20	18	17	21	19
55-64	10	16	13	14	19	17
65-74	10	13	12	13	15	14
75 and ove	17	19	18	14	20	17
All	14	18	16	14	20	17
Bases = 100%						
16-24	*1016*	*1095*	*2111*	*906*	*1058*	*1964*
25-34	*1480*	*1696*	*3176*	*1372*	*1697*	*3069*
35-44	*1331*	*1527*	*2858*	*1343*	*1477*	*2820*
45-54	*1274*	*1344*	*2618*	*1161*	*1351*	*2512*
55-64	*1018*	*1077*	*2095*	*963*	*1093*	*2056*
65-74	*850*	*1021*	*1871*	*879*	*1014*	*1893*
75 and over	*437*	*720*	*1157*	*485*	*754*	*1239*
All	*7406*	*8480*	*15886*	*7109*	*8444*	*15553*

Source:HSE reports

Table 4.11

Mean EuroQol Tariff by age and sex

Age	*Omnibus 1995 male*	Omnibus 1995 female	*Omnibus 1996 male*	Omnibus 1996 female	*HSE 1996 male*	HSE 1996 female
16-24	*0.93*	0.95	*0.96*	0.94	*0.90*	0.87
25-34	*0.93*	0.93	*0.93*	0.94	*0.91*	0.91
35-44	*0.91*	0.89	*0.91*	0.90	*0.89*	0.86
45-54	*0.87*	0.86	*0.85*	0.87	*0.89*	0.82
55-64	*0.82*	0.81	*0.80*	0.83	*0.79*	0.75
65-74	*0.80*	0.73	*0.81*	0.76	*0.78*	0.76
75+	*0.72*	0.68	*0.72*	0.71	*0.72*	0.73
All	*0.87*	0.86	*0.87*	0.86	*0.84*	0.82

Table 4.12

SF-36 Mean scores by age

Mean Scores				Dimension				
Age	PHYSICAL FUNCTIONING	ROLE-PHYSICAL	PAIN	GEN HEALTH	VITALITY	SOCIAL FUNCTIONING	ROLE-EMOTIONAL	MENTAL HEALTH
16-24	90	90	77	73	65	87	86	75
25-34	92	90	78	75	66	87	87	76
35-44	88	87	76	72	63	87	86	74
45-54	85	80	72	70	63	85	83	75
55-64	72	70	66	64	60	79	80	75
65-74	64	66	68	64	61	82	78	76
75+	51	59	67	62	56	78	77	78
All	80	79	73	69	63	84	83	75

Source: HSE 1996

Table 4.13 : Dependency level by year

Adults aged 65 and over ***England***

Dependency level	Score	1980*	1985*	1991*	1994		
					Men	Women	All
		%	%	%	%	%	%
Independent	1	76	73	77	81	69	74
Least dependent	2	16	19	18	13	21	17
	3	5	6	3	4	8	6
	4	1	1	1	1	1	1
	5	1	1	0	1	1	1
Most dependent	6	0	0	0	0	0	0
Base=100%		*3803*	*3155*	*3201*	*1287*	*1771*	*3058*

* Source: Bone M (1995) Trends in Dependency.

Table 4.14

Percentage with dependency score 1 or 2 (least dependent)

Adults aged 65 or over

Age	GHS:1980	GHS:1985	GHS:1991	GHS:1994
65-69	95	97	97	95
70-74	95	96	97	94
75-79	91	90	94	92
80-84	86	84	92	86
85+	68	70	79	78
Total	92	92	95	92

Table 4.15 - Eta coefficients for EuroQol against SF-36 dimensions

	EuroQol Dimesions				
SF-36 Dimensions	Mobility	Pain / discomfort	Usual activities	Self-care	Anxiety / depression
Physical functioning	0.68	0.52	0.64	0.52	0.27
Physical role	0.54	0.48	0.65	0.41	0.29
Bodily pain	0.52	0.70	0.57	0.37	0.30
General health	0.49	0.51	0.54	0.38	0.42
Vitality	0.42	0.44	0.50	0.33	0.46
Social functioning	0.43	0.44	0.55	0.41	0.49
Role emotional	0.31	0.32	0.41	0.26	0.49
Mental health	0.21	0.27	0.30	0.21	0.62

Table 4.16

Percentage with a Long-Standing Illness, Disability or Infirmity by self-reported general health and survey

Health-state	Omnibus '96	GHS '94	GHS '96
Good/V good*	29%	19%	19%
Fair/Fairly good	72%	53%	51%
Bad/V Bad/Not good#	95%	86%	85%

* 'Very good'+ 'good' for HSE; 'Good' for GHS and Omnibus
# 'Bad'+ 'Very bad' for HSE; 'not good' for GHS and Omnibus

