Chapter 1: INTRODUCTION

Several important government policies and initiatives form the context for this set of reviews. The NHS Plan set in place a process of reforms to develop services designed around the patient (NHS Plan, 2000). The NHS Improvement Plan (Secretary of State for Health, 2004) placed special emphasis on the need to develop services more appropriate for individuals with long-term conditions. More recently, the White Paper, Our Health, Our Care, Our Say (DH, 2006) sets out plans to make health and social services more responsive to patients' and users' needs, choices and preferences. An enormous range of developments have been planned to allow individuals with long-term conditions to avoid unnecessary hospitalisation, reduce dependence on the acute care model for services, maintain independence in the community, promote self care and control over their lives and services and reduce disabilities and disadvantages arising from chronic illnesses. All of these ambitious plans require evidence directly from patients and the public that services are having a positive impact in relation to their experience of long-term conditions and of services. With around 6 in 10 adults reporting some form of chronic condition and these individuals making greater use than others of health and social services, there is enormous scope for evidence of patients' and users' experiences to make a difference to the quality of care and to the quality of lives of those with long-term conditions.

At the same time there has been growing recognition of the real absence of evidence outputs and outcomes of public services generally and the NHS specifically. Traditionally the productivity of services has been measured by indicators that might be better thought of as inputs, numbers of procedures carried out, numbers of consultations and admissions etc. The Office of National Statistics commissioned a review of public service performance and productivity that highlighted this lack of evidence and called for the development and use of better measures of outcome to inform decisions about the productivity of public services (Atkinson, 2005). The scope for patients and users directly to report judgements of outcome in relation to health services was emphasised.

The enormous array of patient-reported outcome measures that have been developed over the last thirty years offers clear opportunities to involve patients and users directly in judgements of the outcomes of services. Variously termed measures of 'health status', 'health-related quality of life', 'functional status', 'patient-reported outcome' or often just 'outcome', the common element is an attempt directly to capture the patient's experience of important aspects of health through questionnaire or interview. Considerable resources and effort have been invested to make such 'instruments' valid measures for use in relation to a wide range of decisions and policies in health. One principle problem is that there are large numbers of such instruments from which to choose for any given health problem or context and insufficient guidance to inform choice (Garratt et al., 2002).

Patient-reported health instruments usually take the form of questionnaires containing several items reflecting the broad nature of health status, disease, or injury, which are most often summed to give a total score. The term 'patient-reported health instrument' will be used throughout this review to refer to patient-completed instruments.

There are two broad categories of patient-reported health instrument: generic and specific. Generic instruments are not age-, disease-, or treatment-specific and contain multiple concepts intended to be relevant to a wide range of patients and the general population. Specific instruments may be specific to a particular disease (for example, diabetes), a patient population (for example, older people), a specific problem (for example, pain), or a described function (for example, activities of daily living). Disease-specific instruments may have greater clinical appeal due to their specificity of content, and associated increased responsiveness to specific changes in condition.

The broad content of generic instruments enables the identification of co-morbid features and treatment side-effects that may not be captured by specific instruments, which suggests they may be useful in assessing the impact of new health-care technologies where the therapeutic effects are uncertain. However, the broad content may reduce responsiveness to small but important changes. It has therefore been recommended that a combination of generic and specific measures be used in the assessment of health outcomes.

Patient-reported health instruments have been increasingly applied in a range of settings including routine patient care, clinical research, audit and quality assurance, population surveys, and resource allocation. However, consensus is often lacking as to which instrument to use; this has important implications for the evaluation of clinical effectiveness. Structured reviews of measurement properties are a prerequisite for instrument selection and standardisation, and instruments with measurement properties that support their application in specific populations and across a range of evaluation settings need to be identified.

Selection criteria have been defined for assessing the quality of patient-reported health instruments (Streiner and Norman, 1995; McDowell and Newell, 1996; Fitzpatrick et al., 1998). These include measurement issues, such as reliability, validity, responsiveness, and precision, as well as practical issues, such as acceptability and feasibility. These criteria are briefly summarised since they directly inform the reviews reported here.

