Listening to respondents: a qualitative assessment of the Short-Form 36 Health Status Questionnaire

Sara Mallinson*

Institute of Public Health Research and Policy, University of Salford, Humphrey Booth House, Hulme Place, Salford M5 4AQ, UK

Abstract

Standardised health status questionnaires are widely used to obtain subjective assessments of health. However, little research has investigated the meaning of the data they produce. Statistical tests will highlight some problems with the structure and wording of a questionnaire but they cannot shed any light on the way in which respondents interpret questions or their intended meaning when they select a response. Various qualitative techniques are being used within disciplines such as sociology and psychology to test both the language of survey instruments and the cognitive bases of surveys. This paper outlines some of these methods and reports findings from a qualitative research study in the UK with a widely used questionnaire - the Short-Form 36 Health Status Questionnaire. The value of including in-depth, qualitative validation techniques in the development and testing of surveys used to collect subjective assessments of health is clearly demonstrated by the findings of the study. © 2001 Elsevier Science Ltd. All rights reserved.

Keywords: Subjective health measurement; Qualitative methods; SF-36

Introduction

During the last decade there has been considerable interest in capturing subjective views of health status and health related quality of life, particularly within the field of health services research. Standardised health status questionnaires are the method of choice for much of this research and considerable energy has been directed towards evaluating and refining them in various settings. After initial piloting ‘post-mortems’ (when interviewers report on the problems they encounter using a questionnaire) the testing process is almost exclusively a quantitative activity. While psychometric assessments have an important role in developing health questionnaires they shed little light on the meaning of the questions and response options to respondents and, therefore, the meaning of respondents’ answers. The validity of survey data depends upon shared understandings of questions and response options and yet research evidence has shown that people interpret survey questions in unexpected ways (Tanur, 1992). In the field of health status assessment there have been some papers which raise questions about the validity of widely used health surveys. For example, Donovan, Frankel, and Eyels (1993) reported that people answering the Nottingham Health Profile (NHP) often misunderstood the NHP’s questions and had problems with the simple yes/no response options and this undermined the validity of the data. Mallinson (1998) found that people self-completing the Short-Form 36 Health Status Questionnaire (SF-36) wrote comments on their questionnaires which suggested their interpretation of some items differed from the surveyor’s intended meanings. They also had difficulty understanding the wording of some items and found some of the response options inadequate to describe their views. The effect of unexpected variations or of flawed design is to create uncertain data whose validity is questionable. Given that health services researchers are keen to encourage
the widespread use of subjective health measurement in health care the problem of meaning has to be addressed. This paper discusses some of the methods which could be drawn into the testing process and illustrates the importance of undertaking further research with qualitative findings from an in-depth assessment of the SF-36.

**Questionnaires and the problem of meaning**

An active industry within health services research in the UK and internationally is devoted to the development and application of subjective health measures. The technical skill and intellectual input which informs these researches is considerable and yet it seems that the issue of meaning is largely ignored in most of the papers and text books produced (Hunt, 1997). Indeed, the complex testing typologies which are being developed to evaluate health questionnaires fail to consider subjective interpretation at all (Donovan et al., 1993). Interestingly, this apparent avoidance of issues around meaning has occurred despite a wealth of sociological debate about the challenges of conducting survey research (Cicourel, 1964; Pawson, 1989) and largely without reference to work by psychologists and survey methodologists working in other fields of research.

Evidence from psychology and sociology has shown that the processes involved in interpreting a question and formulating an answer are complex. For instance, if you reword or reorder questions in a survey peoples’ responses change. If you provide even slightly amended response options then people will give different answers (Clarke & Schober, 1992). This is largely due to the nature of all interviews (structured or unstructured) as interactions and the reliance of questionnaires on natural language. Although standardized questionnaires have been developed to avoid potential sources of bias that arise when questions are reworded, the standardisation of the survey text does not automatically lead to standardisation of meaning. In natural language the meanings of words does not **inhere** in the words themselves but is a **product** of the situation and the relationship between those interacting and can be affected by a range of social and cultural factors. In addition to the problem of natural language there are also other factors related to the cognitive bases of surveys which also have to be taken into account.