Table 4.17 : Dependency level by self-reported general health

Adults aged 65 and over

England

Dependency level	1	2	3	4	5	6
Self-reported general health	%	%	%	%	%	%
Good	48	13	12	7	20	20
Fairly good	39	39	31	40	20	0
Not good	13	48	57	53	60	80
Base=100%	*2265*	*531*	*193*	*30*	*25*	*5*

Source: GHS 1994

Table 4.18 : Number of ADLs cannot manage by self-reported general health

Adults aged 65 and over ***England***

Number of ADLs cannot manage	0	1	2	3	4
Self-reported general health	%	%	%	%	%
Good	41	12	7	20	20
Fairly good	39	31	40	20	0
Not good	19	57	53	60	80
Base=100%	*2796*	*193*	*30*	*25*	*5*

Source: GHS 1994

Table 4.19 : Dependency level by self-reported morbidity

Adults aged 65 and over ***England***

Dependency level	1	2	3	4	5	6
Whether has a long-standing illness	%	%	%	%	%	%
Yes	48	85	87	97	88	80
No	52	15	13	3	12	20
Base=100%	*2268*	*532*	*195*	*30*	*25*	*5*

Source: GHS 1994

Table 4.20 : Number of ADLs cannot manage by self-reported morbidity

Adults aged 65 and over ***England***

Number of ADLs cannot manage	0	1	2	3	4
Whether has a long-standing illness	%	%	%	%	%
Yes	55	87	97	88	80
No	45	13	3	12	20
Base=100%	*2800*	*195*	*30*	*25*	*5*

Source: GHS 1994

Table 4.21 - Eta coefficients for Long standing Illness against SF-36 dimensions

SF-36 Dimensions	Self-rated general health	Long standing Illness
Physical functioning	0.57	0.42
Physical role	0.52	0.38
Bodily pain	0.49	0.39
General health	0.75	0.46
Vitality	0.53	0.32
Social functioning	0.51	0.3
Role emotional	0.38	0.23
Mental health	0.37	0.19

Table 4.22 Self-reported general health by EuroQol dimensions

Adults aged 16 and over

Mobility

Whether has problems	No probler	Some problems	Severe problems
Self-reported general health	%	%	%
Very good	40	9	17
Good	44	28	17
Fair	14	40	0
Bad	1	17	17
Very bad	0	6	50
Base = 10	*3122*	*685*	*6*

Source: 1996 HSE

Pain and discomfort

Whether has problems	No probler	Some problems	Severe problems
Self-reported general health	%	%	%
Very good	45	19	4
Good	44	40	12
Fair	11	32	33
Bad	1	8	35
Very bad	0	2	16
Base = 10	*2431*	*1238*	*142*

Source: 1996 HSE

Usual activities

Whether has problems	No probler	Some problems	Severe problems
Self-reported general health	%	%	%
Very good	41	9	4
Good	44	30	14
Fair	13	41	33
Bad	1	17	27
Very bad	0	4	22
Base = 10	*3088*	*618*	*105*

Source: 1996 HSE

Self-care

Whether has problems	No probler	Some problems	Severe problems
Self-reported general health	%	%	%
Very good	36	7	5
Good	43	15	35
Fair	17	38	5
Bad	3	29	25
Very bad	1	11	30
Base = 10	*3586*	*203*	*20*

Source: 1996 HSE

Anxiety and depression

Whether has problems	No probler	Some problems	Severe problems
Self-reported general health	%	%	%
Very good	40	19	5
Good	43	37	28
Fair	14	32	29
Bad	2	9	28
Very bad	0	3	11
Base = 10	*2910*	*823*	*80*

Source: 1996 HSE

Table 4.23 Self-reported morbidity by EuroQol dimensions
Adults aged 16 and over

Mobility

Whether has problems	No problen	Some problems	Severe problems
Whether has a long-standing illness	%	%	%
Yes	35	85	100
No	65	15	0
Base = 10(	*3124*	*685*	*6*

Source: 1996 HSE

Pain and discomfort

Whether has problems	No problen	Some problems	Severe problems
Whether has a long-standing illness	%	%	%
Yes	29	67	96
No	71	33	4
Base = 10(	*2433*	*1238*	*142*

Source: 1996 HSE

Usual activities

Whether has problems	No problen	Some problems	Severe problems
Whether has a long-standing illness	%	%	%
Yes	34	65	94
No	66	15	6
Base = 10(	*3090*	*618*	*105*

Source: 1996 HSE

Self-care

Whether has problems	No problen	Some problems	Severe problems
Whether has a long-standing illness	%	%	%
Yes	41	93	90
No	59	7	10
Base = 10(	*3588*	*203*	*20*

Source: 1996 HSE

Anxiety and depression

Whether has problems	No problen	Some problems	Severe problems
Whether has a long-standing illness	%	%	%
Yes	29	67	96
No	71	33	4
Base = 10(	*2912*	*823*	*80*