Criteria for assessing patient-reported health instruments

Reliability is concerned with whether measurement is accurate over time and, for multi-item instruments, whether they are internally consistent. Test-retest reliability usually involves instrument self-completion on two occasions separated by a suitable time-period and, assuming no change in the underlying health state, measures the temporal stability of the score (Fitzpatrick et al., 1998). A test-retest period of between two days and two weeks has been recommended for most conditions (Streiner and Norman, 1995). Too short a period may be associated with patient recall of answers, which may artificially inflate reliability (Nunnally and Bernstein, 1994; Streiner and Norman, 1995); too long a period may be associated with actual change in health.

Health transition questions, which invite patients to indicate whether their general or specific health has changed between instrument administrations, are often included in evaluations. The correlation coefficient is the most frequently used method for calculating estimates of test-retest reliability; the intra-class correlation coefficient

(ICC) is used to identify group shift over time as a measure of reliability (Streiner and Norman, 1995). For group comparisons, levels of reliability over 0.70 are required (Streiner and Norman, 1995; Fitzpatrick et al., 1998). For the evaluation of individuals, levels above 0.90 have been recommended (Nunnally and Bernstein, 1994; Fitzpatrick et al., 1998).

Internal consistency reliability of multi-item instruments that adopt a traditional summated rating scale format is tested following a single application. The relationship between all items, and their ability to measure a single underlying domain is assessed using Cronbach's alpha: alpha levels of between 0.70 and 0.90 have been recommended (Nunnally and Bernstein, 1994; Streiner and Norman, 1995; Garratt et al., 2001). Homogeneity at the item level can be assessed using item-total correlation: levels above 0.40 have been recommended (Ware, 1997).

Validity assesses whether an instrument measures what is intended in the different settings in which it may be applied (McHorney, 1996; Fitzpatrick et al., 1998). Instrument validity is not a fixed property. The process of validity testing is ongoing, informing instrument application and interpretation in different settings and with different populations (McHorney, 1996; Ware, 1997). Hence, new and refined instruments, and those applied in different settings or with different populations require evidence of validity. Both qualitative and quantitative methods can be used to assess validity.

Face and content validity require appraisal of item content, and assessment of its relationship to the instrument's proposed purpose and application (Fitzpatrick et al., 1998). Methods of item generation and instrument development may influence this assessment. Literature reviews, theoretical propositions, and interviews or focus groups with patients or health-care professionals may all inform this process. However, for patient-reported instruments to have content validity and relevance to the recipients of care, patients should be involved in item derivation (Fitzpatrick et al., 1998).

The quantitative assessment of validity requires comparison of the scores produced using patient-reported health instruments with those derived from other measures of health, clinical, and socio-demographic variables. Patient-reported instruments measure hypothetical constructs which are by definition non-observable, for example, HRQL and pain, and address a more general hypothesis than that supported by a specific behaviour (Nunnally and Bernstein, 1994). However, by reference to established evidence and the instrument's underlying theoretical base and item content, quantifiable relationships with a range of other instruments and clinical and socio-demographic variables can be expected (Ware, 1997; Fitzpatrick et al., 1998).

Expected correlations between variables should be presented to allow validity to be disproved (McDowell and Jenkinson, 1996). The strength of correlation between variables, be they small (less than 0.30), moderate (less than 0.50), or large (greater than 0.70), indicates that the instrument measures the construct in a manner founded on theory or established evidence (McHorney et al., 1993). For example, two patient-reported measures of functional disability with similar content would be expected to correlate strongly. Construct validity may also be assessed using 'extreme groups', which theorises that one group will possess more or less of a construct (Streiner and

Norman, 1995). For example, compared to the general older population, older people who are hospitalised following a hip fracture may be expected to report greater pain and worse HRQL.

