Patient-reported Health Instruments Group

Health Status and Quality of Life in Older People: a review

Kirstie L. Haywood
Andrew M. Garratt
Louise J. Schmidt
Anne E. Mackintosh
Ray Fitzpatrick

Report to the Department of Health
April 2004

health outcome indicators

PATIENT-REPORTED HEALTH INSTRUMENTS GROUP
(formerly the Patient-assessed Health Outcomes Programme)

HEALTH STATUS AND QUALITY OF LIFE IN OLDER PEOPLE

A STRUCTURED REVIEW OF PATIENT-REPORTED HEALTH INSTRUMENTS

Kirstie L. Haywood
Andrew M. Garratt
Louise J. Schmidt
Anne E. Mackintosh
Ray Fitzpatrick

Patient-reported Health Instruments Group
National Centre for Health Outcomes Development (Oxford site)
Unit of Health-Care Epidemiology
Department of Public Health
University of Oxford

April 2004

This report should be referenced as follows:

Haywood KL, Garratt AM, Schmidt LJ, Mackintosh AE, Fitzpatrick R. *Health Status and Quality of Life in Older People: a Structured Review of Patient-reported Health Instruments* Report from the Patient-reported Health Instruments Group *(formerly the Patient-assessed Health Outcomes Programme)* to the Department of Health, April 2004.

Copies of this report can be obtained from:

Dr Kirstie Haywood
Co-Director, Patient-reported Health Instruments Group
National Centre for Health Outcomes Development (Oxford site)
Unit of Health-Care Epidemiology
Department of Public Health
University of Oxford
Old Road
Headington
Oxford OX3 7LF

tel: +44 (0)1865 227157
e-mail: kirstie.haywood@uhce.ox.ac.uk

Alternatively, it can be downloaded free of charge from the PHIG website:

http://phi.uhce.ox.ac.uk/

LIST OF CONTENTS

ABBREVIATIONS

Table I Generic instruments reviewed

Instruments	*Domains*	
Assessment of Quality of Life Instrument: AQoL	Illness	Ill
	Independent Living	IL
	Physical Ability	PA
	Psychological Well-Being	PWB
	Social Relations	SR
COOP Charts: COOP **WONCACOOP***	Bodily Pain*	BP
	Daily Activities*	DA
	Emotional Condition*	EC
	Physical Fitness*	PF
	Quality of Life	QL
	Social Activities*	SA
	Social Support	SS
	Overall Health*	OH
	Change in health status*	
European Quality of Life instrument (EuroQol): EQ-5D	Anxiety/Depression	AD
	Mobility	M
	Pain/Discomfort	PD
	Self-Care	SC
	Usual Activities	UA
	EuroQol thermometer	EQ thermometer
Functional Status Questionnaire: FSQ	Activities of Daily Living	ADL
	Instrumental ADL	IADL
	Psychological Function	PsychF
	Work performance	WP
	Social Function	SF
	Quality of Social Interaction	QSI
Goteborg Quality of Life Instrument: GQL	Social Well-Being	SWB
	Physical Well-Being	PWB
	Mental Well-Being	MWB
Health Status Questionnaire-12: HSQ-12	Bodily Pain	BP
	Energy/fatigue	E
	Mental Health	MH
	Physical Functioning	PF
	Perceived Health	PH
	Role Limitation-Mental	RM
	Role Limitation-Physical	RP
	Social Functioning	SF
Index of Health Related Quality of Life: IHQL	Disability	
	Discomfort	
	Distress	
Nottingham Health Profile: NHP	Bodily Pain	BP
	Emotional Reactions	ER
	Energy	E
	Physical Mobility	PM
	Sleep	S
	Social Isolation	SI
Quality of Life Index: QLI	Family	FAM
	Health and Functioning	HF
	Psychological/Spiritual	PP
	Social/Economic	SE
Quality of Well-being Scale: QWB	Mobility and Confinement	MOB
	Physical Activity	PAC
	Social Activity	SAC

Medical Outcomes Study (MOS) Short Form 12-item Health Survey: SF-12	Bodily Pain	BP
	Energy/Vitality	V
	General Health	GH
	Mental Health	MH
	Physical Functioning	PF
	Role Limitation-Emotional	RE
	Role Limitation-Physical	RP
	Social Functioning	SF
SF-12 summary scores	Mental Component Summary	MCS
	Physical Component Summary	PCS
Medical Outcomes Study (MOS) Short Form 20-item Health Survey: SF-20	Bodily Pain	BP
	General Health	GH
	Physical Function	PF
	Mental Health	MH
	Social Function	SF
	Role Function	RF
Medical Outcomes Study (MOS) Short Form 36-item Health Survey: SF-36	Bodily Pain	BP
	General Health	GH
	Mental Health	MH
	Physical Functioning	PF
	Role Limitation-Emotional	RE
	Role Limitation-Physical	RP
	Social Functioning	SF
	Vitality	V
SF-36 summary scores	Mental Component Summary	MCS
	Physical Component Summary	PCS
Sickness Impact Profile: SIP	Alertness Behaviour	AB
	Ambulation	A
	Body Care and Movement	BCM
	Communication	C
	Eating	E
	Emotional Behaviour	EB
	Home Management	HM
	Mobility	M
	Recreation and Pastimes	RP
	Sleep and Rest	SR
	Social Interaction	SI
	Work	W
SIP summary scores	Physical(A,BCM,M)	SIP-PhysF
	Psychosocial (AB,C,EB,SI)	SIP-PsychF
Spitzer Quality of Life Questionnaire: SQL	Activity level	AL
	Activities of Daily Living	ADL
	feelings of Healthiness	H
	quality of Social Support	SS
	psychological Outlook	O

Table II Older people-specific instruments reviewed

Instruments	***Domains***	
Brief Screening Questionnaire: BSQ	Activities of daily living	ADL
	Cognitive impairment	CI
	Financial impact	FI
	Functional mobility	FM
	Hearing impairment	HI
	Mental health	MH
	Polypharmacy	PP
	Social contact	SC
	Symptoms	S
	Visual impairment	VI
Comprehensive Assessment and Referral Evaluation: CARE	Psychiatric	
	Medical/physical/nutritional	
	Social needs	
	Service needs	
CORE-CARE	Depression	
	Dementia	
	Disability	
	Subjective memory	
	Sleep	
	Somatic symptoms.	
CORE-CARE Summary scores	Psychiatric	
	Physical	
	Service Needs	
	Social	
SHORT-CARE	*as CORE-CARE*	
SHORT-CARE Diagnostic scales	Depression	
	Dementia	
	Disability	
Elderly Assessment System: EASY-Care	General health	
	Disability: activities of daily living, instrumental ADL	ADL, IADL
	Memory	
	Home/Safety/Support	
	Health-care services	
	Looking after your health	
Functional Assessment Inventory: FAI	ADL impairment	ADL
	Economic resources	ER
	Mental health	MH
	Physical health	PH
	Social resources	SR
Geriatric Postal Screening Survey: GPSS		
Specific conditions	Falls/balance	
	Functional impairment	
	Depression	
	Cognitive impairment	CI
	Urinary incontinence	
General health status	Health Perception	HP
	Polypharmacy	PP
	Bodily Pain	BP
	Weight Loss	WL
Geriatric Quality of Life Questionnaire: GQLQ	Activities of Daily Living	ADL
	Symptoms	S
	Emotional Function	EF

Geriatric Screening Questionnaire: GSQ	Cognitive Impairment	CI
	Daily Activities	ADL
	Economic Status	ES
	General Health status	GH
	Mental Health	MH
	Social Support	SS
IOWA Self-Assessment Inventory (Revised): ISAI	Anxiety/Depression	AD
	Alienation	A
	Cognitive Status	CS
	Economic Resources	ER
	Mobility	M
	Physical Health	PH
	Social Support	SS
LEIPAD	Cognitive Function	CF
	Depression/Anxiety	DA
	Life Satisfaction	LS
	Physical Function	PF
	Self-Care	SC
	Social Functioning	SocF
	Sexual Functioning	SexF
Older Americans Resource Study (OARS) Multi-dimensional Functional Assessment Questionnaire: OMFAQ	Activities of daily life (with Instrumental ADL)	ADL (IADL)
	Economic Resources	ER
	Mental Health	MH
	Physical Health	PH
	Social Resources	SR
Perceived Well-being Scale: PWBS	Psychological Well-Being	Psych-WB
	Physical Well-Being	Phys-WB
	General Well-Being	GWB
Philadelphia Geriatric Centre Multilevel Assessment Instrument: PGCMAI	Activities of Daily Living	ADL
	Cognition	C
	Perceived Environment	PE
	Personal Adjustment	PA
	Physical Health	PH
	Social Interaction	SI
	Time Use	TU
Quality of Life Cards: QLC	Affect	
	Life experience	
	Satisfaction/happiness	
Quality of Life Profile-Senior Version: QOLPSV	Being: physical, psychological, spiritual	
	Belonging: physical, social, community	
	Becoming: practical, leisure, growth	
Quality of Life - well-being, meaning and value: QLWMV	Well-being	
	Meaning	
	Value: self-worth	
	Health	
	Functional capacity	
	External factors	
Self-Evaluation of Life Scale: SELF	Depression	D
	Personal Control	PC
	Physical Disability	PD
	Self-Esteem	SE
	Social Satisfaction	SS
	Symptoms of Ageing	SA

SENOTS programme and battery	Activity Limitation	AL
	Activity Propensity	AP
	Financial Hardship	FH
	Happiness/depression	H
	Physical Symptoms	PS
The Wellness Index: WI	Activities of Daily Living, Instrumental ADL	ADL,IADL
	Economic Resources	ER
	Morale	M
	Physical Health	PH
	Religiosity	R
	Social Resources	SR

Executive Summary

Background

This report presents a review of generic and older people-specific self-reported instruments measuring aspects of health and quality of life (HRQL) that have been evaluated for use with older people. This review will provide potential users with information guiding the selection and application of these instruments in clinical trials, routine practice, and population surveys.

Research aims:

1. to identify generic self-reported, multidimensional instruments that measure HRQL and have been applied in the assessment of older people;

2. to identify older people-specific self-reported, multidimensional instruments that measure HRQL;

3. to extract and assess evidence relating to the development and evaluation of these instruments, and make recommendations as to their application.

Methods

Electronic databases from their inception to September 2003 were searched using keywords relevant to the development and testing of self-reported instruments that measure HRQL in older people. Several other sources, including journals, were also searched. The names of instruments were then used in a second search strategy. Studies describing instrument development and evaluation were retrieved. Instrument reviews were also retrieved.

After retrieving published papers, the following information was extracted relating to instrument development and evaluation:

- instrument purpose, including the underlying conceptual base being measured and proposed application;
- instrument development, content, and scoring;
- older populations and study settings in which the instrument was developed and tested;
- measurement properties of reliability, validity, responsiveness, and precision;
- instrument acceptability, including response rates and missing data.

Key findings

Generic instruments

15 generic instruments met the review inclusion criteria. The SF-36, Sickness Impact Profile (SIP), and EuroQol EQ-5D have undergone more evaluations following the assessment of older people than the others. Most instruments were developed and evaluated in North America. The COOP and WONCA/COOP charts, EuroQol, Health Status Questionnaire-12 (HSQ-12), Index of Health-related Quality of Life (IHQL), Nottingham Health Profile (NHP), SF-12 and SF-36 have published UK evaluations.

Evaluations for several instruments reflect a range of settings, including the community, primary care, hospital, day-care, and residential institutions.

All generic instruments assess physical function; most assess psychological and social well-being. Three instruments assess cognitive well-being, namely the Goteberg Quality of Life questionnaire (GQL), SIP, and the Spitzer Quality of Life index (modified) (SQL). The COOP, SF-36, and SIP assess the widest range of health domains.

The SF-36 has the most extensive evidence of reliability. Four generic instruments, namely the NHP, SF-12, SF-20, SF-36, have evidence of internal consistency and test-retest reliability. The range of reliability estimates support application at the group level and, in some instances, at the individual level. There is limited evidence supporting the application of the COOP and EuroQol EQ-5D at the group level. Four instruments do not have evidence of reliability, namely HSQ-12, IHQL, Quality of Well-being Scale (QWB), and SQL.

Patients and members of the general population were involved in item generation for the NHP, SIP and a modified version of the SQL. However, it is not clear that older people were involved in this process.

Empirical evidence supports the internal construct validity of three instruments, namely the AQoL, SF-12, and SF-36.

With the exception of the Quality of Life index (QLI) and SF-12, all generic instruments have evidence for validity through comparison with instruments that measure similar or related constructs; this is most extensive for the SF-36. With the exception of the COOP, GQL, IHQL, and SQL, all generic instruments have evidence to support their ability to discriminate between groups defined by a range of socio-demographic, health, and health-service use variables; this is most extensive for the EQ-5D, HSQ-12, NHP, SF-12, SF-36, and SIP. The AQoL, COOP, SF-20, SF-36, and SIP have evidence of predictive validity.

With the exception of the GQL, HSQ-12, IHQL, QLI and QWB, all generic instruments have some evidence of responsiveness; this is most extensive for the SF-36 across a range of settings. Strong levels of responsiveness were found for the EQ-5D and NHP where change in health was substantive, for example, following the surgical repair of hip fracture.

Ceiling and floor effects were reported for several domains within the COOP, SF-20, SF-36 (role limitation domains), and SIP. Ceiling effects were reported for domains within the AQoL, FSQ, HSQ-12, and NHP (all domains).

Completion rates were higher with interview administration than with self-completion and ranged from 75% (IHQL) to 100% (COOP charts and NHP). Self-completion rates ranged from 43% (SIP) to 95% (NHP). Completion rates were not reported for the AQoL, GQL, or SQL. Mean completion times for interview administration ranged from ten minutes (NHP) to 35 minutes (SIP). Self-completion times were frequently not reported.

Older people-specific instruments

18 older people specific instruments met the review inclusion criteria. The OARS Multidimensional Functional Assessment Questionnaire (OMFAQ), the Comprehensive Assessment and Referral Evaluation (CARE), the Functional Assessment Inventory (FAI), and the Quality of Life Profile - Senior Version (QOLPSV) have undergone the highest number of evaluations. However, most evaluations for the CARE and the QOLPSV refer to the same older population. The majority of instruments were developed and evaluated in North America; most have one published evaluation. The CARE, EASY-Care, and Brief Screening Questionnaire (BSQ) have published UK evaluations; the CARE was developed in the USA and UK, and the EASY-Care in the UK and other European countries. Most instruments were evaluated in community settings.

Most instruments assess physical function, psychological well-being, and social well-being; seven instruments also assess cognitive function, namely the BSQ, EASY-Care, Geriatric Postal Screening Survey (GPSS), Iowa Self-Assessment Instrument (ISAI), LEIPAD, OMFAQ, and the Philadelphia Geriatric Centre Multilevel Assessment Inventory (PGCMAI). The EASY-Care assesses the widest range of health domains.

There is limited evidence of reliability for most instruments. Four instruments, namely LEIPAD, PGCMAI, Perceived Well-being Scale (PWB), and the Wellness Index (WI), have evidence of internal consistency and test-retest reliability supporting their application in the assessment of groups and, for the PGCMAI and WI, in the assessment of individuals. The BSQ and Geriatric Quality of Life Questionnaire (GQLQ) do not have evidence of reliability.

Older people were involved in item generation for three instruments, namely the GQLQ, QOLPSV, and WI.

Empirical evidence supports the proposed health domains assessed by six instruments, namely the ISAI, LEIPAD, OMFAQ, PWB, Self-evaluation of Life Scale (SELF), and WI.

With the exception of the BSQ, EASY-Care, GSQ, and SELF, all instruments have evidence for validity through comparison with instruments that measure similar or related constructs. With the exception of the BSQ, GQLQ, GSQ, LEIPAD, and Quality of Life Cards (QLC), all instruments have evidence to support their ability to discriminate between groups defined by a range of socio-demographic, health and health-service use variables; this is most extensive for the FAI, GPSS, ISAI, PWB, PGCMAI, QOLPSV, the SENOTS battery, and WI. The CARE, GPSS, OMFAQ, and SELF have evidence of predictive validity.

Evidence of responsiveness was found for only five instruments, namely the GQLQ, OMFAQ, PGCMAI, QOLPSV, and SELF, and this was limited.

Ceiling effects were reported for the OMFAQ ADL and IADL domains. Response distributions were not reported for the remaining instruments.

Although infrequently reported, completion rates were generally higher with interview administration than with postal self-completion; the QOLPSV had the lowest reported self-completion rate. Evidence of acceptability is lacking for the PWB, QLC, and WI.

General

There are few concurrent instrument evaluations, particularly in relation to responsiveness. Most evaluations include the SF-36. Similar levels of reliability and evidence for validity are reported for the SF-36 and EuroQol EQ-5D, and for the SF-36 and NHP. The SF-36 appears to be more responsive across lower levels of morbidity; the EQ-5D and NHP may be more responsive where substantive changes in health are expected.

Seven concurrent evaluations comparing generic and older people-specific instruments were reviewed; reliability and content validity were frequently not evaluated. Higher or comparable levels of responsiveness were reported for two older people-specific instruments, the OMFAQ and Geriatric Quality of Life Questionnaire. However, higher levels of responsiveness were reported for the SF-36 when compared to the OMFAQ and QOLPSV.