Survey methodologists have developed a number of techniques to explore the way people interpret survey questions and to check the cognitive processes which come into play. For example ‘think-aloud’ protocols are often used to explore individual interpretations of questions in detail. People are asked to describe what they are thinking of when listening to each question in a survey and how they interpret the questioners intentions (Hak, 1999; Suchman & Jordan, 1992). The meaning the respondent intends to convey with their reply is also explored through in-depth probing. In some ways this process mimics, albeit in a rather formalised and extended way, the process of natural conversation where people automatically probe to establish intended meaning of speakers or to check that a listener has heard their speech as intended. These interactional techniques to clarify meanings and avoid misunderstanding come naturally to most of us and we are often not even conscious that we use them. Natural conversational flow allows us to repair misapprehensions and establish common ground so that meaningful communication can proceed.

In the interactional strangeness of the survey interview most of the mechanisms used to check meanings are suppressed. The role of questioner and respondent are clearly defined because the format of the interview is predetermined. The direction of the questioning is one-way and (in theory) wordings should not be altered. Indeed, the well trained interviewer should not provide any elaborations to help respondents establish the intended meanings of the surveyor. Of course, the extent to which survey interviewers reach text-book standards in practice is questionable. Nevertheless, where in an ordinary conversation a person might ask questions in response to a question they are unsure of, respondents in the structured interview cannot use the same techniques to assist their interpretation.

The problems with language interpretation and the suppression of conversational repair may present challenges to the respondent who has to make judgements and suppositions in order to comply with the survey. The problem for the survey researcher is that routine data processing methods may not find errors of comprehension or variations in the interpretation of questions which arise. Only if people decide that they are unable to provide an answer will there be missing data and the evidence from cognitive research suggests that many people will respond even if they do not understand a question (Clarke & Schober, 1992). Testing the internal consistency of questions aiming to tap a single construct should highlight inconsistencies in responses and therefore alert the surveyor to comprehension problems (Bowling, 1991). However, there is also evidence that people strive to be consistent when they answer a survey and will therefore choose logically consistent responses even if this does not reflect their views (Clarke & Schober, 1992). Although one could argue that some of these problems

---

1 In this paper the **surveyor** is taken to be the survey author/designer. The **interviewer** is often just a 'technician' and has had little input into the design of the survey. The surveyor is therefore the (absent) questioner, the interviewer their intermediary (see Suchman & Jordan, 1992).
are ironed out in a large sample, at some level, especially if the measure is being used to monitor individual health, the validity of the data rests upon shared meanings or at least stable meanings within the individual across time and place. How can one know if people hear the questions in the manner they were intended without doing some kind of data checks between respondents, or in different situations, or within individuals over time? These are questions which prompted the development of a study which used qualitative data to explore the various interpretations which arose during the administration of the SF-36 Health Status Questionnaire.

Study design

Participants and data collection

This research was conducted in the community with older people who were being treated by community health services in two health authorities in the North West of England. People aged 65 yr or more who were newly referred to two teams of community physiotherapists and one team of rehabilitation occupational therapists during a six-month period were invited to join the study. The therapists provided patients with information on the study and obtained verbal consent for an interview before treatment commenced. The final sample included fifty-six people aged between 65 and 89 yr (mean age 77 yr). 12 (21.4%) of the interviewees were men and 44 (78.5%) were women. All the participants had at least 1 chronic health problem (self-defined) in addition to the one they had been referred for. The median number of health problems was 6 and the range was 1–13.

People were interviewed face-to-face by the author in their own homes on two occasions: first interviews took place before their treatment started and second interviews took place 6 months after this baseline assessment. The questionnaire used in the face-to-face interviews covered socio-demographic characteristics, self-reported health problems, expectations of treatment, the SF-36 health status questionnaire. Face-to-face interviews were tape-recorded and transcribed. To assess the effect of administration technique on completion rates, all participants were also sent a shorter postal questionnaire for self-completion three months after their first face-to-face interview (see Mallinson (1998) for a discussion of the findings).

2 Initially 71 people were involved in the study but 9 died and 6 became too ill to continue and asked to be withdrawn from the study before the final interview.