The dimensionality or internal construct validity of a multi-item instrument can be assessed using factor analysis or principal component analysis. Principal component analysis can be used to assess the underlying structure of a multi-item instrument through the identification of components, or domains, into which items may group (McDowell and Newell, 1996). This form of analysis adds empirical weight to a hypothesised domain structure. For example, principal component analysis has supported the hypothesised eight-domain structure of the SF-36 (McHorney et al., 1993).

Responsiveness is considered a necessary measurement property of instruments intended for application in evaluative studies measuring longitudinal changes in health (Beaton et al., 2001; Liang et al., 2002). The numerous approaches to evaluating responsiveness have recently been reviewed by a number of authors (Liang, 1995; Wyrwich et al., 2000; Beaton et al., 2001; Liang et al., 2002; Terwee et al., 2003).

Responsiveness has been described as the ability of an instrument to measure clinically important change over time, when change is present (Deyo et al., 1991; Fitzpatrick et al., 1998). It has also been argued that responsiveness can be viewed as longitudinal validity or as a measure of treatment effect (Terwee et al., 2003). Patient-reported health instruments have had by far the greatest application in clinical trials and most of the literature on responsiveness relates to the measurement of change in health for groups of patients (Fitzpatrick et al., 1998).

There are two broad approaches to assessing responsiveness: distribution-based and anchor-based (Wyrwich et al., 2000; Norman et al., 2001). Distribution-based approaches relate changes in instrument scores to some measure of variability, the most common method being the effect size statistic. The three widely-reported effect size statistics use the mean score change in the numerator, but have different denominators (Fitzpatrick et al., 1998). The effect size (ES) statistic uses the standard deviation of baseline scores (Liang, 1995). The standardised response mean (SRM) uses the standard deviation of the change score to incorporate the response variance in change scores. However, both the ES and SRM may be influenced by natural variance in the underlying state and by measurement error. The modified standardised response mean (MSRM), or responsiveness index, addresses the inherent natural variance that may occur in patients who otherwise report their health as unchanged, and non-specific score change by using the standard deviation of change in patients who are defined as stable (Deyo et al., 1991). In demonstrating responsiveness to clinically important change, instruments should detect change above the non-specific change incorporated in the MSRM (Deyo et al., 1991).

It has been suggested that statistical measures of responsiveness are an insufficient basis for assessing responsiveness and that patients' views on the importance of the change should inform testing (Liang et al., 2002; Terwee et al., 2003). Anchor-based approaches assess the relationship between changes in instrument scores and an external variable (Norman et al., 2001). This includes health transition items or global

judgements of change used to estimate the Minimal Important Difference (MID), the instrument change score corresponding to a small but important change (Jaeschke et al., 1989; Juniper et al., 2002). The MID can inform sample size calculations but consideration must be given to specific groups of patients and specific settings (Terwee et al., 2003). Score interpretation may be improved through the provision of evidence relating to score variation (Terwee et al., 2003) or a score range against which real change may be assessed (Streiner and Norman, 1995; Beaton et al., 2001).

External variables including transition ratings have also been compared to instrument score changes using correlation. This form of longitudinal validity (Kirshner and Guyatt, 1985; Terwee et al., 2003) assesses the extent to which changes in instrument scores concord with an accepted measure of change in patient health (Deyo et al., 1991; Fitzpatrick et al., 1998).

The ability of an instrument to distinguish clearly and precisely between respondents in relation to reported health or illness is referred to as **precision** (Fitzpatrick et al., 1998). Ideally, items within an instrument should capture the full range of health states to be measured, supporting discrimination between respondents at clinically important levels of health (Fitzpatrick et al., 1998). Precision is influenced by several factors including response categories and item coverage of the defined concept of health purportedly measured by the instrument. Limited response categories lack precision and detail, whereas increased gradations of response increase measurement precision (Streiner and Norman, 1995; Fitzpatrick et al., 1998).

Modern psychometric methods, including Rasch analysis, are also used to assess item distribution. Where there is an uneven distribution of items across the proposed hierarchy of health, for example, item grouping in the middle range of functional ability, score change may be influenced by baseline scores and should be considered when interpreting changes in health.