Source: 1996 HSE

Table 4.24 Euroqol tariffs by self-reported general health and self-reported morbidity

Adults aged 16 and over ***England***

	Omnibus 1995		Omnibus 1996			HSE 1996
	Mean scor	Median sc	Mean scor	Median score		Mean score
Self-reported general health						
					Very good	0.93
Good	0.94	0.95	0.95	0.96	Good	0.86
Fairly good	0.83	0.84	0.82	0.83	Fair	0.71
Not good	0.55	0.57	0.55	0.57	Bad	0.48
					Very bad	0.33
Self-reported long-standing illness						
Yes	0.73	0.75	0.72	0.74		0.7
No	0.94	0.95	0.94	0.95		0.92

Table 4.25 Visual Analogue Scale (VAS) by self-reported general health

	Self-reported general health					
	Omnibus 1995			Omnibus 1996		
VAS	Good	Fairly Good	Not Good	Good	Fairly Good	Not Good
	%	%	%	%	%	%
10	1	1	8	2	3	8
20	0	0	5	0	0	5
30	0	1	10	0	1	13
40	0	3	14	0	3	12
50	2	13	24	1	12	23
60	2	10	11	2	10	11
70	8	17	12	8	18	10
80	22	29	11	22	28	9
90	33	20	4	33	18	6
100	33	7	2	32	9	2
Base=100%	2,900	1,660	496	2,732	1,756	496

Table 4.26 - % at the ceiling for each instrument

Instrument	% at floor	% at ceiling
EuroQol	**0**	**52**
Mobility	0	82
Self-care	0	95
Usual activities	2	81
Pain / discomfort	3	63
Anxiety / depression	1	77
SF-36	**0**	*****
Physical functioning	1	29
Physical role	10	57
Bodily pain	1	33
General health	0	4
Vitality	1	2
Social functioning	1	47
Role emotional	8	61
Mental health	0	3
Long-standing illness	**39**	**61**
Self-rated general health	**1**	**38**
GHQ12	**0**	**55**

Table 4.26a Percentage rating their health as bad or very bad by type of long-standing illness

% with bad or very bad health	Condition
50	
49	
48	
47	
46	
45	
44	
43	
42	
41	Bronchitis/emphysema
40	
39	
38	
37	
36	
35	
34	
33	
32	stroke/cerebral haemorrhage/cerebral thrombosis
31	
30	
29	
28	heart attack/angina
27	
26	Mental illness/anxiety
25	
24	Cancer (neoplasm) ; stomach ulcer/abdominal hernia
23	
22	Diabetes
21	
20	Epilepsy/fits/convulsions ; Arthritis/rheumatism/fibrositis
19	
18	
17	
16	Asthma
15	Cataract/poor eye sight
14	Varicose veins/phlebitis; Back problems/slipped disk
13	
12	
11	Hypertension
10	Poor hearing/deafness
9	
8	
7	
6	Migrane/headaches
5	
4	
3	Hayfever
2	
1	
0	

Table 4.27 Mean EuroQol tariff Score by type of long-standing illness

Mean Euroquol tariff score	Condition	
55		
56		
57		
58		
59	Stroke/cerebral haemorrhage/cerebral thrombosis	
60		
61		
62		
63	Bronchitis/emphysema	
64	Arthritis/rheumatism/fibrositis	
65		
66	heart attack/angina	
67		
68	stomach ulcer/abdominal hernia	
69		
70	Back problems/slipped disk	Cancer
71		
72	Diabetes	
73	Catarac/Poor eye sight	
74		
75	Epilepsy/fits/convulsions	
76	Varicose veins/phlebitis	
77	Poor hearing/deafness	
78	Hypertension	Asthma
79		
80	Migraine/headaches	Mental illness/anxiety
81		
82		
83		
84		
85		
86		
87		
88	Hayfever	
89		
90		
91		
92		
93		
94		
95		
96		
97		
98		
99		
100		