The Short-Form 36 health status questionnaire

The SF-36 is the most widely used measure of its kind in the UK. In fact it is now a global phenomenon. It has been translated into 45 languages and there is an ongoing international research programme “norming the SF-36 for use in multinational clinical trials” (http://www.sf-36.com/general/iqola.html). Its national and international status is the main reason for its inclusion in this study.

The SF-36 has 36 questions which are scored into 8 dimensions. It aims to assess health concepts that represent basic human values and are relevant to everyone’s health status and well-being and is not limited to use with a specific age, disease or treatment group (Ware, 1992). It can be administered by post or in face-to-face interviews and the manual suggests that it should take about 10 min to complete. It has been used for population health assessment, in clinical trials, and as a patient/doctor communication aid (Jenkinson, Layte, Wright, & Coulter, 1996a). It is designed for use with all populations, although its suitability for use with older people or those with severe illness has been queried (see Hill, Harries, & Popay, 1996; Mallinson, 1998; McHorney, Ware, & Lu, 1994). Overall, the extensive SF-36 literature suggests that it is one of the most reliable and valid short-form questionnaires available. In tests of reliability Brazier et al. (1992) reported excellent test-retest reliability and good internal consistency. Comparison of SF-36 dimensions with other established measures (criterion validity) and examination of how well the SF-36 scores match hypothesised score patterns relating to age, sex and social class (construct validity) have been excellent (Brazier et al., 1992; Jenkinson, Coulter, & Wright, 1993; Jenkinson, Wright, & Coulter, 1994).

The type of validity assessment most pertinent to this paper is that used to establish ‘face validity’. Face validity (do the questions make sense) is often assessed by ‘experts’ — perhaps a panel of survey methodologists or experienced interviewers during pilot studies. Remarkably, despite the SF-36 being a measure of subjective health status, patients have had little direct input into it’s design or validation. In one of the few UK studies to assess the face validity of the SF-36, Jenkinson, Peto, and Coulter (1996b) found two main areas of confusion for respondents. A sub-sample of 50 women from a postal survey of 348 women with menorrhagia who were attending their GP were asked to highlight any problems they had understanding the SF-36. Firstly, women experienced problems with the concept of general health in the context of the research. They were aware that they had been selected for the research because of their menorrhagia and were confused about whether questions about general health should therefore be answered in relation to the
menorrhagia or more broadly. They also experienced difficulty with time-scales in the SF-36 because their particular condition is prone to fluctuations. Jenkinson et al. (1996a, b) suggest that people with no health problems or longstanding limiting illness may have fewer problems thinking about long-term health than people with intermittent problems. I would add that the problem relates not only to time-scales but to the concept of health and the underlying models of linearity and consistency in measures like the SF-36, which do not map onto the complex reality of health experiences for many people.

Some other problems with the content and layout of the SF-36 have been identified and a revised version (SF-36-II) has recently been produced (Jenkinson, Stewart-Brown, Petersen, & Paice, 1999). Changes have been made to the number of response options on some questions (physical and emotional role-functioning, energy/fatigue, and mental health dimensions) and some questions have had their wording altered (physical and emotional role functioning, mental health dimensions). At present most research is still concerned with the original UK SF-36 and the material presented in this paper relates to questions which have not been altered.

Methodological rationale

The qualitative methods used by psychologists and survey methodologists to test survey questions have influenced the aims of the research reported in this paper and the analysis of the data. However, unlike the various forms of cognitive interview, there was no in-depth probing by the researcher during the interview (beyond what was required to code a response). This approach was adopted to try and avoid altering the interview dynamic in any significant way which might affect the study’s comparability with ‘normal’ usage of the SF-36. The exploration of meaning was therefore conducted by the researcher during analysis of the data rather than in collaboration with the respondent during the interview.

In order to allow peoples’ interpretations of questions to be explored the interviews were tape-recorded and then transcribed. The content of the transcripts have been analysed qualitatively using a manual data indexing technique to identify key themes (Spencer & Ritchie, 1994).