For the most extensively studied instruments, evidence suggests that completion difficulties increase with age, declining cognitive ability, and deteriorating health status. Interview administration generally yielded increased completion rates and associated increased completion times when compared to self-completion.

The point at which an individual with cognitive impairment is unable validly to report on their health is not known. The majority of studies excluded cognitively impaired respondents.

Evidence from proxy completion of the EQ-5D, NHP, and SF-36, suggests that informed health professionals are better able to interpret an individual's health status than nominated lay proxies. There is greater agreement between proxies and patients for the assessment of more observable health states; proxies may overestimate health limitations, particularly for less observable health constructs such as emotions and mental status. High participation rates were found for proxy and respondent completion of the OMFAQ and FAI.

Three generic instruments were evaluated for screening purposes, namely the CARE, COOP, and SIP (mobility). The SHORT-CARE and SIP had good sensitivity for levels of depression and poor function, respectively, but had poor specificity. Three older people-specific screening instruments, the BSQ, GPSS, and GSQ, were reviewed; all three instruments require further evidence of measurement properties.

Key conclusions and recommendations

There has been a huge growth in the availability of patient-reported health instruments over the last decade. There are many from which to choose for the assessment of older people.

Two broad approaches to measuring health from the perspective of the older person have been reviewed: generic instruments and older people-specific instruments. Generic instruments aim to cover aspects of health status and quality of life of relevance to the general population. Older people-specific instruments aim to cover issues of specific relevance to the older population.

Generic instruments are suitable for comparisons across general and specific populations; the availability of normative data supports the interpretation of data. Generic instruments are also particularly relevant to economic evaluation. The broad nature of generic instruments facilitates the identification of co-morbid features and treatment side-effects that may not be captured by specific instruments; however, this may also reduce responsiveness.

Their use in general population surveys and the results of this review support the application of several generic instruments in community-dwelling older people, particularly in those with lower levels of morbidity. For example, evidence suggests that the SF-36 is more responsive than older people-specific instruments, namely the OMFAQ and QOLPSV, in community-dwelling adults with acute or chronic illness. However, item relevance may reduce acceptability and responsiveness in the very old, and those with physical disabilities.

Older people-specific instruments have greater clinical appeal due to their specificity of content. Greater respondent acceptability may be associated with the relevance of items to immediate health concerns. Instruments may have an associated increased responsiveness to specific changes in health. However, few older people-specific instruments included older people in item derivation, and evidence of responsiveness is limited.

Generic instruments have undergone more evaluations in the older population than older people-specific instruments, and have more evidence for measurement and practical properties. There is insufficient evidence from concurrent evaluations to indicate whether older people-specific instruments perform better than generic instruments.

The most extensive evidence for measurement properties, offering some support for its application in the assessment of individuals, and responsiveness to change in health across several settings was found for the generic SF-36. There is also good evidence for the reliability of the EQ-5D and NHP, supporting their application in the assessment of groups, and for their validity and responsiveness . Evidence is more limited for the COOP charts, SF-12, and SIP. With the exception of the SIP, all instruments have been evaluated in UK populations. The SF-12 and SF-36 version 2 have yet to be evaluated in an older population. The IHQL and QWB lack evidence for reliability and responsiveness and cannot be recommended for assessing older people.

Where a more detailed and broad ranging assessment of HRQL is required, particularly in older people with lower levels of morbidity, the SF-36 is recommended. Where a more succinct assessment of HRQL is required, particularly for patients in whom a substantive change in health is expected, the EuroQol EQ-5D is recommended; however, further evidence of its reliability and acceptability to respondents is required.

The greatest evidence for measurement properties of older people-specific instruments, with support for application of the ADL domain in assessing individuals, was found for the OMFAQ. However, most evidence is for the ADL/IADL domain only; evidence for reliability and responsiveness is limited. There is limited evidence of reliability, validity, and responsiveness for the PGCMAI, QOLPSV, and SELF. None of these instruments has been evaluated in a UK population. The CARE and EASY-Care are the most widely evaluated in UK populations. The EASY-Care has limited evidence of validity and both CARE and EASY-Care lack evidence of responsiveness.

Several older people-specific instruments, namely the BSQ, EASY-Care, GPSS, and GSQ, are relatively new and further evidence of their performance is required. The EASY-Care is an important development in the comprehensive assessment of older people and in the single assessment process. The BSQ, GPSS, and GSQ are new, self-completed instruments for the postal screening of community-dwelling older people, which aims to identify those who would most benefit from a comprehensive assessment.

When selecting an instrument, the appropriateness of item content, relationship to the proposed application and population group, and level of respondent and clinician/researcher burden in terms of time, cost, and feasibility of application should be considered. The EASY-Care covers the broadest range of domains when compared to both generic and older people-specific instruments, and has an economical number of items (total: 85). Undue length may limit the scope for application of several instruments, for example, the generic SIP and older people-specific CARE. The shortest instruments are the generic EQ-5D and older people-specific GSQ. Several instruments cover similar domains with a limited number of items (less than 38): the generic AQoL, COOP, EuroQol, HSQ-12, NHP, SF-12, and SF-36, and the older people-specific BSQ, GPSS, GSQ, and LEIPAD.

Interview administration generally increases instrument completion rates, but at increased cost. Practical considerations, for example, larger typeface and greater use of white space in the questionnaire format, and persuasive methods, for example, telephone contact and home visits, may be required to increase response rates following postal self-completion.

The application of patient-reported health instruments across the spectrum of cognitive impairment in older people is required to evaluate instrument performance further.

Responsiveness has been the most neglected area of instrument evaluation with older people. Moreover, the level of change in HRQL that is important to the respondent, the Minimal Important Difference (MID), has not been addressed. Instruments should be administered longitudinally before and after changes in treatment known to improve HRQL, and health transition questions should be included as external criteria of change in health. Where possible, the relative responsiveness of instruments should be assessed concurrently.

Further evaluation and, where appropriate, refinement of existing instruments is required before new instruments are developed; seeking the views of older people with regard to instrument format, relevance, and mode of completion is strongly recommended. Where it is deemed necessary to develop new instruments, the close involvement of older people in instrument development is recommended.

Supported by recommendations from this review, comparative empirical evaluations of widely used generic and new or widely used older people specific instruments, global assessments and domain-specific instruments are required across the wide range of settings in which older people may be invited to report on their health status. This research will inform decisions regarding the selection of instruments for future application in research and clinical practice.

Chapter 1: INTRODUCTION

a) Older people

In the United Kingdom (UK), older people represent a growing proportion of the population: the National Service Framework for Older People (NSF-OP, 2001) suggests that 20% of the population of England is over 60 years of age. Compared with an anticipated increase of 8% in the size of the total population of England and Wales by 2031 (NHS R&D Strategic Review, 1999: p.2), a disproportionate increase is predicted in the number of old (43% in those aged 60-74 years, 48% in those aged 75-84 years) and very old (138% increase in those aged 85 years and older). Reflecting the increasing need for health and social service provision with age, older people, when defined as those aged over 60 years, are the main users of health and social services in the UK. Effective and appropriate assessment of the need for and outcomes of health- and social care in older people is therefore a significant issue. Patient-reported measures of health status are an important aspect of this process (Heyrman and van Hoeck, 1996; Albert, 1997; Garratt et al., 2002a).

Older people, thus defined, represent a diverse population differing not only in age and health status but also, for example, in cultural background and ethnicity (NSF-OP, 2001). The NSF-OP (2001) defines three broad categories of age: entering old age (generally 60 years and above), a transitional phase (the seventh or eighth decade), and frail older people (late old age). These groupings define a continuum of general health status, distinguished by levels of activity and independence, and corresponding need or demand for health service provision. However, it is recognised that certain individuals will remain active and independent well into old age, whereas others may experience significant illness earlier in their old age, which will inevitably affect their demand for health- and social care. Following the definition provided by several authors (for example, Arnold, 1991; Albert, 1997) and guidance from the NSF-OP (2001), this review refers to individuals entering old age as young-old, and those aged over 85 years as old-old. Where authors specifically describe the population as ‘frail elderly’ this is indicated.

b) Patient-reported health instruments

Patient-reported health instruments aim to include in the assessment process the patient’s perspective across a range of health-related concerns, from symptoms and physical functioning to well-being and quality of life (Fitzpatrick et al., 1998). These instruments are usually self-completed and provide a measure of an individual’s experiences and concerns in relation to their health status. Where necessary, others, for example, nominated relatives or clinicians, may complete the instrument on behalf of the individual. This is often referred to as proxy completion, and can be an important source of health information particularly in the case of chronically debilitated or cognitively impaired individuals (Neumann et al., 2000).

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 (Fitzpatrick et al., 1998; Garratt et al., 2002b). The term ‘patient-reported health instrument’ will be used throughout this review to

refer to patient-completed instruments variously defined as measures of functional status, well-being, health status, or health-related quality of life (HRQL).

As a result of the increasing focus on patient-reported health, several hundred instruments are now available, and for many health states and diseases there is often a choice of instrument (Garratt et al., 2002a). Several factors have led to the increased use of patient-reported health instruments. With the advance of medical technology and an ageing population, incurable chronic disease and long-term illness currently dominate the health-care environment of the developed world, entailing a change of emphasis in assessment towards quality of survival and HRQL (McDowell and Newell, 1996).

When mortality is no longer the main concern of outcome assessment, the holistic view of health defined by the World Health Organisation (WHO) as 'a state of complete mental, physical, and social well-being, not merely the absence of disease and infirmity' has increased relevance (WHO, 1947). This statement views health as a complex construct, the measurement of which should include issues of relevance to patients, health-care professionals and providers (McDowell and Newell, 1996; Ware, 1997). Increasingly recognised as the best judge of disease impact, the patient's perception of their health status is recommended as a core component in clinical assessment (Albrecht, 1994; Fitzpatrick et al., 1998), and treatment that improves only traditional biomedical features without benefiting HRQL may be considered to have only limited success.

The role of the older individual in health-care evaluation has recently been highlighted in the single assessment process for the assessment of older people's health- and social care status (Single Assessment Process, Department of Health [DH], September 2002). This approach to evaluation accords with increasing expectations by patients of their role as active partners in medical care, exemplified in chronic disease management through the Expert Patient Agenda (The Expert Agenda, DH, 2001). At a policy level, prioritisation within health-care is the inevitable consequence of limited resources, and the use of appropriate measures of health outcome can enhance the efficiency of resource allocation (Guyatt et al., 1993a; Ware, 1997).

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 of HRQL relevant to both patients and the general population, supporting their application with both populations (Guyatt et al., 1993a; Ware, 1997). Population-based normal values can be calculated which support the interpretation of data from general population and disease-specific groups (Ware, 1997).

There are two classes of generic instrument: health profiles and utility measures. Scores on different domains of HRQL covered by a health profile are presented separately to support data interpretation. Sometimes a single or summary score may be generated, but proponents for profiles argue that measurement is most meaningful within separate domains. The Medical Outcomes Summary 36-item Short Form Health Survey questionnaire (SF-36) is a widely used example of a generic health profile (Ware, 1997). The items cover eight domains of HRQL, including physical and social functioning and mental health. Population norms have been calculated in several countries (McDowell and Newell, 1996; Ware, 1997).

The values and preferences for outcome generated by the patient (direct weighting) or the general population (indirect weighting) provide weightings for utility measurement (Garratt et al., 2001). Although utility measures usually cover several domains of HRQL, the weighting generates a single index that relates HRQL to death (0) or perfect health (1) (Guyatt et al., 1993a). The EuroQol (EQ-5D) is an example of a utility measure that incorporates indirect valuations of health states (EuroQoL Group, 1990). An advantage of utility measures is their recommended use in cost-utility economic analysis; a disadvantage is that the single score limits data interpretation (Kind et al., 1998; Garratt et al., 2001).

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) (Guyatt et al., 1993a). Disease-specific instruments may have greater clinical appeal due to their specificity of content, and associated increased responsiveness to specific changes in condition (Guyatt et al., 1993a). Increased item relevance may also enhance their acceptability to respondents.

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 (Guyatt et al., 1993a; McDowell and Newell, 1996).

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 (Jenkinson and McGee, 1998). 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 (Garratt et al., 2002a), and instruments with measurement properties that support their application in specific populations (for example, older people) and across a range of evaluation settings need to be identified.

c) Criteria for instrument selection

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** (Patrick and Erickson, 1993; McHorney, 1996; Fitzpatrick et al., 1998).

Reliability is concerned with whether measurement is accurate over time and, for multi-item instruments, whether they are internally consistent (Garratt et al., 2002b). 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 (Fitzpatrick et al., 1998). 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 (for example, an older population), 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 (Garratt et al., 2002b). 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 (Garratt et al., 2002b). 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 (Kosinski et al., 1999). 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 (Husted et al., 2000; Liang, 2000; Wyrwich et al., 2000; Wells et al., 2001; 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, also referred to as measures of internal responsiveness (Husted et al., 2000), 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 (Liang, 1995). 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 (Bland and Altman, 1986; Streiner and Norman, 1995; Beaton et al., 2001). Calculating the 95% Limits of Agreement as an estimate of test-retest reliability gives a range of values that is expected to describe the agreement between two observations for most patients indicating no change in health (Bland and Altman, 1986; Altman, 1996). Few repeat observations will be identical due to random error, and score changes above this range support the interpretation of real change, or responsiveness.

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; et al., 1998; Husted et al. 2000). Instruments demonstrating strong cross-sectional validity should also be valid for measuring within-person change over time (Katz et al., 1992; Ware, 1997). However, it is argued that both these measurement properties should be assessed for evaluative instruments (Kirshner and Guyatt, 1985; Deyo et al., 1991; Terwee et al., 2003).

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 (Garratt et al., 2003).

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). However, older people are frequently unrepresented in this process (Walters et al., 2001).

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). Several instruments, for example, the Older Americans Resources and Services (OARS) Multidimensional Functional Assessment Questionnaire (OMFAQ) (George and Fillenbaum, 1985), require detailed training for interviewers, adding to the cost of application. Other instruments, for example, the COOP charts, can be self-completed or interview-administered and require minimal additional time and effort (Nelson et al., 1990; Fitzpatrick et al., 1998).

d) Assessment of older people

Although some individuals may experience a relatively healthy old age with few health impairments, evidence suggests that individuals aged over 75 years suffer from an average of seven significant disease states (Heyrman and van Hoeck, 1996). Most of

these are chronic and incurable conditions, for example, arthritis, congestive heart failure, and chronic lung disease (Bowling, 1995; Heyrman and van Hoeck, 1996). Therefore, in addition to the need for many condition-specific instruments to be acceptable to a wide range of age-groups, including older people, specific impairments may hinder the ability of an individual to complete questionnaires. For example, 80% of individuals over the age of 60 years are visually impaired and 75% are hearing-impaired, while 22% are expected to suffer from impairment of both vision and hearing (NSF-OP, 2001). These impairments may hinder self- or telephone-completion of questionnaires.

Impaired cognitive functioning is also more common among older people (Kirby et al., 1998). It is estimated that 5% of people aged 65 years and above experience some degree of dementia; this increases to 20% in people aged over 80 years (Hofman et al., 1991 cited by Kirby et al., 1998). In the course of a dementing illness, the stage at which an individual becomes unable to report on their individual health state is unknown (Fletcher et al., 1992; Albert, 1997). These issues highlight some of the practical demands involved in selecting methods of assessment for older people. In addition, due to the wide range of co-morbidity in the older population, instruments such as generic and older people-specific measures of HRQL that support the assessment of broader concepts of health status provide an important source of comparative data across older population groups.

Older people demonstrate great heterogeneity in the constructs that underpin HRQL including emotional well-being, self-esteem, and satisfaction with social support. The need for a multidimensional assessment of health status, which may include disease-specific and generic instruments, as well as those specific to older people, has been described by several authors (for example, Fletcher et al., 1992; Bowling, 1995; McHorney, 1996), and acknowledged within guidance on the Single Assessment Process (SAP) for older people (Single Assessment Process, DoH, September 2002). The SAP was first detailed within the NSF-OP (2001), with the general aim of providing ‘person-centred, effective, and co-ordinated’ assessment and planning for the care of older people (Single Assessment Process, DH, September 2002: p.3). Although specific instruments were not recommended, the SAP defines four types of assessment differentiated by the level of detail required, namely contact, overview, specialist, and comprehensive. Health professionals are responsible for determining the type of assessment required. More detailed assessments aim to describe an individual’s strengths, abilities, and needs, and to discriminate between the perceived quality of life of assessed individuals (NHS R&D Strategic Review, 1999; Single Assessment Process, DH, September 2002).

Comprehensive Geriatric Assessment

Comprehensive geriatric assessment (CGA) has been defined as ‘a systematic method of assessing the physical, mental, and social functioning of older people’ (Philp, 2000: p.15). Such comprehensive assessment should recognise the complexity and diversity of the physical, mental, and social needs of this specific group, and the impact of health- and social care utilisation (Rockwood, 1995). This requires a multidimensional approach to assessment.