Table 5.28. Mean SF36 dimension score by type of long-standing illness

Mean SF36 profile score	Physical functioning	Role - physical	Bodily Pain
30			
31	Stroke / cerebral haemorrhage / thrombosis		
32			
33		Stroke / cerebral haemorrhage / thrombosis	
34			
35			
36			
37			
38			
39			
40	Bronchitis / emphysema		
41		Bronchitis / emphysema	
42			
43			
44	Heart attack / angina		
45			
46			
47		Arthritis / rheumatism / fibrosis	Mental illness / anxiety
48			
49	Arthritis / rheumatism / fibrosifiz		
50			
51			
52		Cancer (neoplasm)	Epilepsy / fits / convulsions
53			
54			
55	Cancer (neoplasm)	Diabetes	Bronchitis / emphysema
56		Mental illness / anxiety	Stomach ulcer / abdominal hernia
57	Diabetes	Stomach ulcer / abdominal hernia	Migraine / headaches
58		Cataract / poor eye sight	Asthma
59	Stomach ulcer / abdominal hernia	Back problems / slipped disk	Stroke / cerebral haemorrhage / thrombosis
60	Cataract / poor eyesight		
61			
62		Epilipsy / fits / convulsions	Cancer (neoplasm)
63		Poor hearing / deafness	Heart attack / angina
64	Hypertension		Back problems / slipped disk
65	Poor hearing / deafness	Hypertension	Arthritis / rheumatism / fibrosis
66	Back problems / slipped disk	Varicose veins / phlebitis	Diabetes
67	Epilepsy / fits / convulsions	Asthma	Cataract / poor eye sight
68	Mental illness / anxiety		Varicose veins / phlebitis
69	Varicose veins / phlebitis	Migraine / headaches	Poor hearing / deafness
70	Asthma		Hypertension
71			
72			
73			
74			
75			
76			Hayfever
77			
78			
79	Migraine / headaches		
80			
81			
82			
83		Hayfever	
84			
85			
86			
87			
88	Hayfever		
89			
90			
91			
92			
93			
94			
95			
96			
97			
98			
99			
100			

General Health	Vitality	Social Functioning
Mental illness / anxiety		
	Mental illness / anxiety	
	Epilepsy / fits / convulsions	
Epilepsy / fits / convulsions		
Bronchitis / emphysema		
	Bronchitis / emphysema	
	Stomach ulcer / abdominal hernia	
Stomach ulcer / abdominal hernia		
Migraine / headaches	Migraine / headaches	
Asthma	Asthma	
Stroke / cerebral haemorrhage / thrombosis	Stroke / cerebral haemorrhage / thrombosis	
Cancer (neoplasm)	Cancer (neoplasm)	
	Heart attack / angina	
Heart attack / angina	Back problems / slipped disk	
Back problems / slipped disk	Arthritis / rheumatism / fibrosis	Mental illness / anxiety
Arthritis / rheumatism / fibrosis	Diabetes	
Diabetes	Cataract / poor eye sight	
Cataract / poor eye sight	Varicose veins / phlebitis	
Varicose veins / phlebitis	Poor hearing / deafness	
Poor hearing / deafness	Hypertension	Epilepsy / fits / convulsions
Hypertension	Hayfever	Bronchitis / emphysema
Hayfever		
		Stomach ulcer / abdominal hernia
		Migraine / headaches
		Asthma
		Stroke / cerebral haemorrhage / thrombosis
		Cancer (neoplasm)
		Heart attack / angina
		Back problems / slipped disk
		Arthritis / rheumatism / fibrosis
		Diabetes
		Cataract / poor eye sight
		Varicose veins / phlebitis
		Poor hearing / deafness
		Hypertension
		Hayfever

Role-emotional	Mental Health
Mental illness / anxiety	Mental illness / anxiety
Epilepsy / fits / convulsions	
Bronchitis / emphysema	
Stomach ulcer / abdominal hernia	Epilepsy / fits / convulsions
	Bronchitis / emphysema
	Stomach ulcer / abdominal hernia
	Migraine / headaches
Migraine / headaches	Asthma
Asthma	Stroke / cerebral haemorrhage / thrombosis
Stroke / cerebral haemorrhage / thrombosis	Cancer (neoplasm)
Cancer (neoplasm)	Heart attack / angina
Heart attack / angina	Back problems / slipped disk
	Arthritis / rheumatism / fibrosis
Back problems / slipped disk	Diabetes
Arthritis / rheumatism / fibrosis	Cataract / poor eye sight
Diabetes	Varicose veins / phlebitis
Cataract / poor eye sight	Poor hearing / deafness
Varicose veins / phlebitis	Hypertension
	Hayfever
Poor hearing / deafness	
Hypertension	
Hayfever	

Table 4.29 Chi Squared Goodness of Fit statistics for Logistic Regression Models

Longstanding Ilness	Age & self reported general health	Age and EuroQol tariff score	Age and eight SF36 dimension scores	Age Only
Cancer	28.7	25.7	13.6	372.3
Diabetes	27.5	98.7	11.4	579.49
Mental Illness/Anxiety	130.9	46.9	9.5	2085.6
Epilepsy/fits/convulsions	22	23.8	16.2	435.5
Stroke/cerebral haemorrhage/cererbral thrombosis	38.4	11.4	12.9	298.1
heart attack/angina	53.5	103.4	13	802.9
Bronchitis/emphysema	23.9	62.1	5.2	680.7
Arthritis/rheumatism/fibrostis	458.2	115.9	46.2	1659.1

Table 4.30 General Health measures by blood pressure and lung function (age-standardised)