This paper reports spontaneous contributions respondents made during their interviews and shows how such material can give insight into the problems people experience. In the section below extracts from the data are used to illustrate key issues ranging from simple technical aspects of question construction to problems with the conceptual basis of questions and implicit notions of linearity and stability in peoples’ conceptions of health. Quotations (showing respondent ID) have been used to illustrate key themes wherever possible, along with additional discussion and case summaries.

Two of the SF-36’s eight dimensions have been selected to focus the discussion. The first dimension is the 10 item physical functioning scale (Table 1) which, on the face of it, seems a very straightforward matrix of questions about the limitations health place on physical activities. The second set of illustrations are drawn from the general health perceptions dimension (Table 2) whose 5 questions aim to establish peoples’ views about their current and future health status.

### Table 1

<table>
<thead>
<tr>
<th>Type of activity</th>
<th>Yes, limited a lot</th>
<th>Yes, limited a little</th>
<th>No, not limited at all</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Vigorous Activities, such as running, lifting heavy objects, participating in strenuous sports</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>B. Moderate activities, such as moving a table, pushing a vacuum cleaner, bowling or playing golf</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>C. Lifting or carrying groceries</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>D. Climbing several flights of stairs</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>E. Climbing one flight of stairs</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>F. Bending, kneeling or stooping</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>G. Walking more than a mile</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>H. Walking half a mile</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>I. Walking 100 yards</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>J. Bathing and dressing yourself</td>
<td>1</td>
<td>2</td>
<td>3</td>
</tr>
</tbody>
</table>
Findings

The physical functioning dimension

The questions on physical activity included in the SF-36 (Table 1) aim to represent distinct aspects of physical activity and the different severities with which limitations might be experienced. The authors define physical functioning as the ‘performance or capacity to perform a variety of physical activities normal for people in good health’. (Stewart & Kamberg, 1992, p. 86). They have attempted to separate simple physical activity from role activities so that the assessment shows how limited people are regardless of their life situation. It therefore asks about self-care activities (bathing and dressing) and other physical activity (such as walking, climbing stairs, lifting). The three tier response options were included to allow the measure to detect the degree of limitation — not just whether someone can do an activity or not. Although the questions seem fairly simple and the three tier responses options are not complex, a number of problems became apparent during the course of the research.

Double questions

It is one of the most basic principles of question construction that they should never contain a double question (Gerber & Wellens, 1997). The SF-36 has several because there is a tendency to include several illustrative activities in one question for the sake of brevity. For example, the items below (3F and 3J respectively) ask:

Does your health limit you in these activities?

IV: Can you bend, kneel or stoop?
IE: I can bend but I can’t kneel. (wo20)

When a survey contains double questions it places an unnecessary and inappropriate coding responsibility on people who have no guidance on how to deal with these issues. They have to make a choice between ignoring some of the illustrative activities and answering in relation to one or averaging out their problems by selecting the middle response option. Respondents had different ways of dealing with the dilemma but quite often they would select ‘limited a little’ in order to average out the problems. The uncertainty people expressed and the decision making processes they used are lost in the coding process.

Unfamiliar terms or phrases

Words or phrases which may appear meaningful to survey designers may not have common currency in other cultures or sub-cultures. Sometimes this is a quite subtle problem. On the physical functioning scale several questions ask people to judge how far they could walk (questions 3G, 3H, 3I). The questions use familiar distances but people often asked for clarification. In everyday conversation people did not talk in measured distances but in terms of landmarks. They wanted me to provide those landmarks before they could make a judgement.

IV: Can you walk more than a mile?
IE: (pause) Well I don’t know.
IV: You don’t?
IE: How far is a mile?
IV: I don’t know, erm, probably from here to Tesco and back.
Although additional clarification should not be provided in the highly standardised interview, intervention was necessary in order to progress the interview. This may not be within the ‘rules’ but it is often the only solution if one wants to proceed without alienating the respondent. However, in many surveys, especially if people are given questionnaires for self-completion, respondents would have to make a guess. An appreciation of how accurate these guesses are requires further research, but survey methodologists would point out that surveys should not force people into guesswork as it is unreliable (Gerber & Wellens, 1997).