Several authors refer to assessments invariably described as providing a ‘comprehensive geriatric assessment’ (Stuck et al., 1993; Rockwood, 1995; Philp, 2000; Repetto et al., 2001; Ingram et al., 2002). Although referring to multidimensional assessments,

specific item content varies between approaches. This lack of standardisation across assessments is familiar in health-care evaluation. Most CGA describe a battery of items, selected often from established instruments or on the basis of expert opinion. For example, the Gero-Oncology Health and Quality of Life Assessment includes items from the OMFAQ, namely instrumental activities of daily living, co-morbidity, and financial well-being; the Medical Outcomes Study social support scale; the Hospital Anxiety and Depression Scale; and a cancer-specific quality of life questionnaire, the European Organisation for Research and Treatment of Cancer Quality of Life Questionnaire EORTC QLQ-C30 (Ingram et al., 2002).

The EASY-Care assessment (reviewed in Chapter 5) is described as a CGA, and provides a multidimensional assessment specific to older people that is not disease-specific (Philp, 2000). Developed across several European countries, the EASY-Care is recommended for use within the single assessment process (SAP), supporting patient-centred and standardised assessment (Single Assessment Process, DH, September 2002). The EASY-Care provides a contact and overview assessment, serving as a foundation for the specialist or comprehensive assessments (The Single Assessment Process: EASY-Care Training Pack, 2003: p.17).

However, a CGA can be time-consuming and costly, and hence is most cost-effective when targeted on older people at risk of frailty or functional decline (Fernandez Buergo et al., 2002; Alessi et al., 2003). Hence, screening programmes to identify older people who would most benefit from CGA and associated health-care management plans have been proposed.

Screening in older people

The UK's NSF-OP (2001) recommends that all older people should receive a single annual assessment. The benefit to be gained, in terms of providing appropriate and timely health- and social care, from screening the wider ageing population to identify those most in need of more detailed or comprehensive geriatric assessment, and the financial cost, has been described (Smeeth et al., 2001; Alessi et al., 2003). Several screening instruments specific to older people have been included in the review: the Brief Screening Questionnaire (UK; Smeeth et al., 2001), the Geriatric Postal Screening Survey (USA; Alessi et al., 2003), and the Geriatric Screening Questionnaire (Spain; Fernandez Buergo et al., 2002). These are reviewed in Chapter 5. Limited evidence suggests that postal administration of screening instruments is an acceptable mode of administration. However, non-responders may have greater levels of impairment, and persuasive methods to increase response rates may be required, for example, telephone contact and home-visits (Alessi et al., 2003).

The sensitivity and specificity of screening instruments is often reported. Sensitivity is the proportion of truly diseased persons in the screened population who are identified as such by the test, i.e. the true positive rate (Last, 1995: p.154). Specificity is the proportion of truly non-diseased persons who are so identified by the test: the true negative rate.

e) Summary

Evidence for the effective performance of instruments in this diverse population and across the wide range of settings in which older people may receive care or be invited to report their health status will promote evidence-based health-care. It cannot be assumed

that generic or disease-specific instruments with evidence of good measurement properties in a younger population will perform as well with an older population (Bowling, 1997, 1998). Furthermore, in response to the reported need for research programmes to reflect the appropriate demographics of the population and not to exclude older people (NHS R&D Strategic Review, 1999: p.12), instruments with evidence of measurement properties and good acceptability across the age-ranges are required. Hence, when selecting a patient-reported instrument for use in research or clinical practice, the appropriateness of item content, relationship to the proposed application and patient population, and evidence of measurement properties in the chosen setting and population should be considered (Fitzpatrick et al., 1998; Higginson and Carr, 2001).

This review provides a structured synthesis of published evidence for the measurement and practical properties of generic and older people-specific instruments that provide a multidimensional assessment of HRQL and have been completed by older people. The review aims to inform the future selection of instruments for application in research and clinical practice.

Chapter 2: METHODS

a) Search strategy

The search strategy was designed to retrieve references relating to patient-reported health instruments and older people, including the development and testing of instruments, instrument reviews, and conceptual and methodological issues in measurement. The search strategy was not designed to retrieve references relating solely to the application of instruments.

Hosted by the National Centre for Health Outcomes Development (NCHOD) at the University of Oxford, the Patient-reported Health Instruments (PHI) website (http://phi.uhce.ox.ac.uk/) includes a bibliography of over 6500 records relating to published instrument evaluations found on the following electronic databases: Allied and Alternative Medicine (AMED), Biological Abstracts, British Nursing Index, Cumulative Index to Nursing and Allied Health Literature (CINAHL), Econlit, EMBASE, Medline, PAIS International, PsycInfo, System for Information on Grey Literature in Europe (SIGLE), and Sociological Abstracts. At the time of this review, the bibliography comprised references up to December 2002. The primary search of the bibliography used the terms 'old' (old*) or 'elderly' (elder*) or 'senior' (senior*) across all types of publication, including research reviews and primary studies. A secondary search of the database used the names of identified instruments.

Further searches of five electronic databases, namely AMED, CINAHL, EMBASE, Medline, and PsycIinfo, extended the search period to September 2003 (January 2001 to September 2003). These searches used the terms 'old' (old*) or 'elderly' (elder*) or 'senior' (senior*), combined with the names of identified outcome measures. All searches were restricted to English language publications.

The reference lists of included articles were reviewed for additional articles (Hayes et al., 2000; Garratt et al., 2002b). The journal Quality of Life Research was hand-searched, and texts and compendia were consulted (McDowell and Newell, 1996; Spilker, 1996; Bowling, 1997). The reference lists of existing reviews of outcome measures used in the assessment of older people (Kane and Kane, 1984; Rubenstein et al., 1989; Fletcher et al., 1992; Bowling, 1995; Heyrman and Van Hoeck, 1996; Albert, 1997; Kliempt et al., 2000; Philp et al., 2001) and manuscripts discussing the assessment of older people (NHS R&D Strategic Review, 1999; NSF-OP, 2001; Single Assessment Process, DH, September 2002) were also reviewed.

Authors of instruments identified as specific to older people were contacted for further information about instrument development and testing.

b) Inclusion criteria

Titles and abstracts of all articles were assessed for inclusion/exclusion by two independent reviewers and agreement was checked. Articles included were retrieved in full. Published articles were included if they provided evidence relating to criteria considered important in instrument evaluation (Fitzpatrick et al., 1998), specifically for generic or older people-specific instruments assessing multidimensional aspects of health status and quality of life in older people (those aged 60 years and over). Studies

were restricted to English language publications, and instrument evaluations to those conducted in populations within Europe, North America, and Australia. Clinician-assessed instruments, single-item and anthropometric measures, and radiographic and imaging techniques were excluded. Instruments without empirical evidence of reliability or validity were excluded (Eiser and Moorse, 2001).

c) Data extraction

Data extraction followed pre-defined criteria (Chapter 1) and included both study-specific issues, such as study design and respondent characteristics, and instrument-specific issues, for example, type and description of instrument, including the domains of health status covered, response format, length, and evidence of measurement and practical properties (McDowell and Newell, 1996; Fitzpatrick et al., 1998; Garratt et al., 2002b) (Table 2.1).

d) Format of the reviews

The summary of evidence follows that of previous reviews (McDowell and Newell, 1996; Fitzpatrick et al., 1998). Detailed reviews of generic and older people-specific instruments are found in Chapters 4 and 5, respectively. The following information is provided for each instrument:

Title
The instrument title as given by the original developer. Instrument developers, year of original publication, and subsequent revision.

Description
The purpose and proposed application of each instrument as defined by the developers.

Instrument development, including item derivation, is summarised where available. Instrument content, the domains of health status covered, for example, pain and social well-being, the number of items, response options, and method of scoring are reported (Table 2.1). Instrument modifications are described. Where inconsistencies in the reporting of instruments were identified, contact with the development authors (older people-specific instruments only) was sought; where clarification was not possible, inconsistencies are highlighted.

Measurement properties
For all instruments published evidence of measurement properties is summarised under the following sub-headings:
- reliability
- validity: i. socio-demographic variables and health-service use
  ii. construct validity: other instruments
  iii. other types of validity
- responsiveness
- precision

Where evidence is not available, sub-headings are excluded. Unless otherwise stated, all extracted results are significant at the 5% level.

Measurement properties are specific to the population and setting in which an instrument is used (Streiner and Norman, 1995; Fitzpatrick et al., 1998). Study-specific information relating to study design and setting, for example, whether the assessment was community- or hospital-based, population characteristics including inclusion/exclusion criteria, intervention(s), duration of the study and follow-up, and mode of questionnaire administration, informs the interpretation of instrument performance and clinical usefulness (Table 2.1). Study-specific information summarises population and study characteristics, age, sex, and the presence of cognitive impairment or co-morbidity.

Practical properties
Where available, published evidence of acceptability and feasibility is summarised.

e) Review summaries

Reviewed evidence is summarised using the instrument-specific sub-headings shown in Table 2.1. Evidence for measurement and practical properties was assessed using accepted criteria (Streiner and Norman, 1995; McDowell and Newell, 1996; Fitzpatrick et al., 1998) (detailed in Chapter 1). Summaries for generic and older people-specific instruments are found in Chapters 6 and 7, respectively.

Fitzpatrick et al. (1998) list the domains of health status most frequently identified in the literature as relevant to patient-reported health instruments, as shown in Table 2.2. To support comparison between instruments, instrument content was reviewed against this general classification.

The number of studies in which the instrument has been evaluated is provided; where several publications relate to the same study population, this is indicated. The grading scheme in Table 2.3 gives a summary of the thoroughness and results of instrument evaluation, and was informed by previous work (McDowell and Newell, 1996).

The discussion chapter (Chapter 8) summarises the current state of health-related quality of life measurement for older people, and suggests areas for future evaluative work.

Table 2.1 Data extraction

Study-specific		*Instrument-specific*					
Study design	**Population**	**Description**	**Reliability**	**Validity**	**Responsiveness**	**Precision**	**Acceptability**
Design Setting Intervention(s) Duration Follow-up Mode of administration	Age Sex Co-morbidity Inclusion/ exclusion criteria	Title Purpose Application Development: item derivation Item content Domains Scaling Scoring Modifications	Internal consistency: item-total correlation, Cronbach's alpha Test-retest: health transition, retest period, correlation	Content Face Construct: internal, external[1] Other: domain, predictive	Distribution-based change Anchor-based change Time period	End effects: floor, ceiling Item/score distribution	Response rate Completion rate, time needed Feasibility

[1] Where hypothesised relationships between variables are stated, these are indicated.

Table 2.2 Domains included in patient-reported health instruments (Fitzpatrick et al., 1998)

I Physical Function Mobility, dexterity, range of movement, physical activity Activities of daily living: ability to eat, wash, dress	
II Symptoms Pain Nausea Appetite	Energy, vitality, fatigue Sleep and rest
III Global judgements of health	
IV Psychological well-being Psychological illness: anxiety, depression Coping, positive well-being and adjustment, sense of control, self-esteem	
V Social well-being Family and intimate relations Social contact, integration, and social opportunities Leisure activities Sexual activity and satisfaction	
VI Cognitive functioning Cognition Alertness Concentration	Memory Confusion Ability to communicate
VII Role activities Employment Household management	Financial concerns
VIII Personal constructs Satisfaction with bodily appearance Stigma and stigmatising conditions Life satisfaction Spirituality	
IX Satisfaction with care	

Table 2.3 Summary of measurement and practical properties of reviewed instruments
(after McDowell and Newell, 1996)

Summary of evidence			
Thoroughness		**Results**	
0	No reported evidence of testing	0	No numerical results reported
+	Basic information only	+	Weak evidence
++	Several types of tests, or several studies reporting evidence	++	Adequate evidence
+++	All major forms of evaluation reported; several good quality studies	+++	Good evidence

Chapter 3: RESULTS

a) Search results: identification of articles

At the time of the review, the PHI database contained 6,555 records (up to December 2002). The primary search strategy, using the terms 'old*, elder*, senior*' generated 784 records, as shown in Table 3.1. All abstracts were reviewed. When assessed against the review inclusion criteria, 113 articles were retrieved and reviewed in full. Of these, 88 articles were included in the review (Table 3.1).

Table 3.1 Number of articles identified by the literature review

Source	*No. of abstracts reviewed*	*No. of articles reviewed*	*Number of articles included in review*
PHI database: original search (up to December 2002)	784	113	88
PHI database: instrument name search (secondary search)	-	-	10
Additional electronic database searches (2001-2003)	Generic search: 443 Older people-specific: 64	Generic search: 71 Older people-specific: 19	Generic search: 23 Older people-specific: 10
Handsearching	-	-	20
TOTAL	-	-	**151**

A secondary search of the PHI database used the names of identified generic and older people-specific HRQL measures. Once overlap with previously identified abstracts was accounted for, ten additional articles were included in the review (Table 3.1).

A further search of electronic databases was conducted to extend the search period to September 2003. A total of 90 articles was retrieved in full and reviewed; 33 articles were included in the review (Table 3.1).

Hand-searching and checking the reference-lists of articles generated a further 20 articles for inclusion in the review. A total of 151 published articles provided evidence of measurement and practical properties for the instruments included: 112 articles reported evidence for the generic instruments, 46 articles reported evidence for the older people-specific instruments. Seven articles reported evidence for both generic and older people-specific instruments.

b) Identification of patient-reported health instruments

15 generic and 18 older people-specific instruments were included in the review. The developmental and evaluative studies relating to the instruments reviewed are listed in Tables 3.2, 3.3, and 3.4. Those relating to generic instruments are shown in Tables 3.2 and 3.3; those for older people-specific instruments are shown in Table 3.4.

The generic and older people-specific instruments are reviewed in Chapters 4 and 5, with related summaries in Chapters 6 and 7, respectively.

c) Existing reviews of patient-reported health instruments applied with older people

Two detailed reviews of patient-reported health instruments applied in the assessment of older people (Fletcher et al., 1992; Albert, 1997) and five summary reviews (Kane and

Kane, 1984; Rubenstein et al., 1989; Bowling, 1995; Kliempt et al., 2000; Philp et al., 2001) were identified and are summarised below. Specific instrument recommendations are generally not given. Most authors indicate that selection should be informed by the needs of the specific population, the setting and purpose of the assessment, the structure and content of the instrument, and available evidence of its performance in the specified context (Kane and Kane, 1984; Fletcher et al., 1992; Albert, 1997; Kleimpt et al., 2000).

The remit of a working group from the Royal College of Physicians and British Geriatrics Society was to identify a set of instruments suitable for application in the audit and evaluation of the health status of older people (Fletcher et al., 1992). Following a structured review of the Medline electronic database (using keywords 'quality of life'), patient-reported instruments used in the evaluation of the health status of older people were identified, and article information was extracted in relation to their measurement properties and practical issues. Instruments had to be acceptable for everyday use in community-based, acute, or long-term geriatric health-care.

Two multidimensional instruments were reviewed: the Nottingham Health Profile (NHP) (Hunt et al., 1980) and the Sickness Impact Profile (SIP) (Bergner et al., 1981). Advantages of the NHP included its six-domain structure and associated short completion time; however, ceiling effects were reported. The SIP, while providing a more detailed and comprehensive assessment, with greater evidence for its measurement properties, was associated with increased completion time. However, uncertainty surrounded the application of both instruments in the acute and longer-term care assessment of older people. In conclusion, the comprehensiveness and responsiveness of the SIP was commended; where the length and inappropriateness of certain domains is cause for concern, the use of selected SIP domains may be appropriate.

Generic instruments were reviewed if they had been applied in the assessment of older patients receiving long-term care and 'minimally addressed' both physical and mental health, or were developed as generic measures of quality of life (Albert, 1997). 13 instruments were reviewed for evidence of instrument development, measurement properties, the assessment of four key 'factors' of HRQL, namely physical function, mental health, social function, and somatic symptoms, and the inclusion of a patient global rating of HRQL. These instruments were the COOP charts, EuroQol, Functional Independence Measure (FIM), Functional Status Questionnaire (FSQ), Health Utilities Index Mark III (HUI3), McMaster Health Index Questionnaire (MHIQ), Minimum Data Set (MDS), NHP, Rosser Index (RI), SF-36, SIP, Quality of Well-being Scale (QWB), and the World Health Organisation Quality of Life questionnaire (WHOQOL).

Most instruments had undergone extensive development and testing, and all used very different response scales. Most instruments met accepted standards for reliability and validity, but no instrument covered all four domains of HRQL. The narrow definition of health status provided by the developers of the HUI was considered too restrictive. All instruments covered broad aspects of physical function in detail, the MHIQ and SIP being the most detailed.

The FIM was the only instrument not to assess mental health. Whilst five instruments, namely the FSQ, SF-36, MHIQ, RI, and WHOQOL, assessed both positive and negative aspects of mental health, the HUI assessed only positive health. The remaining six instruments assessed only negative mental health. Only the FSQ and MDS included the

assessment of both contact and quality of social functioning. The EuroQol, FIM, HUI, QWB, and RI do not assess social functioning.

Although the assessment of somatic symptoms was variable, pain assessment was included in most instruments. Only five instruments, namely the FIM, MDS, SIP, QWB, and the WHOQOL, assessed cognitive function (alertness and communication). Albert (1997) highlights the appreciation of respondents' cognitive ability as an important issue in the utilisation of self-completed questionnaires. Fewer than 50% of the instruments requested patient-reported global ratings of health. Item content and domain coverage was most extensive for the SIP and QWB.