	Blood Pressure		Lung Function	
	Normal	High	Normal	Poor
Mean SF36 dimension score				
Physical functioning	84	73	83	77
Role-physical	82	71	82	74
Bodily Pain	78	73	78	74
General Health	71	59	71	64
Vitality	64	58	64	61
Social Functioning	87	79	87	82
Role-emotional	86	79	86	82
Mental Health	76	72	76	74
Mean EuroQol tariff score	0.87	0.81	0.87	0.83
Percentage with a long-standing illness	41	58	40	56
Percentage rating their health as bad or very bad	4	12	4	8

Table 4.31 Chi Squared Goodness of fit statistics for logistic regression models of blood pressure and lung function

Long-standing illness	Age and long-standing illness	Age and self-rated general health	Age and EuroQol tariff	Age and eight SF36 dimensions	Age only
High blood pressure	54.2	67.7	78.6	67	79.5
Low lung function	300.4	75.1	156.6	23.9	533.1

Appendix B: A description of the main data sources for the secondary analysis

The Health Survey for England

The first Health Survey for England (HSE) took place in 1991, and fieldwork has been carried out continuously since January 1993. A probability sample of people living in private households in England are interviewed about their health and health-related behaviour. Between 1991 and 1995, the survey covered adults aged 16 and over; in 1995, it was extended to include children aged two and over. Approximately 16,000 adults and 4,000 children were interviewed in 1995. The fieldwork for each year runs from January to December.

The HSE has included at least one general health measure since its inception. The interview has included questions on self-reported general health and self-reported long-standing illness in all years. From 1991 to 1995, it carried the General Health Questionnaire 12 (GHQ12); in 1996, both the SF-36 and the EuroQol instrument (without the visual analogue scale) were included in the survey.

The General Household Survey

The General Household Survey (GHS) is a multi-purpose survey which has been carried out continuously since 1971[1]. A probability sample of adults aged 16 and over living in private households in Great Britain are interviewed about a number of topics, including their health and health-related behaviour. Information about children's health is collected from a parent or other adult in the household. Approximately 18,000 adults in 9,700 households are interviewed each year. The fieldwork for each year runs from April to the following March.

Questions on long-standing illness or disability have been included in the General Household Survey since 1971. A question on self-reported general health was introduced in 1976 and has been included in every subsequent year. The EuroQol instrument was included in the interview for the 1996-7 survey. Questions on Activities of Daily Living and Instrumental Activities of Daily Living are included periodically in a section of the interview addressed to respondents aged 65 and over.

The ONS Omnibus Survey

The ONS Omnibus Survey is a multi-purpose survey which started in 1989, for which interviews are carried out at approximately 1900 addresses in Great Britain each survey month. One person aged 16 and over is randomly selected for interview at each address.

The Omnibus Survey frequently includes modules on health, health-related behaviour and Activities of Daily Living. The EuroQol instrument was included in

[1] Fieldwork was not carried out in 1996/7.

the Omnibus interview for the first three months of 1995 and of 1996, together with questions on self-reported general health, long-standing illness and limiting long-standing illness. In 1996, the interview also included questions on some Activities of Daily Living, and use of health services. In May 1997, the survey carried a module which included both the GHS and HSE questions on self-reported general health, together with questions on long-standing illness and use of services.

The ONS Census question-testing programme

In preparation for the 2001 Census, SSD has been carrying out a Census Testing Programme on behalf of Census Division of ONS in order to explore public responses both to general aspects of form design and to selected questions, including questions on self-reported long-standing illness and general health.

Purposive samples of households were chosen for the Test, which targeted sections of the population known from the 1991 Census Validation Survey (Heady et al. 1996) to have experienced particular difficulties with some of the questions in 1991. Each household was asked to complete a Census test form. One to two weeks later, the households were re-visited and residents were interviewed in depth about their experience of filling in the form. Interviews focused on aspects of question design, and on respondents' understanding of, ability and willingness to answer individual items. All interviews were tape-recorded and transcribed for analysis, and interviewers were also asked to summarise the results of individual interviews using Report Forms designed for the purpose.

Qualitative work carried out for the Health Education Monitoring Survey

The Health Education Monitoring Survey (HEMS) is designed to measure health-related knowledge, attitudes and behaviour. The survey is commissioned by the Health Education Authority (HEA) and carried out by Social Survey Division of ONS; the first survey took place in 1995. Approximately 4,700 adults aged 16-74 living in private households in England are interviewed each year. Since its inception, the survey has included questions on self-reported general health and long-standing illness. As part of the pilot study for the 1997 HEMS, the HEA commissioned the Qualitative Methods Unit of SSD to carry out some qualitative question-testing on these questions.

A sample of 14 men and 16 women aged between 18 and 75 was selected to ensure an even mix of age and sex from among the respondents to the 1997 HEMS pilot[2].

[2] In the context of the survey sample, the number of respondents participating in the cognitive question testing is small. However, the aim was not to select a large representative sample, but to explore the meanings and processes which respondents used when answering questions.