Vague questions

Linked to the above comments, the data also raised some concerns about the vagueness of the questions. Although everyday language is filled with vagueness people use their knowledge of the situation and common ground to fill-in information not explicitly given in a verbal exchange. The structured interview may restrict their sense-making activities. It is unclear how many people now have experience of survey interviews but it is becoming an increasingly popular medium for public consultation. It is likely that people have an expectation of what a survey entails. They are self-consciously aware that it has an order and theoretical linkages which if not immediately apparent, must still be present. They strive to see consistency and be consistent themselves but when faced with vague terms this may prove difficult. In ordinary circumstances people would be able to ask for clarification but as we noted before, in the standardised survey interview a trained interviewer should not offer clarification.

Looking again at the walking items (3G-I), a number of respondents pointed out the problem of providing a reasonable response to a question which gave no contextual information about the circumstances of the walking in which the surveyor wanted the assessment to be framed. Was it hill walking, walking in their street, walking round shops?

IV: Can you walk half a mile?
IE: Where’s half a mile?
IV: Say down to the garden centre, maybe a little bit further than that.
IE: I can walk down to the garden centre but there’s no way I could get back because it’s up-hill, and as soon as I, I can’t walk up that hill so it depends which, if you’re talking about on the flat, slowly, not talking or carrying anything... I can walk around the shopping precinct and round the supermarket because you’re going slowly and you’re stopping and looking at things and you’re not talking to anybody.

IV: What about 100 yards? Would you say you are very limited over 100 yards or just a little bit limited.
IE: 100 yards on the flat, fine. 100 yards up a hill — I wouldn’t tackle it. (sp01)

Similarly, people took issue with the lack of context for the question on lifting and carrying (3C). Their concerns often revolved around whether the question referred to a few groceries (i.e. those likely to be gathered in a quick call in the local shop) or the heavy load likely from a weekly shopping trip.

IV: How about lifting or carrying groceries?
IE: Well I can carry a bag but I can’t carry heavy things? (wp04)

A slightly different problem also emerged in relation to the wording of the question about bathing (3J). Many people took the wording quite literally and referred to using the bath rather than other forms of washing/personal hygiene. For example one person pointed out:

IE: I couldn’t manage a bath. I have a shower. But fortunately I have a seat in the shower. I can sit down (sp16)

Lengthy and detailed questions can be tiresome and may create too much respondent burden, thereby lowering response rates and making them impractical for some routine applications. However, brief questions create another set of problems for respondents and researchers because they are not always easy to answer and the responses they generate are certainly not easy to interpret.

Normative assumptions

Taken as a whole the physical functioning scale sets up a normative level of activity which is well beyond the capacity of most of the participants in this study. Two main problems arise:

- Firstly, in terms of the measurement, there is a floor effect whereby people are relegated to the margins of the measure. It will therefore be relatively insensitive to differences between people which could nevertheless be experientially relevant and especially insensitive to small changes in status within individuals. The SF-36 authors have noted this problem (Ware & Sherbourne, 1992)
- Secondly, such questions may also seem irrelevant to people. This is exacerbated by apparently unhelpful
response options. Evidence from experimental research with survey respondents has shown that people will try to answer questions no matter how difficult they find it (Tanur, 1992). This is partly because respondents experience pressure to collaborate once they have agreed to take part but it is also related to cognitive processes which people use to make sense of the survey. People will try to impute meaning and relevance, structure and order to questions in a survey on the principle that the survey designer is a reasonable professional, that they have thought carefully about what they want to ask, and that if no additional information is supplied then the meaning must be self-evident to most people (Clarke & Schober, 1992). When we look at the response options supplied for the physical functioning scale we can see that there is no ‘opt-out’ response provided, for example ‘I do not do this’ or ‘I don’t know’. The assumption is that everyone either does these activities or wishes to do them. In fact, many of the people I listened to did not do vigorous or moderate activities, they did not climb stairs, do their own shopping or go for walks. Although the selection of ‘limited a lot’ response was a reasonable reflection of their physical functioning, it was a very demoralising experience to find that one could do nothing implied to be ‘normal’. For others non-participation was actually a matter of choice and they felt dubious about having to specify a limitation as examples of responses to questions 3G and 3C illustrate.