Kane and Kane (1984) review three patient-reported multidimensional instruments, namely SIP (Bergner et al., 1981), the OARS-OMFAQ (George and Fillenbaum, 1985), and the Comprehensive Assessment and Referral Evaluation (CARE) (Gurland et al., 1977; Gurland and Wilder, 1984). Due to differential instrument performance in different settings, no single instrument was recommended over another (Kane and Kane, 1984).

Rubenstein et al. (1989) provide a broad overview of the purpose, benefits, and difficulties of health status assessment in older people across different settings. Additionally, a limited list of instruments that are short, easily administered, reliable, valid, and which measure three aspects of daily life (physical, psychological, and social functioning), is provided. Specific instruments are not recommended; rather, settings and core assessment domains with examples of suitable instruments are given. For instance, the SF-36 or SIP may be useful in a screening programme for cognitively intact older people.

Although consensus on instrument selection for the multidimensional assessment of health in older people is lacking, the OMFAQ is highlighted as an older people-specific instrument that addresses most relevant assessment domains (Bowling, 1995). Although the length increases respondent burden, there is good evidence of reliability and validity.

Kliempt et al. (2000) provide a non-critical descriptive summary of generic and older people-specific measures of general health status and quality of life applied in the assessment of older people. Three textbook reviews of outcome measures (Bowling, 1995; McDowell and Newell, 1996; Bowling, 1997) and a Medline search (not detailed) provided the source for instruments. Measurement properties are not summarised; the authors indicate that this should be assessed within the context of instrument application.

Eight generic instruments, namely the Duke Health Profile (DHP), EuroQoL, FSQ, MHIQ, NHP, SF-36, SF-20, and SIP, and four older people-specific instruments, namely CARE, Philadelphia Geriatric Centre Multilevel Assessment Instrument (PGC MAI), OMFAQ, and the Self-Evaluation of Life Function Scale (SELF), are described. Five instruments originally designed for the assessment of chronic disease or illness, but subsequently applied in the assessment of older people, are also described, namely the Arthritis Impact Measurement Scale (AIMS), Karnofsky Performance Index (KPI), London Handicap Scale (LHS), Physical and Mental Impairment of Function Evaluation (PAMIE), and Spitzer Quality of Life Index (SQL).

With the exception of the AIMS, DHP, KPI, LHS, MHIQ, and PAMIE, all of these instruments are included in the present review (see Chapters 4-7).

An unstructured review of three generic instruments, namely COOP charts, SF-36, and SIP, and the older people-specific EASY-Care (Philp, 1997), concluded that all were comprehensive measures of health status suitable for cross-cultural application (Philp et al., 2001). With the exception of the SF-36, all were considered suitable for multidisciplinary use and, except for the SIP, all were considered sufficiently brief. Only the COOP charts had evidence of reliability and validity supporting their application in routine practice. The review team recommended the EASY-Care and COOP charts on the basis that they were comprehensive and had been developed for use in primary health-care.

Table 3.2 Developmental and evaluative studies relating to the generic instruments excluding the SF-36 (see Table 3.3)

Instrument	*Study*	*Administration*	*Population (n)*	*Range/mean age in yrs (sd)*	*Respondent characteristics*
Assessment of Quality of Life (AQoL)	Osborne et al. (2003)[d]	Self	Trial: care co-ordination (526) *vs* usual care (530) Australia	36% 60-75 64% >75.0	Community Chronic disease
COOP charts	Nelson et al. (1990)	Interview	Samples from outpatient clinics: a. veterans clinic (231) b. primary care (2349) USA	a. 70.0 b. 58.0	Community a. 100% males b. 57% females
	Siu et al. (1993a)	-	Residential home: concurrent admissions (155), 6-month follow-up USA	84.3	84% females Independent in basic ADL; able to walk independently to a dining room.
	Siu et al. (1993b)	-	*as above*	*as above*	*as above*
	Doetch et al. (1994)	Self	Family practice outpatient clinic: consecutive patients (100) USA	71.6 >65 yrs	Primary care 48% females
	Jenkinson et al. (1997)	Interview	General practitioner [GP] referral to hospital for congestive heart failure: consecutive patients (61) UK	81.0	Hospital outpatients 70% females
Bodily Pain chart only	Manz et al. (2000)	Interview	Nursing facility residents: random sample (100) USA	83.0	74% females
WONCA/COOP	Kempen et al. (1997)	Interview	Sub-sample of Groningen Longitudinal Ageing Study (575) Netherlands	74.9	Community 75% females
	Coast et al. (1998)	Self or interview	Trial participants (214): hospital at home *vs* routine hospital care UK	79.0	Various settings 70% females Varied case-mix: mainly fractures, stroke, hip/knee replacements; cognitively impaired patients excluded
	Philp et al. (2001)	Interview	GP random sample (595), 9 practices UK	>75 yrs	Community: home or residential care
	Van Balen et al. (2001)	Interview (25% proxy)	Hip fracture patients: post-operative cohort (102) Netherlands	83.0	84% females Cancer and multiple trauma excluded
	Van Balen et al. (2003)	Interview (26% proxy)	Hip fracture patients: post-operative cohort (208) Netherlands	83.0	79% females 20% dementia (proxy completion) Patients with cancer or multiple trauma excluded
EuroQol	Brazier et al. (1996)	Self	GP random sample: trial participants (377) UK	80.1	Primary care 100% females
	Coast et al. (1998)	Self or interview	Trial participants (214): hospital-at-home *vs* routine hospital care UK	79.0	Various settings 70% females Varied case-mix: mainly fractures, stroke,

					hip/knee replacements; cognitively impaired patients excluded
	Burstrom et al. (2001)	Self	General population survey (11,698; 2865 >60 yrs) Sweden	16-84 25% >60.0	-
	Degl'Innocenti et al. (2002)	Self	HRQL sub-study (2791) of multi-centre Study on Cognition and Prognosis in the Elderly [SCOPE] Italy	70.0-89.0	Patients with cognitive impairment, cerebrovascular accident [CVA], myocardial infarction [MI], liver/kidney disease, alcoholism, depression, or psychosis excluded
	Tamim et al. (2002)	Interview or telephone (some proxy)	Trial participants discharged from Emergency Dept. and at risk of functional decline (388) *vs* not at risk (132); subject-proxy pairs (231) Canada	76.4 (6.9)	Subject-proxy group: 53% subjects, 73% proxies female Patients without available caregiver, cognitively impaired patients, nursing-home residents excluded
	Tidermark et al. (2002a)	Self	Prospective study: internal fixation [IF] of femoral neck fracture [FNF] (90) Sweden	80.0 (7.0) >65 yrs	Living independently pre-injury 73% females Cognitively impaired patients excluded
	Tidermark et al. (2002b)	Self	Prospective study: IF of FNF (90; 65 at 1 yr) Sweden	80.0 (7.0)	*as above*
	Tidermark et al. (2003a)	Self	Trial: IF *vs* total hip replacement [THR] for acute displaced FNF (102; 95 at 4 months; 84 at 2 yrs) Sweden	80.0 (6.0) >65 yrs	Living independently pre-injury 79% females Cognitively impaired patients excluded
	Tidermark et al. (2003b)	Self	Post-operative review: THR for acetabular fracture (10) Sweden	57.0-87.0 73.0	30% females Mean follow-up 38 months, range: 11-84 months Cognitively impaired patients excluded
	Hage et al. (2003)	Self	Trial: cardiac rehabilitation *vs* usual care, 1-year follow-up (93) Sweden	76.0	Admitted to coronary care unit with acute coronary event 76% males
Functional Status Questionnaire (FSQ)	Tedesco et al. (1990)	Interview	Convenience sample of hospital patients with aortic stenosis: a. undergoing valvuloplasty (23) b. no operation (14) USA	a. 78.0 b. 75.0	a. 43% females b. 31% females
	Reuben et al. (1995)	Self	Sample (83) drawn from 3 settings: meal sites, seniors' recreation sites, housing units USA	76.0	Community 54% females
	Yarnold et al. (1995)	Self	General medicine: convenience sample (40), non-geriatric (85) USA	72.6	Ambulatory 83% females
	Sherman & Reuben (1998)	Interview	Elders with up to 4 health states: incontinence, depression, reduced function, falls (363) USA	75.9	Community Nursing-home residents and cognitively impaired patients excluded

	Cleary & Jette (2001)	a. Self b. Telephone c. Telephone	a. 6-hospital study: transurethral prostatectomy (2484) b. PORT: cardiac catheterization for MI (n not recorded) c. Cooperative Cardiovascular Project: acute MI (3263) USA	a. 69.0 b. 27% >75 c. 73.4	Hospital a. 100% males b. 40% females c. 42% females
	Brach et al. (2002)	Self	High functioning women: participants in walking intervention trial (170) USA	56-84 74.3 (4.3)	Community 100% females Cognitively impaired and non-community dwelling patients excluded
Goteborg Quality of Life Instrument (GQL)	Andersson et al. (1995)	Self	Convenience sample registered with hearing center (63) Sweden	69.9	Community 44% females Retired and hearing-impaired
	Nygren et al. (2001)	Self	Registered with Occupational or Physical Therapy services (233) Sweden	78.0	Community 59% females Cognitively impaired patients excluded
Health Status Questionnaire-12 (HSQ-12)	Bowling & Windsor (1997)	Interview	Office for National Statistics Omnibus Survey (375) UK	>65.0	-
	Pettit et al. (2001)	Interview	Random sample of community-dwelling older people in London (544) UK	74.0	Community 59% females 9% diagnosed with dementia
Index of Health-related Quality of Life (IHQL)	Livingston et al. (1998)	Interview	Random sample of community-dwelling older people in London (700) UK	75.7	Community 64% females
Nottingham Health Profile (NHP)	Hunt et al. (1980)[d]	Interview	a. physical exercise programme (50) b. GP random sample: no illness (28) c. social services luncheon club (49) d. GP purposive sample: chronic disease (86) UK	a. 68.9 b. 68.5 c. 74.4 d. 73.0	Community a. 100% males Cognitively impaired patients excluded
	Thorsen et al. (1995)	Self	a. fitness class (118) b. outpatients, low back/leg pain (68) c. outpatients, hip osteoarthritis (64) Denmark	a. 70.0 b. 74.0 c. 74.0	Community Females: a. 81% b. 81% c. 77%
	Crockett et al. (1996)	Interview	Outpatients, chronic obstructive airways disease (60) Australia (UK version)	68.6 (6.2)	Community 46.7% females Cognitively impaired patients excluded
	Stadnyk et al. (1998)	Interview	Inpatient *vs* outpatient rehabilitation (146) Canada	57% >80.0	Frail older people 64% females Cognitively impaired patients excluded

	Sharples et al. (2000)	Interview	GP random sample (481) from 78 practices in East Anglia UK	77.0	Community. 48% females Institutionalised or hospitalised patients and those too ill/cognitively impaired for interview excluded
	Mitchell et al. (2001)	-	Surgical fixation proximal femoral fracture (80): trial-specific quadriceps exercise *vs* standard physiotherapy UK	80.0 (1.2)	84% females Patients with cognitive impairment and pre-morbid inability to walk excluded
	Van Balen et al. (2001)	Interview (25% proxy)	Hip fracture patients: post-operative cohort (102) Netherlands	83.0	84% females Cancer and multiple trauma excluded
	Van Balen et al. (2003)	Interview (26% proxy)	Hip fracture patients: post-operative cohort (208) Netherlands	83.0	79% females 20% with dementia (proxy completion) Cancer and multiple trauma excluded
Quality of Life Index (QLI)	Kleinpell & Ferrans (2002)	Self	Follow-up of medical/surgical patients discharge from intensive care unit: a. middle-aged (45-64 yrs) (36) b. young-old (66-79 yrs) (76) c. old-old (>80.0 yrs) (52) USA	total 73.7 (11.4)	56.7% females Included if not hospitalised at time of follow-up and alive 4-6 months following hospital discharge
Quality of Well-Being Scale (QWB)	Andresen et al. (1995)	Telephone	Health plan enrolees from 3 clinics: random sample (200) USA	72.5	Community
	DeBon et al. (1995)	Interview	Residents from convalescent homes and senior centers (71) USA	79.9	Various community settings 79% females
	Groessl et al. (2003)	-	Longitudinal cohort of osteoarthritis patients, 1 year with education (363) USA	69.2 (5.6)	Community 64.2% females
QWB Self-administered (SA)	Andresen et al. (1998b)	Self	3 primary care offices: random sample (282) USA	74.7	Community 59% females
SF-12	Schofield & Mishra (1998)	Self	Medicare database or telephone directory: random sample (221) Australia	range: 70.0-74.0	100% females
	Lim & Fisher (1999)	Self	Hospital Heart and Stroke Register: random sample (2341; 1425 >65yrs) Australia	39% <65.0 37% 65-74 26% >75.0	39% females Heart- and stroke-related conditions
SF-12 and York SF-12	Iglesias et al. (2001)	Self	Trial: hip protectors for fracture prevention (422) UK	>70.0	Community 100% females
	Pettit et al. (2001)	Interview	Random sample of community-dwelling older people in London (544) UK	74.0	Community 59% females 9% diagnosed with dementia

Standard and revised scoring	Resnick & Parker (2001)	a. Interview b. Telephone	a. Independent, retirement community (182) b. Home discharge from acute care (211) USA	a. 86.0 (6.1) b. 73.0 (6.5)	Community Females: a. 78% b. 60% Cognitively impaired patients excluded
Standard and revised scoring	Resnick & Nahm (2001)	Interview	Independent adults from a retirement community (182) USA	86.0 (6.1)	Community 78% females Cognitively impaired patients excluded
	Theiler et al. (2002)	Self	Outpatients with osteoarthritis of hip/knee: prospective evaluation 3-week trial of Rofecoxib (92) Switzerland	69.0 (8.0)	Community 68% females
SF-20	Siu et al. (1993a,b)	-	Residential home: concurrent admissions (155), 6-month follow-up USA	84.3	84% females Independent in basic ADL; able to walk independently to a dining room
	Carver et al. (1999)[f,c]	Interview	Community-dwelling, random sample (333) Canada	76.0	Community 58% females Cognitively impaired patients excluded
Sickness Impact Profile (SIP)	Goldsmith & Brodwick (1989)	-	Family practice patients with chronic illness (62); clinic-based use of SIP: a. 14 clinicians instructed b. 13 clinicians not instructed USA	a. 70.3 (8.7) b. 66.9 (9.2)	Females: a. 76% b. b. 74%
	Rothman et al. (1989)	Interview	Community and Veterans Association nursing-home residents (168) USA	68.0 (11.3)	Various community settings 100% males Patients with psychosis or unable to participate in interview excluded
	Weinberger et al. (1991)[f]	Interview	Veterans in receipt of medical centre care: convenience sample (25) USA	73.5	100% males
	Andresen et al. (1995)[f]	Self	Health plan enrolees from 3 clinics: random sample (200) USA	72.5	Community
	Page et al. (1995)	Self or proxy	Coronary artery bypass surgery (18) Canada	70.4 (5.1)	14.3% females Cognitively impaired patients excluded
	Larson et al. (1998)	Self	Hospital register for chronic obstructive airways disease: (72) USA	69.5 (6.5)	37.5% females Cognitively impaired patients excluded
	Kleinpell & Ferrans (2002)	Self	Follow-up of intensive care unit discharges (medical/surgical): a. middle-aged: 45-64 yrs (36) b. young-old: 66-79 yrs (76) c. old-old: >80.0 yrs (52)	73.7 (11.4)	56.7% females Not hospitalised at time of follow-up. Alive 4-6 months following hospital discharge

			USA		
SIP physical summary only	Morishita et al. (1995)	Interview or telephone	Geriatric outpatient clinic: convenience sample (31) USA	77.3	74% females Cognitively impaired patients excluded
	Liddle et al. (1996)	Self or telephone	Occupational therapy consumers (167) Australia	81.6	Community 68% females
	Andresen et al. (1998a)[f]	Self	3 primary care offices: random sample (282) USA	range: 65.0-96.0	Community 57% females Patients in long-term care or unable to communicate in writing excluded
	Andresen et al. (1998b)[f]	Self	*as above*	*as above*	*as above*
SIP(68) Mobility only	Jannink-Nijlant et al. (1999)	Self	GP independent-living random sample (84) Netherlands	74.1 (3.2)	Community 58% females
Spitzer Quality of Life Index (SQL)	Stadnyk et al. (1998)	Interview	Inpatient or outpatient rehabilitation facilities (146) Canada	57% >80.0	Various settings 64% females
	Simpson (2002)	Not recorded	Post-hip fracture prospective study: usual care (20) *vs* transition rehab. programme (30) Canada	>60.0	4-week follow-up Residents of long-term care facilities and those with cancer excluded
	Carver et al. (1999)[f,c]	Interview	Community-dwelling random sample (333) Canada	76.0	Community 58% females Cognitively impaired patients excluded