Bibliography

Anderson, R.T., Aaronson, N.K., and Wilkin, D. (1993) Critical review of the international assessments of health-related quality of life, *Quality of Life Research*, 2, 369-395.

Barr, R.D., Furlong, W., Dawson, S., Whitton, A.C., Strautmanis, I., Pai, M., Feeny, D. & Torrance, G.W. (1993) An assessment of Global Health Status in Survivors of Acute Lymphoblastic Leukaemia in Childhood. *American Journal of Paediatric Haematology/Oncology* 15(4): 284-290.

Bennett, N. et al (1995) Health Survey for England 1993 (London: HMSO)

Bennett, N., Jarvis, L., Rowlands, O., Singleton, N. and Haseldon, L. (1996) *Living in Britain: results from the 1994 General Household Survey* (London: HMSO)

Berthelot, J.M., Roberge, R. & Wolfson, M.C. (1993) The Calculation of Health-Adjusted Life Expectancy for a Canadian Province Using a Multi-Attribute Utility Function: A First Attempt. In Robine, M.J. Mathers, C.D., Bone, M.R. & Romieu, I. eds. *Calculation of Health Expectancies: Harmonization, Consensus Achieved and Future Perspectives*. Colloque INSERM / John Libbey Eurotext Ltd. 226: 161-172.

Blaxter, M. (1987) *'Self-reported health' in The Health and Lifestyles Survey* (London: Health Promotion Research Trust)

Blaxter, M. (1990) *Health and Lifestyles*. (London: Routledge)

Bone, M. (1995) *Trends in dependency among older people in England*. (London: HMSO)

Bone, M., Bebbington, A.C., Jagger, C., Morgan, K. and Nicolaas, G. (1995b) *Health expectancy and its uses*. (London: HMSO)

Bowling, A. (1991/1997) *Measuring Health: a review of quality of life measurement scales.* (Milton Keynes: Open University Press)

Boyle, M.H., Furlong, W., Feeny, D., Torrance, G.W. & Hatcher, J. (1995) Reliability of the Health Utilities Index - Mark III Used in the 1991 cycle 6 of the Canadian General Social Survey Health Questionnaire. *Quality of Life Research* 4: 249-257.

Brazier J, Jones N and Kind P (1993) Testing the validity of the EuroQol and comparing it with the SF-36 health survey questionnaire *Quality of Life Research* 2 pp169-180.

Brazier, J.E., Harper, R., Jones, N.M.B., O'Cathain, A., Thomas, K.J., Usherwood. T. and Westlake, L. (1992) Validating the SF-36 health survey questionnaire: new outcome for primary care, *British Medical Journal*, 305, pp. 160-164.

Breeze, E. et al (1994) *Health Survey for England 1992* (London: HMSO)

Bridgwood, A. (1993) *Baseline '93: health status and performance monitoring: feasibility study*. (Unpublished paper: Social Survey Division for the Welsh Office)

Bridgwood, A. and Malbon, G. (1995) *Survey of the Physical Health of Prisoners 1994*. (London: HMSO)

Bush, J.W., Chen, M.M., Patrick, D.L. (1972) Social Indicators for Health Based on Function Status and Prognosis. *Proceedings of the American Statistical Association Social Statistics Section*: 71.

Cadman D, Boyle MH, Offord DR, Szatmari P, Rae-Grant NI, Crawford J, Byles J (1986) Chronic illness and functional limitation in Ontario children: findings of the Ontario Child Health Study *CMAJ* 135(7):761-7

Calhoun, H. et al (1996) *Health Survey for England 1994* (London: HMSO)

Cox, D.R., Fitzpatrick, R., Fletcher, A.E., Gore, S.M., Spiegelhalter, D.J., and Jones, D.R. (1992) Quality-of-life assessment: can we keep it simple? *Journal of the Royal Statistical Society*, 155, pp. 353-393.

Crosnick J (1999) Survey Research *Annual Review of Psychology* 50 537-567.

Department of Health (1992) *The Health of the Nation: a strategy for health in England* (London: HMSO)

Dixon, P., Heaton, J. and Long, A. (1994) *Reviewing and applying the SF-36.* (UK Clearing House)

Dolan, P, Gudex, C, Kind, P and Williams, A (1995) *A Social Tariff for EuroQol: results from a UK general population survey*. Discussion paper 138, Centre for Health Economics, York.

Donovan, J.L., Frankel, S.J. and Eyles, J.D. (1993) Assessing the need for health status measures, *Journal of Epidemiology and Community Health*, 47, pp. 158-162.

Essink-Bot ML. Krabbe PF. Bonsel GJ. Aaronson NK. (1997) An empirical comparison of four generic health status measures. The Nottingham Health Profile, the Medical Outcomes Study 36-item Short-Form Health Survey, the COOP/WONCA charts, and the EuroQol instrument. Medical Care. 35(5):522-37, 1997 May

Feeny, D., Torrance, G.W., Goldsmith, C.H., Furlong, W. & Boyle M. (1994) A Multi-Attribute Approach to Population Health Status. In: *1993 Proceedings of the Social Statistics Section of the American Statistical Association. Alexandria, Virginia: American Statistical Association*: 161-166.