IV: How are you for walking more than a mile?
IE: Well I haven’t tried it really. I’m lazy. I get in the car. (sp12)
‘IE: Oh well I haven’t carried a lot of groceries. My husband always comes with me for groceries so he carries them. (wp09)

When one tries to squeeze answers into response options which do not quite fit two problems may arise. In circumstances where respondents are completing a survey by themselves a poor fit between questions and response options may lead to higher levels of missing data, although many people will try to answer despite such problems. In a face-to-face interview, rather than refusing to answer, respondents may begin to disengage from the interview. Surveyors depend upon people taking the survey seriously and providing meaningful and realistic (though quickly considered) responses. The constrained interaction of the interview presents a challenge to the flow of normal communication anyway. If one also adds questions that seem irrelevant and poorly conceived the interviewee may feel less inclined to co-operate. Once an interview has started it is rare for interviewees to ask for the process to be stopped altogether but they may show signs of irritation, boredom or withdraw intellectually. Interviewees in this study displayed many of these signs of disengagement and some were assertive enough to vocalise their concerns about the relevance of questions.

Most of the problems discussed above can be traced directly to problems with construction of items and response options and it is surprising to find such elementary flaws in an instrument which has been in use for over 8yr in the UK (Jenkinson et al., 1999). Before the discussion moves on to some of the challenges associated with researching conceptually contested areas such as health and illness, problems around judgments of severity which arose in the study will be outlined.

Relativism: the problem of adaptation to limitation and response-shift

It has been noted that as peoples’ health changes over time so do the scales by which they make judgements (beta change), and occasionally their entire conceptualisation of a target concept (gamma change) (Allison, Locker & Feine, 1997). This is referred to as response shift (Schwartz & Sprangers, 1999). Response shift is a particular problem when one is doing longitudinal research on changes in health status because the implication is that changes might not be ‘real’ (alpha) but forms of beta or gamma change instead (Hak, 1999). Although the main purpose of this study was not to see how well the SF-36 measured changes over time (which is when problems around response shift become really interesting) if one is trying to assess meaning equivalence of questions across a group or a population these processes of concept reconceptualisation and the recalibration are also important.

After the onset of illness and disability many people adapt to their physical limitations and find new ways of achieving their objectives. Some of the comments made by respondents hinted at adaptive changes and a recalibration of their judgements about severity of limitation. In the illustration below a man who described himself as having ‘some limitations’ in climbing stairs (3E) describes the climbing technique he has developed over time.

IV: Can you climb a flight of stairs?
IE: Err, oh yes I can do. Like everything else I have a system. Stand at the bottom of the stairs (of course there are times when I forget) breathe in and out on each step and I can manage it. But sometimes if I go somewhere and I haven’t climbed stairs for a while I
find I am half way up the stairs and I forget the system.

IV: That’s because you’ve spent most of your life charging up the stairs?

IE: That’s right, but there again I always think it’s nice at time to forget rather than having it on your mind all the time. It’s nice to forget. But the stairs, I can manage them. (sp16)

As an external observer I would have rated him as very limited and although he made a different assessment on this occasion, contextual information he gave during the interview suggested that his current perspective reflected his adaptation to his health state. This illustrates two points: firstly, it serves to illustrate why self-assessment is important because the interviewer and respondent had different opinions formed on the basis of differently calibrated judgement scales; secondly, it raises questions about the relativism of peoples’ severity ratings over time as circumstances change. If we wish to understand subjective perceptions of health further research is needed to explore the extent to which variations such as these occur within and across individuals.

General health perceptions dimension

There are numerous technical flaws in the physical functioning dimension but other more subtle problems arose when items belonging to the general health perceptions dimension were asked. Health is difficult to define — in fact it is probably one of the best examples of an essentially contested concept (Bryant, 1995). Where any attempt is made to define it the talk often revolves around what it is rather than what it is. Extensive research within the social sciences has shown the depth and breadth of concepts used by lay people to describe states of health and how the representations people use to explain it are shifting (e.g. Herzlich, 1973). A questionnaire like the SF-36 aims to reduce complexity into some ‘essential’ representation — the very core of those beliefs.