Key: [d] developmental [f] floor, [c] ceiling effects reported

Table 3.3 Developmental and evaluative studies relating to the SF-36

Study	*Administration*	*Population (n)*	*Range/mean age in yrs (sd)*	*Respondent characteristics*
Anderson et al. (1996)[c]	Interview	Patients post-stroke (90) Australia	36-92 72.0 (12.0)	47% females 23% from hostel or nursing homes
Andresen et al. (1999)[f,c]	Interview	Nursing-home residents (97) USA	80.1	80% females Patients with severe dementia or communication problems excluded
Andresen et al. (1998a,b)[f,c]	Self	3 primary care offices: random sample (282) USA	>65	Community 57% females Patients in residential care excluded
Andresen et al. (1996)	Self	2 primary care offices: random sample (253) USA	76.5	Community 63% females Patients in residential care excluded
Andresen et al. (1995)[c]	Self	Health plan enrolees: random sample from 3 clinics (200) USA	72.5	Community
Baldassarre et al. (2002)	Self	Patients undergoing coronary artery bypass surgery: elective (15) *vs* emergency (15) Canada	63-87 69.3 (6.0)	100% females Cognitively impaired patients excluded
Ball et al. (2001)	Interview (and proxy)	Day-hospital patients (134), inpatients (30) UK	58-93 79.0	Various hospital and community settings Cognitively impaired patients excluded
Berkman et al. (1999)	Self	2 primary care offices: random sample (313) USA	>65.0	Community
Beusterien et al. (1996)[f,c]	Self	Depressed older people: drug trial participants (532) USA	60-86 67.0	Community 54% females
Bombardier et al. (1995)[f,c]	Self	Patients post-knee arthroplasty (1404) USA	67-99 74.8 (6.5)	72% females
Brazier et al. (1996)	Self	GP random sample of trial participants (377) UK	80.1	Community 100% females
Cochrane et al. (1998)	Self	a. Previously sedentary people participating in an exercise programme (55) b. Matched controls (55) UK	65-87 a. 74.4 b. 73.4	Community Females: a. 64% b. 58%
Crockett et al. (1996)	Interview	Chronic obstructive airways disease (60) Australia (UK version)	68.6 (6.2)	46.7% females 59/60 completion (98.3%) Cognitively impaired patients excluded
Dexter et al. (1996)	Interview	Outpatients in trial of patient management (1053) USA	64.0	65% female Patients with dementia or difficulty communicating and those resident in nursing homes excluded

Doraiswamy et al. (2002)	Self	Outpatients with moderate to severe depression in drug trial (100): use of Hamilton Rating Scale for Depression (HAMD) USA	60-88 70.2 (7.8)	57% females Patients with HAMD score <18 and actively suicidal excluded
Ekman et al. (2002)[c]	Interview	Inpatients receiving acute care for chronic heart failure (158); general population age- and sex-matched controls (94) Sweden	80.9	Hospital 45% females Cognitively impaired patients excluded
Fowler et al. (2000)	Interview Self 3/6 months	Geriatric day hospital: multidisciplinary rehabilitation (99) UK	>65.0	Rehabilitation for medical conditions Cognitively impaired patients excluded
Girotto et al. (2003)	Interview	Female patients post-mastectomy (372 aged <65 yrs, 28 aged >65 yrs) USA	93% <65 7% >65	100% females with breast cancer
Hage et al. (2003)	Self	Cardiac rehabilitation trial: intervention (44) *vs* control (44), mean follow-up 4.4 yrs Sweden	65-84 71.0	80% males Following acute cardiac event
Hamilton et al. (1996) (Fowler et al., 2000)	Interview Self 3/6 months	Geriatric day hospital: multidisciplinary rehabilitation for medical conditions (99) UK	66-99 81.5	74% females Cognitively impaired patients excluded
Harada et al. (2001) *PF, GH, MH, BP only*	Self	Convenience sample of more active elders from community centres (51), less active from retirement homes (36) USA	65-89 75 (6.0)	Various community settings 62% females Cross-sectional instrument evaluation Cognitively impaired patients excluded
Hayes et al. (1995)	Interview or self	GP outpatients and hospital inpatients (195) UK	65-103 77.0	Various settings 62% females
Heslin et al. (2001)	Interview	Population registers of those aged >70 yrs (4004) 6 European countries	78.0	62% female
Hill & Harries (1994) Hill et al. (1996)	Interview	Convenience sample of patients with mental health or continence problems (47) UK	majority 75-85	Community Females: males approx. 3:1 Cognitively impaired patients excluded
Ho et al. (2001)	Interview	GP random sample of patients with self-reported dyspnoea (452) UK	>70.0	Community Cognitively impaired patients excluded
Hobson & Meara (1997) *revised format*	Self	Parkinson's disease (66) UK	74.5	48% female Cognitively impaired patients excluded
Inaba et al. (2003)	Telephone	Following traumatic injury (128), mean follow-up 2.8 yrs (range: 1.5-4.5 yrs) Canada	74.0	41% female Cognitively impaired patients excluded
Irvine et al. (2000)	Interview	Regional home-care nursing agency: convenience sample (50) Canada	61.0	Community 60% females

Jaglal et al. (2000)	Interview	Post-hip fracture: convenience sample (43) Canada	80.9 (8.3)	Community 81.0% females Cognitively impaired patients excluded
Jenkinson et al. (1995)	Self	Patients with Parkinson's disease [PD], Oxford (95) and GP random sample, Sheffield (103) UK	65-74	-
Jenkinson et al. (1997)	Interview	Trial participants with symptomatic congestive heart failure (61); baseline and 1-month data UK	81.0	70% females
Larson et al. (1998)	Self	Hospital register for chronic obstructive airways disease (72) USA	69.5 (6.5)	37.5% females Cognitively impaired patients excluded
Lisse et al. (2001) [f]	Self	Osteoarthritis of knee or hip: pooled data from three Celecoxib trials (768), 12-week follow-up USA	74.8	67.1% females
Lyons et al. (1994) [c]	Interview	Random sample from local Family Health Services Authority [FHSA] register (216) UK	73.9	Community
Lyons et al. (1997)	Interview	Random sample from FHSA register (1608) UK	>70	63% females Some care-home residents 8% with cognitive dysfunction
Mallinson (1998)	Self	Physio-/occupational therapy consumers (56) UK	77.1	79% females
Mangione et al. (1993)	Self	Patients receiving major elective surgery for a range of conditions: a. 50-70 yrs (479) b. >70yrs (276) USA	67.0 (9.0)	Hospital Cognitively impaired patients excluded
McHorney et al. (1990)	Self	General population survey (623) USA	>60	5.8% (n=36) cognitively impaired
McHorney et al. (1994a)[f,c]	Self	Hospital and general practice clinics: random sample of patients with chronic medical and psychiatric conditions (total 3445) a. <65 years (2456) b. 65-74 yrs (700) c. >75 yrs (287) USA	18-98 58.0	Chronically ill: Medical Outcomes Study [MOS] longitudinal panel survey Females 61.7%.
McHorney et al. (1994b)[c]	Self (mail) or Telephone	General population survey random sample (2474); over-sampling of >65 yrs: self completion (533), telephone (184) USA	71% 18-65 29% >65	Community
McHorney (1996)	Self	Chronically ill (877) USA	>65	Chronically ill: MOS longitudinal panel survey

Morgan et al. (2002)	Interview	Neurologically healthy (93) Australia	72.37 (7.44)	Community Patients with cognitive or neurological impairment excluded
Murray et al. (1998)	Interview	Patients with chronic pain: a. community (15) b. low care (15) c. high care (15) d. institutions (45) Canada	79.0	73% females Patients with major cognitive deficits excluded
O'Mahony et al. (1998)[f,c]	Self	Post-stroke patients (73) UK	>45.0	Community Hospitalised patients and nursing-home residents excluded
Osborne et al. (2003)[f,c]	Self	Trial: care co-ordination (526) *vs* usual care (530) a. 36% 60-75 yrs b. 64% >75 yrs Australia	77.0 (9.7)	Community Patients with chronic diseases 63% females
Overcash et al. (2001)	Self	Older people with cancer (85), or living in the community (27) USA	65-90 75.0	Community 61% females Cognitively impaired patients excluded
Parker et al. (1998)	Interview or self	Inpatients [IP]: interview/self (533), outpatients [OP]: self (57), patients in general practice [GP]: self (37) UK	*median:* IP 76.0 OP 80.0 GP 77.0	Some patients with cognitive impairment
Pierre et al. (1998)	Patient: interview Proxy: self/telephone	Day hospital [DH] and Rehabilitation unit [RU]: patient and proxy (lay/professional) completion Canada	DH: 79.8 RU: 75.8	Various hospital settings 120 respondent pairs (proxy age not recorded) Cognitively impaired patients excluded
Rebello et al. (2001)	Interview	Dialysis-transplant programme (483); 183 >65 yrs) Spain	68-76 *median:* 72.0	Cognitively impaired patients excluded
Reuben et al. (1995)[f,c]	Self	Meal sites or recreation centres (53) USA	76.0	Community 54% females
Reza et al. (2002)	Interview	Trial: insulin therapy (30) *vs* oral glucose-lowering drugs (10) UK	72.5 (4.5) all >65.0	Patients with type 2 diabetes and poor glycaemic control; patients with acute illness during previous 6 months excluded
Schofield & Mishra (1998)	Self	Medicare database or telephone directory: random sample (221) Australia	70-74	100% females
Seki et al. (2003)	Self	Trial: a. cardiac rehabilitation (20) *vs* b. usual care (18) Japan	a. 69.3 (2.9) b. 70.1 (3.7)	100% males 6 months following major coronary event
Seymour et al. (2001)	Interview and proxy	Day hospital or rehabilitation wards (314) UK	79.7	68% females 33% cognitively impaired

Sharples et al. (2000)[c]	Interview	Random sample (481) from 78 GP practices in East Anglia UK	77.0	Primary care 48% females Cognitively impaired patients and those in residential care excluded
Sherman & Reuben (1998)[f,c] *PF only*	Interview	Elders (363) with up to 4 health states: incontinence, depression, reduced function, falls USA	75.9	Community Cognitively impaired patients and nursing-home residents excluded
Stadnyk et al. (1998)[f,c] *acute format*	Interview	Inpatient *vs* outpatient rehabilitation (146) Canada	57% >80.0	64% female Frail older people Cognitively impaired patients excluded
Suzuki et al. (2002)	Self	Day care: convenience sample (135) Japan	males 76.1 females 82.6	Community 68% female Cognitively impaired patients excluded
Tidermark et al. (2003a)	Self	Trial: IF vsTHR for acute displaced femoral neck fracture (102; 95 at 4 months; 84 at 2 yrs) Sweden	80.0 (6.0)	>65 yrs and living independently pre-injury 79% females Cognitively impaired patients excluded
Walters et al. (2001)[f,c]	Self	12 general practices: random sample (8117) UK	74.6	Community 58% females
Weinberger et al. (1991)	Interview	Medical centre care: convenience sample (25) USA	73.5	Community 100% males
Weinberger et al. (1994)	Telephone or interview	Patients prescribed >5 medications (31) USA	68.5	Community 100% males Cognitively impaired patients and those in residential care excluded
Wildner et al. (2002)	Interview	Survey: extremity fracture previous 10 years (146); age- and sex-matched controls (311) Germany	66.8	Community 57% females
Wolinsky & Stump (1996)	Interview	Serious chronic conditions or very old (1051) USA	64.0	66% females Cognitively impaired patients and those in residential care excluded
Wolinsky et al. (1998)[f,c]	Telephone or interview	Follow-up of very old patients or those with serious chronic conditions (786): see Wolinsky & Stump, *above* USA	64.0	*as above*
Wood Dauphinee et al. (1997)	Self or telephone	Outpatients and community-dwelling older people (120) Canada	70.1	Community 51% females Patients in residential care excluded
Yip et al. (2001)	Interview: patient, proxy	Older people and their proxies, geographically proximate and seen within previous week (32 pairs) USA	respondents: 63-94; 78.4 proxy: 19-86; 64.2	Community Females: 56.3% Cognitively impaired patients excluded.

Key: [d] developmental [f] floor, [c] ceiling effects reported

Table 3.4 Developmental and evaluation studies relating to older people-specific instruments

Instrument	*Study*	*Administration*	*Population (n)*	*Mean age in yrs*	*Respondent characteristics*
Brief screening questionnaire (BSQ)	Smeeth et al. (2001)	Randomised: postal *vs* interview	Random sample from 106 general practices. Postal survey (5277), interviews: lay (4893), nurse-led (6033) UK	>75.0	Community Terminally ill patients or those resident in long-stay hospitals or nursing homes excluded
Comprehensive Assessment and Referral Evaluation (CARE)	Gurland et al. (1977)	Interview	Randomly selected community residents: USA (445), UK (396)	>65.0	Community
CORE-CARE	Golden et al. (1984) Teresi et al. (1984a) Teresi et al. (1984b)	Interview	*as above*	>65.0	Community
SHORT-CARE	Gurland et al. (1984)	Interview	*as above*	>65.0	Community
Elderly Assessment Summary (EASY-Care)	Bath et al. (2000)	Interview	GP random selection: a. Belfast (179) b. Hampshire (238) UK	mean 81.0	Community: area of deprivation Females: Belfast 73%, Hampshire 65%
	Philp (1997)[d]	*not applicable (n/a)*	*n/a*	*n/a*	*n/a*
	Philp et al. (2001)	Interview	9 General practices: random sample (595) UK	>75.0	Community (home or residential care)
	Philp et al. (2002)	Interview	Day rehabilitation unit (50) UK	78.5	Community 72% females Patients with dementia, communication difficulties, or unstable medical conditions excluded
Functional Assessment Inventory (FAI)	Pfeiffer et al. (1981)[d]	Interview Some proxy	4 settings: a. nursing homes (63) b. congregate living (62) c. day-care (60) d. senior centers (59) USA	76.5	Community 74% females Interview 63.9%, proxy (those with low SPMSQ [Short Portable Mental Status Questionnaire] scores) 36-1%
	Cairl et al. (1983)	Interview	a. domiciliary care [DC] (57) b. nursing home [NH] (81) USA	a. 67.5 b. 77.5	Community Females: a. 9% b. 17%
	Robinson et al. (1986)	Interview	Hospital-at-home care scheme (30) USA	60-70	Community Housebound patients with multiple chronic illnesses; terminally ill patients excluded 6.6% females

	Pfeiffer et al. (1989)	Interview	5 settings: a. mental health facility [MH] (25) b. nursing home [NH] (25) c. visiting nurse service (25) d. senior center (25) e. well elderly: control (25) USA	mean range: 72.0 (e) to 83.0 (b)	Various settings Females: 52% [MH] to 80% [NH] Interview 59%, proxy (those with low SPMSQ scores) 41%
Geriatric Postal Screening Survey	Alessi et al. (2003)[d]	Self: postal	a. Survey development (2545) b. Initial testing (2382) USA	a. >65.0 b. 73.5	Community 96.8% males (veterans) Cognitive impairment not reported.
Geriatric Quality of Life Questionnaire	Guyatt et al. (1993b)[d]	Interview	Clinical trial (76): a. day hospital [DH] b. conventional care [CC] Canada	DH 79.6 CC 78.2	Community Frail elderly Cognitively impaired patients excluded
Geriatric Screening Questionnaire	Fernandez Buergo et al. (2002)[d]	Interview	GP primary care: random sample (300) Spain	74.0 (6.4)	Community 57.3% females
IOWA Self-Assessment Inventory (ISAI) (*preliminary*)	Morris & Buckwalter (1988)[d]	Self	Meal programme: attendants (63), housebound, receiving home-delivered meals (23) USA	-	Community
	Morris et al. (1989)	Self	4 settings (1153): public housing projects, meal sites, community groups, retirement homes USA	*mode:* 75.0-79.0	Community 77% females
Revised ISAI	Morris et al. (1990)	Self	*as above*, plus residents of local hospital auxiliary units and older people living independently in the community (420) USA	-	Various settings
LEIPAD	De Leo et al. (1998)[d]	Interview	Cross-cultural sample (586) Italy, Netherlands, Finland	-	Community Living in own home
	Condello et al. (2003)	Interview	Psycho-geriatric outpatients (60) and matched controls (50) Italy	74.7 (8.1)	Community 66.7% females Cognitively impaired patients excluded
OARS Multidimensional Functional Assessment Questionnaire (OMFAQ)	Fillenbaum & Smyer (1981)	-	49 proxy representatives for 130 patients (OMFAQ completed on joining a family medicine center) USA	70.2	Community 64-70% females
	Cairl et al. (1983)	Interview	Domiciliary care [DC] (57), Nursing home [NH] (81) USA	DC 67.5 NH 77.5	Community Females: DC 9%, NH 17%
	Harel & Deimling (1984)	Interview	Random sample (1834) from one geographic location USA	74.0	Community 65% females