Feeny, D.H., Lieper, A., Barr, R.D., Furlong, W., Torrance, G.W., Rosenbaum, P. & Weitzman, S. (1992) The Comprehensive Assessment of Health Status in Survivors of Childhood Cancer: Application to High-Risk Acute Lymphoblastic Leukaemia. *British Journal of Cancer* 67: 1047-1052

Feeny D, Furlong W, Boyle M, Torrance GW (1996) *Multi-attribute health status classification systems. Health Utilities Index*. Pharmacoeconomics 7(6):490-502.

Foster, K., Wilmot, A. and Dobbs, J. (1990) *General Household Survey 1988* (London: HMSO)

Franks P, Gold MR, Clancy CM (1996) *Use of care and subsequent mortality: the importance of gender*. Health Serv Res Aug;31(3):347-63.

Garratt, A., Ruta. D.A., Abdalla, M.I., Buckingham, J.K., Russell, I.T. (1993) The SF-36 health survey questionnaire : an outcome measure suitable for routine use within NHS?, *British Medical Journal*, 306, pp. 1440-1404.

General Household Survey 1972 (1975) (London: HMSO)

Goddard, E. (1990) *Measuring morbidity and some of the factors associated with it', in Health and Lifestyle surveys: towards a common approach*: report of a workshop held on 7 November 1989 organised by the HEA and OPCS. (London: HEA and OPCS)

Goddard, E. and Savage, D. (1994*) General Household Survey: People aged 65 and over*: GHS No. 22 Supplement A. (London: HMSO)

Goldberg, D.P. (1972) *The Detection of Psychiatric Illness by Questionnaire*. Maudsley Monograph No. 21 Oxford: Oxford University Press.

Goldberg, D.P. and Williams, P. (1988) *Users' Guide to the General Health Questionnaire*. NFER-Nelson, Windsor.

Goodchild, M.E. & Duncan-Jones, P. (1985) Chronicity and the General Health Questionnaire *British Journal of Psychiatry* 146: 55-61.

Hays, R.D., Stewart, A.L., Sherbourne, C.D., and Marshal, G.N. (1993) The 'states versus weights' dilemma in quality of life measurement, *Quality of Life Research*, 2, pp. 167-168.

Hill, S. and Harries, U. (1993) The outcomes process: some reflections from research with people in their 60s and 70s, *Critical Public Health*, 4:4, pp. 21-28.

Hollingworth, W., Mackenzie, R., Todd, C.J. & Dixon, A.K. (1995) Measuring changes in quality of life following magnetic resonance of the knee: SF-36, EuroQol or Rosser Index? *Quality of Life Research* 4: 325-334.

Grand A, Grosclaude P, Bocquet H, Pous J, Albarede 1990 Disability, psychosocial factors and mortality among the elderly in a rural French population *Journal of Clinical Epidemiology* 43(8):773-82.

Hunt, A. (1978) *The Elderly at Home*. (London: HMSO)

Idler EL, Angel RJ (1990) Self-rated health and mortality in the NHANES-I Epidemiologic Follow-up Study *American Journal of Public Health* 80(4):446-52.

Jenkinson, C, Layte, R, Wright, L & Coulter A (1996) *The UK SF-36: An Analysis and Interpretation Manual* Health Services Research Unit: Oxford

Jenkinson,-Crispin; Layte,-Richard; Lawrence,-Kate (1997) *Development and testing of the Medical Outcomes Study 36-Item Short Form Health Survey Summary Scale Scores in the United Kingdom: Results from a large-scale survey and a clinical trial* Medical-Care; 1997 Apr Vol 35(4) 410-416

Jenkinson, C., Coulter, A. and Wright, L. (1993a) Short Form 36 (SF-36) health survey questionnaire: normative data for adults of working age, *British Medical Journal*, 306, pp. 1437-1440.

Jenkinson, C. and Wright, L. (1993b) The SF-36 Health Survey Questionnaire, *Auditorium*, 2, pp.7-12

Julious SA, George S, Campbell MJ (1995) Sample sizes for studies using the short form 36 (SF-36) Journal of Epidemiology and Community Health Aug;50(4):473-4

Kahneman D and Tversky A (1972) Subjective Probability: A Judgement of Representativeness. *Cognitive Psychology* 3, 430-454.

Keeney, R. L., & Raiffs, H. (1976) *Decisions with Multiple Objectives: Preferences and Value Tradeoffs*. Wiley New York.