Item content and response format will inevitably influence data quality and scaling, in which floor and ceiling effects are key features. Where more than 20% of responders score at the maximum level of good or bad health, score distribution generally suggests ceiling or floor effects, respectively (Streiner and Norman, 1995; Fitzpatrick et al., 1998). The greater concern is for respondents with already poor health who score at the floor of the instrument range and are consequently unable to report further deterioration in health. Evidence suggests that floor effects are more common with instrument completion by older, sick, or disadvantaged respondents (McHorney, 1996).

Instrument **acceptability** addresses the willingness or ability of patients' to complete an instrument (Fitzpatrick et al., 1998). Although difficult to evaluate directly, this is most readily assessed through instrument completion, response rates, and missing values. Where items within an instrument are consistently omitted, or difficulty is encountered in providing an answer, perhaps due to perceived irrelevance, this would suggest poor acceptability (McHorney, 1996). The font style and size used in questionnaires may also influence completion. Ideally, patients' should be interviewed for their views on instrument completion, content relevance and format during the pre-testing stage of instrument development (Fitzpatrick et al., 1998).

Reading ability is a further consideration regarding instrument acceptability (Streiner and Norman, 1995). A reading level equivalent to that of a 12 year-old has been recommended for questionnaires applicable to the general population (Streiner and Norman, 1995). However, many instruments, including the widely used Nottingham Health Profile (NHP) and the SF-36 have higher reading level requirements (McHorney, 1996; Sharples et al., 2000). It must also be remembered that reading ability may decrease with age (McHorney, 1996). Lack of familiarity with a questionnaire may further reduce response rates in older people (McHorney, 1996).

Instrument completion will also be influenced by mode of administration. Although cheaper than interview or telephone administration, postal administration often results in higher levels of missing values (McHorney, 1996; McColl et al., 2001). Evidence suggests that respondents are more willing to report less favourable health states when completing an instrument themselves than when the instrument is administered by interview (Fitzpatrick et al., 1998; Smeeth et al., 2001). Furthermore, response rates may be influenced by specific item content, for example, items relating to physical or emotional issues; the associated item relevance and appropriateness to the specific population (Bowling, 1998); and response formats, for example, visual analogue scales or Likert scaling (Fitzpatrick et al., 1998). The burden imposed by instrument length and time needed for completion is an important consideration for both respondent and clinician or researcher.

The **feasibility** of instrument administration refers to the time and cost of administration, scoring, and interpretation for clinicians, researchers, and other staff (Fitzpatrick et al., 1998).

REFERENCES

Atkinson T. Atkinson Review: Final Report – Measurement of Government Output and Productivity for the National Accounts. TSO, London 2005.

Beaton DE, Bombardier C, Katz JN, Wright JG. (2001) A taxonomy for responsiveness. *Journal of Clinical Epidemiology.* 54: 1204-1217.

Bowling A. (1995) Measuring Disease. Open University Press, Buckingham.

Bowling A. (1997) Measuring Health. Open University Press, Buckingham.

Department of Health. Our Health, Our Care, Our Say: A New Direction for Community Services. White Paper, London 2006.

Deyo RA, Diehr P, Patrick DL. (1991) Reproducibility and responsiveness of health status measures. Statistics and strategies for evaluation. *Controlled Clinical Trials;* 12: 142S-158S.

Fitzpatrick R, Davey C, Buxton MJ, Jones DR. (1998) Evaluating patient-based outcome measures for use in clinical trials. *Health Technology Assessment;* 2(14).

Garratt AM, Hutchinson A, Russell I. (2001) The UK version of the Seattle Angina Questionnaire (SAQ-UK): reliability, validity and responsiveness. *Journal of Clinical Epidemiology;* 54: 907-915.

Garratt AM, Schmidt L, Mackintosh A, Fitzpatrick R. (2002) Quality of life measurement: bibliographic study of patient assessed health outcome measures. *British Medical Journal;* 324 (7351): 1417-1421.