	Fillenbaum (1985)	Interview	Random samples: a. community (998) b. statewide (1530) c. general elderly (1609) USA	>65.0	Community
Mental health domain only	Liang et al. (1989)	-	Random samples: a. urban setting (1834) b. statewide (2146) USA	a. >65.0 b. >60.0	Community Non-institutionalised
ADL/IADL only	Reuben et al. (1995)	Interview	3 settings (83): meal sites, recreation sites for seniors, housing units USA	76.0	Community 54% females
ADL/IADL only	Stadnyk et al. (1998)	Interview (from records)	Frail inpatients or outpatients (146) Canada	*mode*: >80.0	Community 64% female Cognitively impaired patients excluded
ADL/IADL only	Carver et al. (1999)	Interview	Community-dwelling (333) Canada	76.0	Community 58% females Cognitively impaired patients excluded
ADL/IADL only	McCusker et al. (1999)	Interview	Emergency department: patients or proxy completion (213) Canada	77.0	59% females
Physical health domain only	Jaglal et al. (2000)	Interview	Post-hip fracture, community-dwelling (43) Canada	80.9 (8.3)	81.0% females Cognitively impaired patients excluded
ADL/IADL only	Breithaupt & McDowell (2001)	-	Canadian National Survey: elderly caregivers (1364) Canada	-	-
	Osborne et al. (2003)	Self	Trial: care co-ordination (526) *vs* usual care (530) Australia	36% 60-75 64% >75.0	Community Patients with chronic diseases
Perceived Well-being Scale (PWB)	Reker & Wong (1984)[d]	-	Convenience sample: living in the community (33), community and institutionalised (238) Canada	>60.0	Various settings
	Cousins (1997)	-	Test-retest reliability (18) and validity evaluations (327) Canada	>68.0	Setting not clear 100% females
Philadelphia Geriatric Center Multilevel Assessment Instrument (PGCMAI)	Lawton et al. (1982)[d]	Interview	3 settings: a. independent-living (426) b. high-intensity in-home care: dependent (99) c. institution waiting-list: dependent (65) USA	a. 74-76 b. 75-80 c. 79-80	Community Females: a. 49-57% b. 16-72% c. 77-91%

	Wissing & Unosson (2001)	Interview	Patients with leg ulcers in primary care and dermatology clinics: a. baseline (70) b. 4-yr follow-up (38) Sweden	a. 81.4 b. 78.0	Community 70% females
	Wissing & Unosson (2002)	Interview	Primary care and dermatology clinics: patients with leg ulcers (70) *vs* no ulcer (74) Sweden	81.4	Community 70% females
Quality of Life Cards	Rai et al. (1995)[d]	Interview	Day hospital: a. Continued care (30) b. Acute/rehabilitation wards (30) Netherlands	a. 83.5 b. 79.0	Hospital Females: a. 80% b. 70%
Quality of Life Profile-Senior Version (QOLPSV)	Raphael et al. (1995a,b)[d] Raphael et al. (1997)	Self	Older population (205) Canada	73.0	Community 77% females
	Irvine et al. (2000)	Interview	Regional home-care nursing agency: convenience sample (50) Canada	61.0	Community 60% females
Quality of life: well-being, meaning and value (QLWMV)	Sarvimäki & Stenbock-Hult (2000)[d]	Interview (home)	Non-institutional settings (300) Finland	75.0-97.0	Community 71% females
Self-evaluation of Life Function (SELF) Scale	Linn & Linn (1984)[d]	Self	6 settings (548): hospital, nursing homes, outpatient clinics, residential housing, trailer park residents, patients in psychiatric care or counselling USA	70.4	Various community settings 55% females Cognitively impaired patients excluded
SENOTS program and battery	Stones & Kozma (1989)[d]	Computer (self) or interview	2 settings: community (80), residential institutions (80) Canada	77.8	Various settings 50% females Patients with physical and cognitive impairments excluded
The Wellness Index (WI)	Slivinske et al. (1996)[d]	Self	(463); nursing homes through volunteer programs USA	73.4	Various community settings 77% females

Key: [d] *developmental*

Chapter 4: INSTRUMENT REVIEWS - GENERIC INSTRUMENTS

a) Assessment of Quality of Life Instrument (AQoL) (Hawthorne et al., 1997)

The Assessment Quality of Life Instrument (AQoL) was developed in Australia during the 1990s to provide a generic measure of health-related quality of life (HRQL) suitable for the evaluation of a wide range of health-care interventions and the economic evaluation of health-care programmes (Hawthorne et al., 1997; Osborne et al., 2003).

Literature reviews and existing instruments informed five key domains: illness (Ill), independent living (IL), physical ability (PA), psychological well-being (PWB), and social relations (SR). Item content was derived from the literature and interviews/focus groups with 24 hospital-based clinicians. The initial item pool was administered to patients (n=143) and community residents (n=112). Item analysis, factor analysis, and reliability testing informed item reduction. The final instrument contains 15 items across the five domains (three items per domain), as shown in Table 4.1. The AQoL may be self-administered by the respondent, or administered by interview or telephone.

Statements refer to important aspects of HRQL. Respondents select the response that best describes their current state/ability/relationship, etc. Items sum to give a 0-9 domain score where 9 is the worst HRQL, or an index score of 0-45 where 45 is the worst HRQL. A utility score is derived from four of the five domains, and ranges from –0.04 to 1.00 where 0 is equivalent to death and 1.00 is best HRQL. A computer programme supports this calculation. The Illness domain describes the use of health-care resources and does not contribute to the utility score (Osborne et al., 2003). Factor analysis supported the five domain scores and the inclusion of four domains in the utility index; each domain contained three items.

There has been one evaluation of the AQoL. This included a community-based older population in Australia (Osborne et al., 2003), as shown in Table 3.2. This study calculated a utility score and hence only four domains are used.

Reliability

Internal consistency reliability of the AQoL utility was 0.73, and for four individual domains ranged from 0.43 (PA) to 0.76 (IL), as shown in Table 4.1. There is no evidence of test-retest reliability.

Validity

(i) Socio-demographic variables and health-service use

As hypothesised, those scoring lower HRQL at baseline were greater consumers of health resources at 18 months (1.9 times those with the best HRQL at baseline), as shown in Table 4.3.

(ii) Construct validity: other instruments

Correlations between the AQoL and OMFAQ domain scores that had hypothesised associations ranged from –0.68 (AQoL utility with self-care) to –0.82 (AQoL-IL with self-care): see Table 4.3. Correlation between the AQoL and OMFAQ domains that did not have hypothesised associations ranged from 0.03 (IL with social resources) to –0.40 (SR with self-care).

Correlations between the AQoL utility and SF-36 domain scores that had hypothesised associations ranged from 0.34 (bodily pain) to 0.62 (physical function). Correlations with domains that did not have hypothesised associations ranged from 0.19 (role-physical) to 0.22 (role-emotional). Correlations with the SF-36 mental component and physical component summary scores were 0.41 and 0.37, respectively. Correlations between AQoL and SF-36 domain scores ranged from 0.04 (PA with SF-36 bodily pain and role-physical) to 0.64 (IL with SF-36 physical function).

(iii) Validity: other
Correlation between the AQoL utility score and individual AQoL domain scores were in accordance with hypotheses and ranged from 0.43 (PA) to 0.79 (IL). Correlation between AQoL domains ranged from 0.14 (PA with PWB) to 0.43 (IL with SR).

Responsiveness
Following the assessment of care coordination versus usual care in chronically ill community-dwelling older people, institutionalisation was defined as an external criterion of health deterioration at 18 months. The AQoL (utility and domains) was more responsiveness than the SF-36 and OMFAQ when evaluated by both Relative Efficiency and Receiver Operating Characteristic curves. High levels of responsiveness were also reported for the SF-36 physical component summary scale, physical function, and general health domains.

The AQoL utility index and three of the four domains (excluding psychological well-being) discriminated between baseline differences in people who remained community-dwelling at 18 months versus those requiring institutionalised care: AQoL utility and IL (p less than 0.001), AQoL SR and PA (p less than 0.05). The OMFAQ self-care domain was the most sensitive instrument to baseline differences.

Precision
A mean utility score of 0.33 (SD 0.25) without floor (0.1%) or ceiling effects (0.3%) was reported. Three domains had high mean values (SR and PWB 0.75, PS 0.85), and two domains had ceiling effects (SR 21.7%, PS 24.2%). There are currently no norm-based values available (Hawthorne et al, 1997).

Acceptability
The AQoL reading level of 71% (Flesch Reading Ease score) suggests that the instrument should be acceptable to most literate individuals, taking approximately 5-7 minutes for self-completion (Hawthorne et al., 1997). However, there is no evidence of acceptability in the older population.

b.i) COOP Charts for Primary Care Practice (Nelson et al., 1987)
b.ii) WONCA/COOP Health Assessment Charts (Froom, 1988; Landgraf and Nelson, 1992)

The Dartmouth Primary Care Cooperative Information Project developed the COOP charts in the late 1980s to provide a screening tool for use by doctors in routine practice (Nelson et al., 1987). The charts support the assessment of patient health status and functioning.

The original instrument, developed in the USA, has nine charts, each containing a single question about health, functioning, or quality of life during the previous month (Table 4.1). Eight charts assess bodily pain (BP), daily activities (DA), emotional condition/feelings (EC), physical fitness (PF), quality of life (QoL), social activities (SA), social support (SS), and current overall health (OH) perceptions (Table 4.1). An additional chart assesses change in overall health. Literature reviews, existing instruments, and discussion with practising physicians and experts in health status measurement informed item derivation (Nelson et al., 1990).

Following a multinational feasibility study, item content was revised to seven charts, omitting quality of life and social support, with a reduced recall period of two weeks (World Organisation of National Colleges, Academies and Academic Associations of General Practitioners and Family Physicians [WONCA]: WONCA/COOP Health Assessment Charts. Froom, 1988; Langraf and Nelson, 1992). Each chart within the WONCA/COOP includes a descriptive title, a question, and a pictorially illustrated five-point response scale, where five is the most severe limitation. Each represents a separate domain; an overall score is not calculated (McDowell and Newell, 1996). The charts can be self- or interview-administered.

Five articles describe the evaluation of the nine-chart COOP in older populations (Nelson et al., 1990; Siu et al., 1993a,b; Doetch et al., 1994; Jenkinson et al., 1997) and one article evaluates the BP chart alone (Manz et al., 2000), as shown in Table 3.2. All studies describe a range of care and community settings in US and UK populations. Five articles describe the evaluation of the WONCA/COOP (Kempen et al., 1997; Coast et al., 1998; Philp et al., 2001; Van Balen, 2001, 2003): see Table 3.2. These studies include community-based populations from the Netherlands and the UK, and two post-operative populations from Sweden. The results given below are derived from these articles.

Reliability
The results of reliability testing for the COOP charts are shown in Table 4.2. High levels of test-retest reliability (one-hour retest: mean 0.93, range 0.78-0.98) were found for the nine original COOP charts following completion by older male outpatients (Nelson et al., 1990). There is no evidence for the test-retest reliability of WONCA/COOP charts. Internal reliability testing is not appropriate for the COOP charts.

Validity
(i) Socio-demographic variables and health-service use
The impact of a range of socio-demographic variables and disease states on COOP scores was assessed (Nelson et al., 1990): see Table 4.3. All charts discriminated

between groups defined by sex, with higher scores in males. As hypothesised, the charts were sensitive to the impact of specific disease.

The COOP (nine charts) was completed by old-old people on admission to residential care and at a follow-up assessment (median 557 days) (Siu et al., 1993a,b). Following multivariate analyses, the COOP EC predicted future placement in skilled nursing care and the COOP OH was predictive of future hospitalisation. However, COOP change scores were not associated with subsequent placement in skilled care (Siu et al., 1993a).

At four months after surgical repair of hip fracture, the COOP PF and DA scores of survivors was significantly less than that of an age- and sex-matched reference population (p less than 0.05) (Van Balen et al., 2001).

(ii) Construct validity: other instruments
Correlation between the COOP (seven charts only) and seven RAND health status measures with hypothesised associations ranged from 0.59 (PF) to 0.69 (EC) (Nelson et al., 1990): see Table 4.3. Correlations between domains that did not have hypothesised associations ranged from 0.01 (PF with RAND emotional status) to 0.17 (BP with RAND social support). The correlation between COOP scores and self-reported symptoms in people with chronic illness was assessed. As hypothesised, the strongest correlation was between chronic illness and both PF and DA charts; the greater the number of reported symptoms, the worse the reported level of overall health (0.51).

The scores of several measures of depression suggested a prevalence of 16.5% to 34.7% in community-dwelling older people (Doetch et al., 1994). COOP EC scores suggested possible depression in 32.7% of participants. A concurrent review of medical records revealed that 7% of participants received medical care for depression. Correlation between the COOP EC and measures of depression ranged from 0.70 (Beck Depression Inventory-short form [BDI-SF]) to 0.74 (Brief Carroll Scale [BC]), Durham GRECC Scale (DURHAM). When defined by COOP EC scores as having no/slight emotional problems or moderate/extreme emotional problems, four measures of depression discriminated between groups, namely BC, BDI-SF, DURHAM, and Geriatric Depression Scale. The authors suggest that the COOP EC may be an appropriate screening measure for depression in older people.

In people representing a range of cognitive abilities, correlations between the COOP BP and several measures of pain ranged from 0.75 (Numeric Pain Rating Scale) to 0.89 (FACES scale) (Mantz et al., 2000).

Correlations between the WONCA/COOP PF chart and the Groningen Activity Restriction Scale were 0.50 (no illustrations) and 0.51 (with illustrations). Correlations between the WONCA/COOP OH chart and SF-20 global health were 0.56 (with illustrations) and 0.63 (no illustrations). Correlation between the WONCA/COOP EC chart and the SF-20 mental health was –0.71 (with/without illustration) (Kempen et al., 1997), as shown in Table 4.3.

In patients hospitalised mainly due to orthopaedic conditions, correlation between WONCA/COOP charts and the EuroQol that had hypothesised associations ranged from –0.53 (OH with EQ-5D index) to 0.74 (BP with EuroQol pain/discomfort) (Coast et al., 1998). Correlations between domains that did not have hypothesised associations

ranged from 0.13 (SA with EuroQol pain/discomfort) to 0.42 (OH with EuroQol self-care). Correlation between WONCA/COOP charts and the EQ-5D index ranged from –0.35 (change in health) to 0.59 (DA), and with the EuroQol thermometer ranged from –0.29 (PF, SA) to –0.65 (OH).

Following completion four months post-hip fracture, correlation between WONCA/COOP charts and Nottingham Health Profile (NHP) domain scores that had hypothesised associations ranged from 0.50 (OH with NHP energy) to 0.75 (DA with NHP physical mobility) (Van Balen et al., 2003). Correlation between charts and domains that did not have hypothesised associations ranged from 0.09 (OH with NHP sleep) to 0.38 (EC with NHP physical mobility). Correlation between the WONCA/COOP charts and the Rehabilitation Activities Profile (RAP) ranged from 0.08 (OH with RAP relationships) to 0.79 (DA with RAP mobility and personal care), and with the Barthel Index ranged from 0.18 (BP) to 0.75 (DA).

(iii) Other types of assessment
Correlation between the seven COOP charts (excluding change in health and quality of life) ranged from 0.02 (PF with EC) to 0.59 (DA with SA) (Nelson et al., 1990) see Table 4.3.

Correlation between WONCA/COOP charts with and without illustrations ranged from 0.01 (Change in Health [no illustrations] with BP [with illustrations]) to 0.64 (SA [no illustrations] with DA [with illustrations]) (Kempen et al., 1997). In people with a hip fracture, correlation between WONCA/COOP PA and DA charts was 0.68 (Van Balen et al., 2003).

Responsiveness
Three months following residential home admission, correlations between change scores for COOP charts and SF-20 domains with hypothesised associations ranged from 0.05 (PF with SF-20 physical function) to 0.74 (BP with SF-20 pain) (Siu et al., 1993b). Mean correlation between COOP charts and SF-20 domain change scores ranged from 0.18 to 0.37.

Performance-based tests (gait, balance, 50-foot walk time) were external criteria for improvement or deterioration in the same patient population. Change in gait and balance scores were statistically significant and in the hypothesised direction. The COOP PF had moderate responsiveness (effect size[ES] –0.35) where function had deteriorated (n=43), and poor responsiveness (ES –0.15) where function had improved (n=32); responsiveness was comparable to that of the SF-20 physical function domain. When sensitivity and specificity to change in performance-based tests (external criteria) was assessed, the COOP PF was unable to discriminate better than chance on any performance test. The SF-20 PF discriminated better than chance for deterioration in balance and gait.

Following four weeks of treatment for congestive heart failure, small to moderate ES statistics were found for the COOP charts ranging from –0.17 to –0.40 (Jenkinson et al., 1997), as shown below in Table b.i). There was little increase in ES when recalculated to include the 43% of patients reporting improvement in their global health. In addition, patients completing global items of change reported improved health status. The authors suggest that such standardized measures of health status alone may not usefully reflect changes in health status of importance to patients.

Table 4b Responsiveness of the COOP charts
(Jenkinson et al., 1997)

COOP Charts	*Effect size*
Physical fitness	0.18
Emotional condition/feelings	–0.11
Daily activities	–0.40
Social activities	–0.14
Bodily pain	0.22
Overall health	0.08
Social support	–0.17
Quality of life	0.00

Following the surgical repair of hip fracture, most WONCA/COOP charts were responsive to change over time and discriminated between change at one week, one month and four months (Van Balen et al., 2003). The most sensitive charts were PF (ES range from 0.30 to 1.15), change in health (ES range from 0.46 to 1.8) and OH (ES range from 0.40 to 0.48). The least sensitive chart was SA; a hypothesised change in social activity between one and four months post-fracture was not detected.