The SF-36 asks a total of 5 questions about general health perceptions (Table 2). The authors state that the items are designed to tap the following aspects of health beliefs; general health perceptions (Qu 1.) resistance to sickness (10a): current health (10b and d): health outlook (10c) (Ware & Sherbourne, 1992). The whole process of measuring perceptions of health rests upon the stability of meaning of terms such as ‘illness’ or ‘healthy’ within individual accounts over time and also across different social groupings. However, the evidence from ethnographic research using life-histories and biographies suggests that people do not hold static and unresponsive concepts. Not only do peoples’ views vary because of who they are and where they are, they may also vary because of the purpose of the account they are giving and how they are asked about their views (Cornwell, 1984).

This study has not allowed the exploration of how different settings and different contexts for the research affect the assessments people make but it has highlighted the variations which occur within and across accounts. Two illustrations from the general health perceptions dimension will be presented: Firstly, when asked to rate their health in general (qu1), people seemed to make self-assessments in different ways. Some made explicit social comparisons but others did not; Secondly, when asked about getting ‘ill’ (qu 10a), peoples’ conceptualisation of illness varied in important ways. Some people excluded certain types of illness from their assessments while other took those types of illness into account.

Comparative health

When people were asked to rate their general health a large proportion explicitly used some kind of comparator to make their judgement. However, the type of comparisons made varied and this had an important impact on the responses selected.

It has been noted in psychological research that most peoples’ self-conception is formed in comparison to peers (Gibbons, 1999). In particular, we are likely to compare ourselves to people of the same age. This process is indicated in the comments of some respondents answering Question 1 (See Table 2) who spontaneously refer to age as a key criteria for comparison although the SF-36 does not instruct them to do this.

IE: Well I wouldn’t say I was excellent but... well I’d say for my age, you know I’m 80, turned 80.

IV: Really?

IE: Yeah, so really you don’t know whether its say very good for your age or just good. (wpo2)

This person decided to bound her comparison using age, although she points out that it is not clear whether the surveyor expects this to be done or if they want a comparison to a ‘general population’. By contrast some other respondents adopted a more general perspective and appeared to compare themselves to a general population. In addition some people used an internal comparison, comparing themselves now with their memory of health states when they were younger or before the onset of health problems. Of course, who you compare yourself to (yourself, your peers, or the general population) is likely to have an important impact on your self-rating. One interviewee grasped fully the range of possibilities available to her in interpreting this question.

IE: There are people worse, so I’d say fair. But you see again, If somebody is dying with cancer I’m excellent but compared with you I’m very poor...
It’s like being at an interview panel. You aren’t sure which answers they want you to give! (sp01)

**Conceptualising health: inclusions and exclusions**

Other factors which affected the survey process arose because of the complexity of the dimensions of ‘health perception’ which the SF-36 attempts to assess. The interviewees often gave brief contexts for their responses and in some instances attention to this contextual detail meant that apparently paradoxical responses (for example when people give positive self-ratings in the face of quite severe limitations) become more meaningful. The following extract is taken from an interview with a man who was housebound and who struggled to walk or do any self-care or activities of daily living. Initially he states that his health is good before qualifying that assessment because of his age.

IE: My health is good. Its the spinal atrophy that’s the problem.
IV: Is your health very good, good, fair.
IE: Well I’d say general health, at my time of life, you can’t say very good, I’m 88 you see, so you can’t say your health is good then. I’ll say fair. (wp22)

He separated his specific chronic health problems from his sense of being a ‘healthy’ person, which is what he thought the surveyor was interested in. This decision was also related to his assumption that the surveyor would want his age to be taken into account. This kind of response was not an isolated event. For example, another respondent with similar chronic health problems said

IE: I am feeling quite well in myself — except for these blooming bones. (sp07)

When asked to rate their health status people can often make surprising decisions but that does not mean they have misunderstood the question or that they hold unrealistic self-perceptions. The problem is that they include and exclude problems in different ways and may unexpectedly weigh all kinds of information before answering. In effect, people are responding from different premises to each other and from the surveyor. This inevitably affects respondents’ intentions/meanings in selecting a particular response option and makes it difficult for the surveyor to interpret their answers.