Jaeschke R, Singer J, Guyatt GH. (1989) Measurement of health status. Ascertaining the minimal clinically important difference. *Controlled Clinical Trials;* 10: 407-415.

Juniper EF, Price DB, Stampone PA, Creemers JP, Mol SJ, Fireman P. (2002) Clinically important improvements in asthma-specific quality of life, but no difference in conventional clinical indexes in patients changed from conventional beclomethasone dipropionate to approximately half the dose of extrafine beclomethasone dipropionate. *Chest;* 121(6): 1824-32.

Kirshner B, Guyatt G. (1985) A methodological framework for assessing health indices. *Journal of Chronic Diseases;* 38: 27-36.

Liang MH. (1995) Evaluating measurement responsiveness. *The Journal of Rheumatology;* 22(6): 1191-1192.

Liang MH, Lew RA, Stucki G, Fortin PR, Daltroy L. (2002) Measuring clinically important changes with patient-oriented questionnaires. *Medical Care.* 40(4): Supplement: II45-II51.

McColl E, Jacoby A, Thomas L, Soutter J, Bamford C, Steen N, Thomas R, Harvey E, Garratt A, Bond J. Design and use of questionnaires: a review of best practice applicable to surveys of health service staff and patients. *Health Technol Assess.* 2001;**5**(31):1-256

McDowell I, Jenkinson C. (1996) Development standards for health measures. *Journal of Health Service Research Policy;* October. 1(4): 238-246.

McDowell I, Newell C. (1996) Measuring Health: a guide to rating scales and questionnaires. Oxford University Press, New York.

McHorney CA, Ware JE, Raczek AE. (1993) The MOS-36-item Short-Form Health Survey (SF-36): II. Psychometric and clinic tests of validity in measuring physical and mental health constructs. *Medical Care;* Mar.31(3): 247-63.

McHorney CA. (1996) Measuring and monitoring general health status in elderly persons: practical and methodological issues in using the SF-36 Health Survey. *Gerontologist;* 36: 571-583.

Norman GR, Sridhar FG, Guyatt GH, Walter SD. (2001) Relation of distribution and anchor-based approaches in interpretation of changes in health-related quality of life. *Medical Care;* 39(10): 1039-1047.

Nunnally JC, Bernstein IH. (1994) Psychometric Theory. McGraw-Hill Series in Psychology, McGraw-Hill, Inc. Third Edition.

Secretary of State for Health The NHS Improvement Plan 2004. London, HMSO, 2004.

Sharples LD, Todd CJ, Caine N, Tait S. (2000) Measurement properties of the Nottingham Health Profile and Short Form 36 health status measures in a population sample of elderly people living at home: results from ELPHS. *British Journal of Health Psychology;* 5: 217-233.

Smeeth L, Fletcher AE, Stirling S, Nunes M, Breeze E, Ng E, Bulpitt CJ, Jones D. (2001) Randomised comparison of three methods of administering a screening questionnaire to elderly people: findings from the MRC trial of the assessment and management of older people in the community. *British Medical Journal;* 323: 1403-1407.

Streiner DL, Norman GR. (1995) Health Measurement Scales. A practical guide to their development and use. Oxford Medical Publications, Inc. Second Edition.

Terwee CB, Dekker FW, Wiersinga WM, Prummel MF, Bossuyt PMM. (2003) On assessing responsiveness of health-related quality of life instruments: guidelines for instrument evaluation. *Quality of Life Research;* 12: 349-362.

The NHS Plan. HMSO Stationery Office, London 2000.

Ware JE. (1997) SF-36 Health Survey. Manual and Interpretation Guide. The Health Institute, New England Medical Centre. Boston, MA. Nimrod Press. Second Edition.

Wyrwich KW, Wolinsky FD. (2000) Identifying meaningful intra-individual change standards for health-related quality of life measures. *Journal of Evaluation in Clinical Practice;* 6(1): 39-49.