Precision
Four months post-hip fracture, floor effects indicative of no problems were reported for three WONCA/COOP charts (BP: 21%, EC: 36%, SA: 57%) and ceiling effects (maximum score, i.e. severe problems) for two charts (DA: 39%, PF: 60%) (Van Balen et al., 2003).

Acceptability
Completion of the WONCA/COOP charts with and without illustrations produced similar measurement properties (Kempen et al., 1997). Where 75% of respondents found the pictorial illustrations informative, approximately 17% did not; two participants cited the illustrations as a reason for not completing the charts. The authors concluded that there was no need to include illustrations.

Completion rates of between 92% (Jenkinson et al., 1997) and 100% (Doetch et al., 1994) have been reported. More than 90% of residential-home respondents with good to moderate cognitive status completed the COOP pain chart; however, for those stating a preference, it was not the preferred pain measure (Manz et al., 2000).

Average interview completion time for the WONCA/COOP by community-dwelling adults was 49.0 minutes (range 29 to 65 minutes), compared to 39.0 minutes (range 18 to 50 minutes) for the older person-specific EASY-Care (Philp et al., 2001).

c) EuroQol (The EuroQol Group, 1990; revised 1993)

The European Quality of Life instrument (EuroQol) was developed by researchers in five European countries to provide an instrument with a core set of generic health status items (The EuroQol Group, 1990; Brazier et al., 1993). Although providing a limited and standardized reflection of HRQL, it was intended that use of the EuroQol would be supplemented by disease-specific instruments. The developers recommend the EuroQol for use in evaluative studies and policy research; given that health states incorporate preferences, it can also be used for economic evaluation. It can be self- or interview-administered.

Existing instruments, including the Nottingham Health Profile, Quality of Well-Being Scale, Rosser Index, and Sickness Impact Profile were reviewed to inform item content (The EuroQol Group, 1990). There are two sections to the EuroQol: the EQ-5D and the EQ thermometer. The EQ-5D assesses health across five domains: anxiety/depression (AD), mobility (M), pain/discomfort (PD), self-care (SC), and usual activities (UA), as shown in Table 4.1. Each domain has one item and a three-point categorical response scale; health 'today' is assessed. Weights based upon societal valuations of health states are used to calculate an index score of –0.59 to 1.00, where –0.59 is a state worse than death and 1.00 is maximum well-being. A score profile can be reported. The EQ thermometer is a single 20 cm vertical analogue scale with a range of 0 to 100, where 0 is the worst and 100 the best imaginable health.

Ten articles describe the evaluation of the EuroQol, as shown in Table 3.2. With the exception of one hospital-based evaluation in Canada (Tamim et al., 2002), all studies describe European populations across a range of community and hospital settings. The findings below are derived from these articles.

Reliability

The results of test-retest reliability are shown in Table 4.2. Moderate reliability was reported for older female respondents reporting no change in health over six months (EQ-5D index 0.53; EQ thermometer 0.67) (Brazier et al., 1996). Internal reliability testing is not appropriate for the EuroQol.

Validity

(i) Socio-demographic variables and health-service use

Both the EQ-5D index and the EQ thermometer discriminated between groups of women defined by the affirmation or negation of the following variables: 1) GP visits in the previous two weeks, 2) hospital inpatient stays in the previous 12 months, 3) long-standing illness, and 4) high or low disability severity category (range 0-10) (Disability Survey, Office of Population Census Survey [OPCS]) (Brazier et al., 1996). The EQ-5D index also discriminated between respondents who had or had not attended hospital for an outpatient appointment in the previous three months, but neither instrument discriminated between respondents defined by age-group or attendance at Accident and Emergency departments in the previous three months (Table 4.3).

In patients hospitalised due mainly to orthopaedic conditions, there was a statistically significant baseline correlation between the EuroQol and age (thermometer only) and limiting long-standing illness (LLI) (EQ-5D M and UA only) (Coast et al., 1998), as shown in Table 4.3. At four weeks, all EQ-5D items, except UA and AD, and the

thermometer had a statistically significant correlation with age. The EQ thermometer was the only score to have a significant correlation with LLI at four weeks.

(ii) Construct validity: other instruments
In patients with mainly orthopaedic problems, correlation between the EuroQol and WONCA/COOP charts that had hypothesised associations ranged from –0.53 (EQ-5D index with WONCA/COOP overall health) to 0.74 (EQ-5D PD with WONCA/COOP bodily pain) (Coast et al., 1998). Several correlations were smaller than hypothesised: for example, the correlation between EQ-Mobility and WONCA/COOP Physical Fitness was 0.39.

Correlations between domains that did not have hypothesised associations ranged from 0.13 (EQ-5D PD with WONCA/COOP social activities) to 0.42 (EQ-5D SC with WONCA/COOP overall health). Correlation between the EQ-5D index and WONCA/COOP charts ranged from –0.35 (change in health) to 0.59 (daily activities); and with the EQ-thermometer ranged from –0.29 (Physical Function, Social Activities) to –0.65 (Overall Health).

Statistically significant correlations between the EuroQol and Barthel Index (BI) domains that had hypothesised associations were found between the EQ-5D index M and SC items and BI-Mobility, and the EQ-5D SC item with BI-Dressing and BI index (all p less than 0.05; correlation not reported) (Coast et al., 1998). The EQ thermometer had a statistically significant correlation with the BI index score.

(iii) Validity: other
The EQ-5D was completed at four, 12, 17, and 24 months following the surgical repair of hip fracture (Tidermark et al., 2002a,b). The EQ-5D discriminated between groups defined by their pain level (greater or less than 30mm on a 100mm visual analogue scale [VAS]), state of fracture-healing (healing or complications), mobility (no aid or one stick versus walker or wheelchair), and self-care (low or high activities of daily living [ADL] index) at four and 17 months (Tidermark et al., 2002a), and by severity of initial fracture (primary displaced [lower scores] versus undisplaced femoral neck fractures) at 12 and 24 months (Tidermark et al., 2002b). Where patients with healed undisplaced fractures reported a return to their pre-injury level of health, a statistically significant difference in scores for patients with healed displaced fractures was reported (lower HRQL). An average of 38 months following hip replacement for acetabular fracture, a small group of older people reported lower index scores (0.62; standard deviation [SD] 0.27) than an age-matched reference population (0.78; SD not reported) (Tidermark et al., 2003b).

Following completion of the EuroQol by patients due to be discharged from an Emergency Department and their primary caregivers (proxy), agreement between scores and change in score over time was assessed (at baseline, one month, and four months) (Tamim et al., 2002). Agreement between the EQ-5D and observable items, for example, mobility, ranged from 0.44 to 0.60. Agreement between the EQ-5D and more subjective items ranged from 0.10 to 0.50; this improved over time.

Responsiveness
It was hypothesised that patients admitted for elective surgery (knee arthroplasty) would be relatively fitter that those admitted for trauma-related surgery (fractured femoral neck) (Coast et al., 1998). In addition, orthopaedic patients were expected to recover more rapidly than those admitted following a stroke. Despite considerable patient variability, results followed the expected pattern and were more pronounced for the EQ-5D than for the EQ thermometer, i.e. there was greater and more rapid health improvement following elective knee arthroplasty. Mean change at four weeks was EQ-5D 0.31 (SD 0.50), EQ thermometer 11.9 (SD 30.8); mean change at three months was EQ-5D 0.28 (SD 0.45), EQ-thermometer 15.10 (SD 38.1). The smallest mean change was for stroke survivors: EQ-5D at four weeks was –0.046 (SD 0.42), EQ-thermometer 2.1 (SD 21.7); mean change at three months was EQ-5D –0.005 (SD 0.49), EQ-thermometer –0.091 (SD 19.5).

Strong levels of responsiveness were reported for the EQ-5D following the surgical repair of hip fractures (at four months post-surgery ES 1.37, Standardised Response Mean [SRM] 0.90) (Tidermark et al., 2003a). The EQ-5D discriminated between groups defined by the external criterion 'good versus less good clinical outcome'. Change score correlation between the EQ-5D and SF-36 ranged from 0.03 to 0.45, the strongest being with the SF-36 domains of bodily pain, vitality, and physical function.

Brazier et al. (1996) evaluated the ability of the EuroQol to discriminate between hypothetical health states, and improvement in health state when assessed against change in health-service use, change in long-standing illness, and by age-group. A hypothetical change from having to not having a long-standing illness was associated with large effect sizes (ES): EQ-5D 0.85, EQ thermometer 0.71. Other hypothetical improvements in health status, namely hospitalisation in the last 12 months, GP consultation in the previous two weeks, outpatient attendance in the previous three months, and moving from the over 85-year to the 75 to 79-year age-group, were associated with low to moderate ES in the range 0.07 (EQ-5D for age-group change) to 0.42 (EQ-5D hospital inpatient stay).

Participants in a randomized controlled trial of cardiac rehabilitation following an acute cardiac event completed the EuroQol at baseline, three months, one year, and an average of 4.4 years after randomisation (Hage et al., 2003). There was no statistically significant difference between groups or within groups over time.

Precision
Although end effects have not been reported for the EQ-5D index when applied in an older population (Brazier et al., 1996), a large proportion of older participants in a clinical trial of anti-hypertensive drugs indicated 'no problem' with four out of the five domain items (range: 68.2% for the mobility item to 91.7% for the self-care item), the exception being Pain/Discomfort (Degl'Innocenti et al., 2002). 61% of these respondents reported some or extreme problems with at least one item. Index score distribution was not reported.

Acceptability
EuroQol completion difficulties are associated with both increased age and reduced cognitive function (Coast et al., 1998). Approximately 50% of a population of older acute care patients required interviewer administration; the expected probability of an

acute care patient requiring interview administration was calculated to be 11% at 65 years of age, 37% at 75 years, and 73% at 85 years.

EuroQol completion levels by domain range from 84%-93.5% (Brazier et al., 1996) to 89%-100% (Tidermark et al., 2002a,b; 2003a). Less than 10% missing data has been reported for the EQ-5D (Brazier et al., 1996).

d) Functional Status Questionnaire (Jette, 1986)

The Functional Status Questionnaire (FSQ) is a brief multidimensional instrument for the assessment of physical, role, and social function, and psychological health status in children and adults (Jette, 1986; Yarnold et al., 1995). The developers recommend application in primary care screening for disability and for monitoring change in function.

Instrument content was derived from existing instruments (McDowell and Newell, 1996). The six core domains are activities of daily living (ADL: three items), instrumental activities of daily living (IADL: six items), psychological function (PsychF: five items), work performance (six items), social function (SF: three items), and quality of social interaction (QSI: five items), as shown in Table 4.1. Six additional items include general feelings about health, interpersonal relationships, and days sick. Items use statements, which generally refer to a departure from normal performance. Four-, five-, or six-point response scales record a respondent's level of agreement with each statement with reference to the previous month.

Items sum to give six domain scores ranging from 0 to 100, where 100 is the best function, and six single item scores. Areas of clinical concern are highlighted following the application of a scoring algorithm (Jette, 1986; 1987). The instrument may be self-, interview-, or telephone-administered.

Six articles describe the evaluation of the FSQ, as shown in Table 3.2. All studies describe populations across a range of hospital and community settings in the USA. The results given below are derived from these articles.

Reliability

Internal consistency reliability results are shown in Table 4.2. Alpha levels range from 0.42 (QSI) to 0.91 (IADL and SA) (Yarnold et al., 1995; Cleary and Jette, 2000). There is no evidence for test-retest reliability.

Validity

(i) Socio-demographic variables and health-service use

Individual FSQ items did not discriminate between geriatric and non-geriatric (age-defined) ambulatory respondents (Yarnold et al., 1995), as shown in Table 4.3. Although statistically significant, between-group score differences for the IADL and PsychF domains were considered clinically non-significant. The moderate within-group score variability was hypothesised. Yarnold et al. (1995) recommends further work to enhance the precision and breadth of the IADL scale.

Scores on the ADL, IADL, and SA domains discriminated between groups of women defined by their perceived and actual level of difficulty walking ('yes' or 'no' on both accounts; statistical significance not reported) (Brach et al., 2002).

Following a screening programme for functional change (average of 51 months duration), very low levels of correlation between FSQ domains and mortality rates were found (Reuben et al., 1992, cited by Cleary and Jette, 2000), as shown in Table 4d.i and Table 4.3.

Table 4d.i Correlation between FSQ domain scores and mortaility rates (Reuben et al., 1992, cited by Cleary and Jette, 2000).

FSQ domains	***Correlation with mortality rates***
ADL	–0.11
IADL	–0.22
Psychological function	–0.06
Work performance	-
Social function	–0.25
Quality of social interaction	0.08

(ii) Construct validity: other instruments
The correlation between FSQ IADL and PsychF domains and several symptom-specific and generic instruments of health status were assessed in two patient populations, namely, patients receiving cardiac catheterisation following myocardial infarction (MI) and patients hospitalised following acute myocardial infarction (Cleary and Jette, 2000). Findings are shown below in Table d.ii and in Table 4.3.

Table 4d.ii Correlation between FSQ summary scales and other instruments (Cleary and Jette, 2000)

	Cardiac catheterisation		***Hospitalised MI patients***	
Instrument	***FSQ-IADL***	***FSQ-PsychF***	***FSQ-IADL***	***FSQ-PsychF***
Symptom-specific				
Dyspnoea	0.59	0.40	0.58	0.35
Angina	0.48	0.42	0.39	0.32
Schedule of Specific Activities	0.72	0.37	0.73	0.33
Perceived health status	0.49	0.44	0.51	0.37
Global rating	0.60	0.51	0.54	0.41
Generic				
SF-12 Physical Component Scale	-	-	0.75	0.32
SF-12 Mental Component Scale	-	-	0.21	0.82

Correlations between the FSQ and symptom-specific measures ranged from 0.33 (PsychF with Schedule of Specific Activities [SSA]) to 0.73 (IADL with SSA). Correlations between the FSQ and SF-12 domains that had hypothesised associations ranged from 0.75 (IADL with physical component scale) to 0.82 (PsychF with mental component scale).

Correlation between the FSQ ADL and IADL domains and the Older Americans Resources and Services (OARS) Multidimensional Functional Assessment Questionnaire (OMFAQ) and the Physical Performance Test (PPT) ranged from 0.45 (IADL with PPT) to 0.70 (ADL with OMFAQ-IADL) (Reuben et al., 1995), as shown in Table 4.3. Correlations between the FSQ and SF-36 domain scores ranged from 0.33 (ADL with role: emotional) to 0.76 (IADL with physical function).

Correlations between the FSQ ADL domain and a range of measures of functional ability were 0.57-0.67 for the PPT, 0.58-0.78 for SF-36 physical function and 0.68-0.73 for the National Institute on Aging Battery (Sherman and Reuben, 1998).

(iii) Validity: other
Correlation between the FSQ ADL and IADL domains were reported as 0.73 (Reuben et al., 1995) and 0.85 (Sherman and Reuben, 1998).

Responsiveness
Patients undergoing balloon valvuloplasty (group 1) and those who declined, and those for whom the procedure was deemed unsuitable (group 2), completed both the FSQ and New York Heart Association (NYHA) classification system. Baseline scores did not differ between groups for either instrument. Scores at one month were significantly higher than baseline for group 1, with a statistically significant between-group difference for both instruments (Tedesco et al., 1990). Further score comparison at three months, with a physician's interpretation of patient ability, supported the authors' recommendation of the FSQ as a suitable instrument for identifying cardiac patients with residual functional limitation.

Precision
Completion by community-based respondents demonstrated ceiling effects of 16% for IADL, 41% for ADL (Sherman and Reuben, 1998), and 60% for ADL (Reuben et al., 1995). Completion by high functioning women showed ceiling effects of 61% for IADL, 77% for ADL, and 94% for SA (Brach et al., 2002). Negligible floor effects have been reported (Sherman and Reuben, 1998).

Acceptability
Older and younger people are equally unlikely to generate missing data; the maximum percentage of missing responses for a specific item was 17% and 11%, respectively (Yarnold et al., 1995).

Among 83 respondents, misunderstanding was apparent in only one person (FSQ ADL and IADL), which suggests that the FSQ has a high level of acceptance (Reuben et al., 1995). Self-completion in the general population takes approximately 15 minutes (Yarnold et al., 1991).

e) Goteborg Quality of Life (Tibblin et al., 1990)

The Goteborg Quality of Life (GQL) instrument was developed in Sweden during the 1970s for use in population-based evaluations of general health (Tibblin et al., 1990). The WHO statement of health and literature searches informed item content. The instrument is self-administered in two parts. Part I, the GQL-instrument provides a subjective assessment of well-being across three domains: social (four items), physical (six items) and mental well-being (five items) (Tibblin et al., 1990), as shown in Table 4.1. There are inconsistencies in the number of items reported; Nygren et al. (2001) reported 18 items. Items have ordinal response scale ranging from 'very bad' (one point) to 'excellent, could not be better' (seven points). Although the developers considered all 'well-being' items separately, subsequent authors have summed items to give an index score between 7 and 105, where 105 is the best health (Nygren et al., 2001).