Kind, P (1995) Measuring the reliability of individual assessments of the life quality associated with health states. *Survey Methods Centre Newsletter* Vol. 15 No 2.
Lawton, M.P. and Brody, E.M. (1969) Assessment of older people: self-maintaining and instrumental activities of daily living, in *Gerontologist*, 9, pp. 179-186.

Long, A. (1993) *General Health Measures - an introduction to multidimensional profiles*, Paper prepared for the sub-group of the Chief

Medical Officers' Health of the Nation Survey.

Marmot, M. (Undated) *The reliability and validity of the SF-36 general health survey in the Whitehall II study: inequalities in health status*.

Martin, J., Meltzer, H. and Elliot, D. (1988) *The prevalence of disability among adults: OPCS surveys of disability in Great Britain*; Report 1 (London: HMSO)

McCabe,-C.-J.; Thomas,-K.-J.; Brazier,-J.-E.; Coleman,-P. (1996) Measuring the mental health status of a population: A comparison of the GHQ-12 and the SF-36 (MHI-5) British-Journal-of-Psychiatry; Oct Vol 169(4) 517-521.

McDowell, I and Newell, C. (1987) *Measuring Health: a Guide to Rating Scales and Questionnaires*. New York: Oxford University Press.

McHorney, C.A., Kosinski, M. and Ware, J.E. (1994) Comparisons of the costs and quality of norms for the SF-36 Health Survey collected by mail versus telephone interview; results from a national survey, *Medical Care*, 32, 6, pp. 551-567.

McHorney, C.A., Ware, J.E., and Raczak, A.E. (1993) The MOS -36 item Short Form Health Survey (SF - 36): II psychometric and clinical tests of in measuring physical and mental health constructs, *Medical Care*, 31, 3, pp. 247-263.

Medical Outcomes Trust (1993) *How to score the SF-36 Health Survey*

Meltzer, H., Gill, B., Petticrew, M. and Hinds, K. (1995) *Physical complaints, service use and treatment of adults with psychiatric disorders*. (London: HMSO)

Nunally J (1978) *Psychometric Theory, Second Edition* New York McGraw Hill.

Pluscauskas, M. (1992) *The Measurement of Population Health* (McMaster University, unpublished thesis).

Roberge, R., Bethelot, J.M. & Wolfson, M.C. (1993*) The Impact of Socio-economic Status on Health Status in Ontario*. (Statistics Canada, unpublished).

R A Carr Hill (1992) Health related quality of life Euro style *Health Policy* 20 pp 321-328.

Ruta,-D (1995) "The short form 36 health status questionnaire: Clues from the Oxford region's normative data about its usefulness in measuring health gain in population surveys": Comment. Journal-of-Epidemiology-and-Community-Health; Vol 49(5) 555

Ruta, D., Garratt, A., Abdalla, M., Buckingham, K. and Russell, I. (1993) The SF-36 health survey questionnaire, letter to the *British Medical Journal*, 307, p. 448.

Sundquist J, Johansson SE (1997) Indicators of socio-economic position and their relation to mortality in Sweden. *Social Science and Medicine* 45(12), 1757-66

The Health and Lifestyle Survey (1987) (London: Health Promotion Research Trust)

Thomas, M., Goddard, E., Hickman, M. and Hunter, P. (1994) *General Household Survey 1992* (London: HMSO)

Thomas R. and Purdon S., Survey Methods Centre Newsletter, Vol. 14 No. 2 National Centre for Social Research 1994

Torrance GW, Furlong W, Feeny D, Boyle M (1995) Multi-attribute preference functions. Health Utilities Index Pharmacoeconomics Jun;7(6):503-20

Ware, J.E. (1995a) Self-evaluated transitions in general health, *Medical Outcomes Trust Bulletin*, 3, 2, p. 2.

Ware, J.E., Donald, C.D (1992) The MOS -36 item Short Form Health Survey (SF - 36) conceptual framework and item selection, *Medical Care*, 30, 6, pp. 473-483.

Weinberger,-Morris; Nagle,-Becky; Hanlon,-Joseph-T.; Samsa,-Gregory-P.; et-al (1994) *Assessing health-related quality of life in elderly outpatients: Telephone versus face-to-face administration* Journal-of-the-American-Geriatrics-Society; 1994 Dec Vol 42(12) 1295-1299

Welsh Office (1996) *Welsh Health Survey 1995* (Cardiff: HMSO)

White, A. (1995) *Measuring subjective health status*. (Unpublished paper: Social Survey Division).

White, A. et al (1993) *Health Survey for England 1991* (London: HMSO)

Wilkin, D., Hallam. L., Doggett, M-A. (1992) *Measures of need and outcome for primary health care*. (Oxford University Press)

Ziebland, S. (1995) The Short Form 36 Health Status Questionnaire: Clues from the Oxford Region's Normative Data about its usefulness in Measuring Health Gain in Population Surveys Journal of Epidemiology and Community Health, 49, 102-5.