This complex problem is further illustrated by considering the way respondents reacted to the SF-36 question about resistance to sickness (Table 2, Part 10a). This question asks people whether they think they get ill more easily than other people. There were clear differences in the way people interpreted this question which again raises concerns about the meaning equivalence of questions across groups. A significant proportion of the interviewees thought the term ‘ill’ only referred to everyday problems such as coughs and colds. They therefore discounted other types of illness such as chronic health problems, particularly things they related to age and things that they no longer found ‘bothersome’. This tendency has been noted in other social research settings (Blaxter, 1990). The quotes below illustrate some of the ways in which illness was defined.

IE: Well I don’t think I get ill. They are just common complaints aren’t they? (wp07)
IE: I don’t know what category it comes under really. I’m not bodily ill, its mobility more than anything isn’t it?
IV: You don’t think you get ill more easily than other people?
IE: No, no. It’s just mobility (wp18)
IE: I don’t think so — apart from my diabetes and it doesn’t bother me that. I’ve got used to it. (wp09)
IE: I think it’s just old age. I shake but it’s not bothering me. I’m not letting it bother me. (w008)

There are two main concerns which need to be addressed. Firstly, researchers need to be aware of the tendency to exclude some types of health problem and to consider whether this leads to an underestimation of the limitations some people experience. Secondly, the evidence from this study suggests that there is little consistency in peoples’ approach to these questions. Some people included age-related problems, or ‘mobility’ when they assessed their susceptibility to illness but others did not. The meanings of the aggregated data are therefore uncertain and unless some contextual data are retained, the differences in individual interpretations will be lost when the questionnaire is scored.

**Conclusions**

The practice of health measurement has a long history and considerable energy has been devoted to establishing the conceptual basis for many health questionnaires. However, critics have noted that with the recent upsurge of interest in health measurement some of the more detailed, in-depth research seems to have been abandoned (Hunt (1997) calls it the ‘rush to measurement’). It is easy to fall into the trap of using questionnaires like a form of laboratory equipment (a kind of calibrated dipstick of health) and to forget that, like most social research tools, they are open to interpretation. Unfortunately the continuing emphasis on quantitative testing technologies means that the problem of subjective meaning and interpretation is unlikely to be fully addressed. In fact, I would argue that current practices in validating health
status questionnaires are becoming ritualistic (Gouldner, 1967). They seem to 'exalt the method' without genuinely appraising its ability to produce the knowledge required. As this paper has shown, the validation of the most widely used health status questionnaire, the SF-36, has failed to identify some important problems. Most SF-36 research has relied upon quantitative, psychometric methods without any appraisal of what it means to patients and whether it aids communication processes or systematically distorts them.

The issue of meaning is absolutely central to understanding subjective views and without more assessment of peoples’ understandings of survey questions it is difficult to see how one can establish their validity as subjective health measures. The study reported in this paper is only a starting point for more detailed and comprehensive qualitative validation research. For example, this research has considered the problems encountered by older people with chronic health problems. Further research should explore whether there are any systematic differences between social groups in the interpretation of questions. By drawing in knowledge and expertise from different research disciplines such as sociology and psychology the science of health measurement could be advanced quickly and effectively. In particular more research effort could be channelled into the development of qualitative testing protocols for existing health questionnaires. At the same time it would also be interesting to consider whether the highly standardised questionnaire is the most appropriate way to assess subjective views or whether other forms of survey, perhaps where interviewers could use discretion and be responsive to context, would be better. However researchers wish to proceed, it is certain that the problem of meaning needs to be addressed if they want to hear peoples’ views about their health and the outcome of health care.

Acknowledgements

I would like to thank NWRHA for providing the Health Services Research Training Fellowship which funded this research and Jean Siddall for her assistance. I am grateful to Jennie Popay, Rob Flynn and Sharon Bennett for their comments on drafts of this paper and to Gareth Williams for his supervision. Particular thanks go to all the staff and patients who gave up time and energy to participate in the study.

References