Part II is a Symptom Profile that uses yes/no responses to 30 symptoms, for example, dizziness, abdominal pain, and breathlessness. Subsequent authors have summed items to produce an index score between 0 and 30, where 30 indicates the presence of all symptoms (Andersson et al., 1995).

There have been two evaluations of the GQL. Both studies include community-based older populations in Sweden (Andersson et al., 1995; Nygren et al., 2001), as shown in Table 3.2. The results given below are derived from these articles.

Reliability

A high level of internal consistency reliability was reported for the GQL well-being instrument (Nygren et al., 2001), and similarly for the GQL symptom profile (Andersson et al., 1995), as shown in Table 4.2. There is no evidence of test-retest reliability.

Validity

Completion by people with hearing difficulties demonstrated a moderate to strong correlation between the GQL and the Hearing Coping Assessment (0.34), the Life Orientation Test (–0.50), and the Beck Depression Inventory (0.61) (Andersson et al., 1995). A small correlation was found between the GQL and the Hearing Questions Scale (–0.13), and the assessment of pure tone (Pure Tone Average –0.15).

f) Health Status Questionnaire 12 (Radosevich and Pruitt, 1995; Health Outcome Institute, 1996)

The Health Status Questionnaire 12 (HSQ-12) was developed by the Health Outcome Trust as a generic instrument for the multidimensional evaluation of physical, emotional, and social functioning (Bowling and Windsor, 1997).

Instrument content was derived from the 39-item Health Status Questionnaire (HSQ, version 2), an extension of the SF-36 (Radosevich and Pruitt, 1995; Health Outcome Institute, 1996 - both cited by Bowling and Windsor, 1997). The twelve items of the HSQ-12 assess the same eight domains of health status included in the SF-36, namely bodily pain (BP), energy/fatigue (E), mental health (MH: three items), physical functioning (PF: three items), perceived health (PH), role limitation-mental (RM), role limitation-physical (RP), and social functioning (SF), as shown in Table 4.2. The HSQ-12 explains at least 90% of the variance in the SF-36 (Radosevich and Pruitt, 1995, cited by Bowling and Windsor, 1997). Although the HSQ-12 and SF-12 include different items, the SF-12 also explains 90% of the variance of the SF-36 (Ware et al., 1995).

The HSQ-12 assesses the impact of health on functioning over the previous four weeks. Categorical response options range from three to six options. An average score is calculated for the two multi-item scales (Physical Health and Mental Health); the recoded response is the score for single items. Scores range from 0 to 100, where 100 is the best possible health. The instrument has been interview-administered to older people.

There have been two evaluations of the HSQ-12. Both studies refer to community-based older populations in the UK (Bowling and Windsor, 1997; Petit et al., 2001), as shown in Table 3.2. The results given below are derived from these articles.

Validity

(i) Socio-demographic variables and health-service use

When responses were dichotomised into ‘no limitations or problems’ versus ‘limitations or problems’, the PH, PF, BP, and E domains showed a strong and statistically significant association with age (Bowling and Windsor, 1997). MH and SF domains had a less consistent and non-statistically significant association with age. With the exception of the RM and SF domains, multiple regression analysis identified age as a significant factor influencing all HSQ-12 domains.

Completion by a large community-based population showed that all domains discriminated between groups who did or did not report long-standing illness, disability, or infirmity (Bowling and Windsor, 1997), as shown in Table 4.3. In a subsequent study, the HSQ-12 discriminated between groups with and without self-reported health problems, and between groups defined by the presence of depression or ADL limitation (Pettit et al., 2001). Furthermore, with the exception of RM, all domains discriminated between groups defined by the receipt of health services and impaired vision. With the exception of MH and BP, all domains discriminated between groups with or without dementia. PH, MH, and RM discriminated between groups defined by psychiatric problems, and PF, RP, SF, and E between groups defined by hearing impairment. The HSQ-12 did not discriminate between groups defined by the presence of organic brain syndrome.

(ii) Validity: other
Correlations between HSQ-12 domains that had hypothesised associations ranged between 0.45 (BP with RP) and 0.72 (PF with RP) (Pettit et al., 2001), as shown in Table 4.3. Correlations between domains that did not have hypothesised associations ranged from 0.19 (PF with RM) to 0.28 (E with RM).

Following completion of the HSQ-12 and SHORT-CARE at baseline and 18 months, regression analysis of change scores indicated that change in SHORT-CARE Activities of Daily Life (ADL) domain was predicted by the baseline ADL score and by change in HSQ-12 PF, RP, and SF scores, explaining 56% of the variance in change score (Petit et al., 2001). Change in SHORT-CARE depression was predicted by the baseline depression score, and change in HSQ-12 MH and RM scores, explaining 41% of the variance in change score.

Precision
High mean values with wide score variation, and hence concern over ceiling effects, were reported for the BP, MH, PF, RP, RM, and SF domains (Bowling and Windsor, 1997). High mean values for RM and SF domains were also reported in a separate community-based study: in those over 65 years of age, 88.9 (SD 24.0) for RM, 77.1 (SD 35.3) for SF; in those over 75 years of age, 88.6 (SD 24.0) for RM, 71.5 (SD 38.8) for SF (Petit et al., 2001).

Acceptability
Following interview administration of a package of instruments including the HSQ-12, SF-12, and the SHORT-CARE, 94.4% correctly completed the HSQ-12 (Pettit et al., 2001). Respondents with self-reported depression and those with dementia had completion rates of 91.5% and 78.8%, respectively. Community-dwelling independent people with less severe dementia had higher completion rates (91.1%) than institutionalised individuals with severe depression (56%). The majority of instrument completers (97.5%) found it to be acceptable. Three independent predictors for non-completion of the HSQ-12 were the SHORT-CARE dementia score, first language, and ethnicity.

g) Index of Health-related Quality of Life (Rosser et al., 1992)

The Index of Health-related Quality of Life (IHQL), also known as the Health-related Quality of Life Questionnaire, was developed by Rosser et al. (1992; 1993). Instrument content was derived from the Rosser Index (Rosser and Watts, 1978). Pain and emotional distress domains were added to provide a multidimensional and hierarchical assessment of health-related quality of life (HRQL) for application across different disease states (Livingstone et al., 1998).

The IHQL has three core domains with seven sub-domains, namely disability: dependency and dysfunction, discomfort: pain/discomfort and symptoms, and distress: dysphoria, disharmony, and fulfilment, as shown in Table 4.1. The sub-domains comprise 44 items, which in turn include 107 descriptors and a further 225 descriptor levels. Most items have a five-option response scale. A scoring algorithm is used to generate an index score between 0 and 1, where 0 is equivalent to death and 1 to 'no impairment'. Alternatively, a five-level multidimensional classification across the three domains may be produced (Bowling, 1995; Rosser et al., 1993). The instrument is interview-administered.

There has been one evaluation of the IHQL. This included a community-based older population in the UK (Livingston et al., 1998), as shown in Table 3.2.

Validity

Construct validity: other instruments

A package of instruments including the IHQL, the SHORT-CARE, and an Anxiety Disorder Scale were completed. Correlation between the IHQL and instruments with hypothesised associations ranged from 0.08 (IHQL discomfort with SHORT-CARE depression) to 0.14 (IHQL disability with SHORT-CARE somatic symptoms). The authors concluded that the lack of evidence in support of the instrument's convergent validity restricted its usefulness for informing decision-making in older people.

Precision

Following completion by a largely independent population, with only a minority experiencing depression or dementia, similar scores with a narrow range of values and uniform distribution were reported for each IHQL domain.

Acceptability

75% of participants completed an interview-administered IHQL. Although completion rates may have been adversely affected by positioning at the end of a long interview, several respondents indicated that they did not find the IHQL applicable to their life.

h) Nottingham Health Profile (Hunt et al., 1980)

The Nottingham Health Profile (NHP) was developed in the UK during the 1970s for use in the evaluation of medical or social interventions (Hunt et al., 1980). Instrument content was derived from over 2000 statements given by 768 patients with a variety of chronic ailments and other lay people.

Part I of the instrument has 38 items across six domains: bodily pain (BP), emotional reactions (ER), energy (E), physical mobility (PM), sleep (S), and social isolation (SI), as shown in Table 4.1. All items are statements that refer to departures from normal functioning, and relate to feelings and emotional state rather than change in behaviour. Respondents answer 'yes' or 'no' according to whether or not they feel the item applies to them in general. Positive responses are weighted and summed to give six domain scores between 0 and 100, where 100 denotes maximum limitation.

Part II of the NHP is less widely used and provides a brief indicator of handicap. The instrument may be self-, interview-, or telephone-administered.

There have been eight evaluations of the NHP. Three studies are of community-based older populations in the UK (Hunt et al., 1980; Sharples et al., 2000; Mitchell et al., 2001), as shown in Table 3.2. The remaining studies are of patients from Europe (Thorsen et al., 1995; Van Balen et al., 2001, 2003), Canada (Stadnyk et al., 2000), and Australia (Crockett et al., 1996) in a variety of community- and hospital-based settings (see Table 3.2). The results given below are derived from these articles.

Reliability

The results with regard to test-retest reliability and internal reliability are shown in Table 4.2. Agreement between the six NHP domains was assessed using Cronbach's alpha (0.82) (Sharples et al., 2000). With the exception of the SI domain (0.52), at four months post-hip fracture, high levels of internal consistency reliability (greater than 0.70) were reported for five out of the six NHP domains (Van Balen et al., 2003).

High levels of one-month test-retest reliability were found for all NHP domains, ranging from 0.81 (SI) to 0.97 (PM) (Sharples et al., 2000).

Although detail is limited, item-total correlation for each NHP domain ranged from 0.61 (SI) to 0.85 (other domains not specified) (Sharples et al., 2000).

Validity

(i) Socio-demographic variables and health-service use

All NHP domains discriminated between groups defined by their fitness levels, well-being or general practitioner consultations (Hunt et al., 1980), as shown in Table 4.3. As hypothesised, respondents who were fit and without long-standing illness had lower mean scores across all domains (score less than 10) than those with chronic illness or disability. Although not discriminating between groups defined by social class, age (under or over 70 years), or living status (living alone or not), several domains discriminated between groups defined by marital status and sex. High-scoring divorcees or widowers were more likely than their married counterparts, and women more likely than men, to score highly on SI. For those from the low-scoring group, women were less likely than men to affirm scores on the sleep domain.

Patients with chronic obstructive airways disease had lower scores than the general population on the ER, E, and SI domains (Crockett et al., 1996). As hypothesised, patients with osteoarthritis of the hip awaiting hip replacement surgery had higher scores (greater distress) across all domains than outpatients with back pain (Thorsen et al., 1995). However, both groups scored more highly across all domains than participants in a fitness class ('fit elderly'). BP and PM domains discriminated between members of the fit elderly group defined by self-reported musculoskeletal problems. The E, S, and SI domains discriminated between older men and women (higher scores) attending a back pain clinic, and E and PM domain scores discriminated by age for patients with osteoarthritis of the hip; the older age-group reported worse health.

(ii) Construct validity: other instruments
Correlation between NHP domains and a battery of lower extremity performance tasks, the Guralnik Performance Test (Guralnik et al., 1994) and activities of daily life (ADL) items adapted from the Katz ADL scale (Katz et al., 1963) that had hypothesised associations ranged from 0.51 (BP with ADL) to 0.74 (PM with ADL) (Sharples et al., 2000), as shown in Table 4.3 and in Table 4h.i below.

Table 4h Correlation between the NHP and other instruments (Sharples et al., 2000)

NHP	***Guralnik Performance Test (Guralnik et al., 1994)***	***Katz ADL scale (mean score) (Katz et al., 1963)***
Physical mobility (PM)	0.70	0.74
Energy (E)	0.53	0.58
Bodily Pain (BP)	0.54	0.51

Following completion by patients with lung disease, correlations between the NHP and SF-36 ranged from 0.00 (BP with SF-36 general health; sleep with SF-36 role-physical) to –0.88 (E with SF-36 vitality) (Crockett et al., 1996). In patients undergoing rehabilitation, correlations between NHP and SF-36 domain scores that had hypothesised associations ranged from –0.25 (SI with SF-36 social functioning) to –0.76 (PM with SF-36 physical functioning) (Stadnyk et al., 1998) (Table 4.3). Correlations between domains that did not have hypothesised associations ranged from 0.02 (ER with SF-36 role-physical) to –0.41 (E and ER with SF-36 general health).

At four months post-hip fracture, correlation between the NHP and WONCA/COOP charts that had hypothesised associations ranged from 0.50 (E with WONCA/COOP overall health) to 0.75 (PM with WONCA/COOP daily activities) (Van Balen et al., 2003), as shown in Table 4.3. Correlation between domains and charts that did not have hypothesised associations ranged from 0.09 (Sleep with WONCA/COOP overall health) to 0.38 (PM with WONCA/COOP emotional condition). Correlations between the NHP and the Rehabilitation Activities Profile (RAP) ranged from 0.01 (S with RAP relationships) to 0.87 (PM with RAP Mobility and Personal Care), and with the Barthel Index ranged from 0.04 (S) to 0.79 (PM), in accordance with study hypotheses (Van Balen et al., 2003).

(iii) Validity: other

The SI and ER domains discriminated between groups with and without anxiety or depression, and between people with possible and probable morbidity (Sharples et al., 2000). The relationship between NHP scores and self-reported morbidity and symptoms recorded during an interview was assessed. Expected relationships were demonstrated. For example, those with self-reported nervous or emotional problems scored more highly on the ER domain than those without self-reported problems, and those reporting hypertension. In addition, those reporting breathlessness or tiredness scored more highly on the E domain than those not reporting these symptoms, and those reporting heart or chest problems, or hypertension.

At four months post-hip fracture the ER, PM, S, and SI domain scores for survivors were less than that of a reference population matched for both age and sex (Van Balen et al., 2001).

Correlation between NHP domains ER and SI was 0.65 (Van Balen et al., 2003).

Responsiveness

Following the rehabilitation of frail older people with mostly medical conditions, small to moderate levels of responsiveness were found between hospital admission and discharge (duration not reported): effect size (ES) ranged from 0.00 for BP and ER to –0.40 for ER (Stadnyk et al., 1998). Mean domain scores generally improved for patients receiving specific quadriceps training versus those receiving standard physiotherapy, but between-group difference was only statistically significant for the E domain (Mitchell et al., 2001).

Following hip fracture repair, all NHP domains were responsive to change and discriminated between change at one week, one month, and four months (Van Balen et al., 2001,2003). The most responsive domains were BP (ES 0.35-0.95) and PM (ES 0.57-1.48); the least responsive was ER (ES 0.02-0.11) (Van Balen et al., 2003).

Precision

A skewed response distribution towards fewer affirmations was found following completion by community-based patients in the UK. Results were: actual affirmation range 0-32 compared with a possible 0-38 (median affirmations: 4), no affirmation 19%, and affirmation of one or two statements 21% (Hunt et al., 1980). For the group with reported better health (low-scoring group) the affirmation range was 0-14 (no affirmation: 35.6%), compared to 0-32 (no affirmation: 9.7%) for the high-scoring group. A similar result was found following completion by Dutch patients: fitter people had fewer affirmative answers when compared to people with osteoarthritis of the hip (Thorsen et al., 1995).

Following completion by a community-based population, ceiling effects (minimum score, i.e. no limitation) all NHP domains had ceiling effects ranging from 38% (PM) to 68% (SI) (Sharples et al., 2000). Floor effects were not reported in this population. At four months post-hip fracture, four domains had ceiling effects, namely ER 27%, E 34%, SI 36%, and S 44% (Van Balen et al., 2003). The Energy domain had floor effects (27%).

Acceptability

Interview administration of the NHP with older people took approximately 10-15 minutes (Van Balen et al., 2003). Interview participation rates ranged from 73% (n=511) (Sharples et al., 2000) to 84% (Stadnyk et al., 1998). Refusal rates increased with age (Sharples et al., 2000). NHP item completion was high following interview administration (over 99.6% Sharples et al., 2000; 100% Crocket et al., 1996 and Van Balen et al., 2003).

When defined by fitness levels or well-being, non-completion with self-administration (including non-response) ranged from 4.7% (five out of 64 clinic patients with hip osteoarthritis) to 7.4% (five out of 68 clinic patients with back pain) (Thorsen et al., 1995).

A high level of reading ease (according to the Flesch formula; Todd and Bradley, 1994) was reported for the NHP (part I), indicating that 88% of individuals would understand the instrument (Sharples et al., 2000). Respondent burden may be reduced by the use of dichotomous response options (Sharples et al., 2000). Alternatively, the lack of response discrimination may reduce instrument responsiveness.

i) Quality of Life Index (Ferrans and Powers, 1985; Ferrans and Ferrell, 1990)

The Quality of Life Index (QLI) was developed in the USA during the 1980s as a measure of morbidity for application in both normal and unwell populations (Ferrans and Powers, 1985; Bowling, 1995).

Instrument content was informed by literature reviews, which considered quality of life across all age-groups and different illnesses (Kleinpell and Ferrans, 2002). Quality of life was defined as a multidimensional construct with four key domains: family, health and function, psychological and spiritual, and social and economic. The instrument comprises two sections assessing respondent satisfaction and relative importance of each domain, respectively. Each section has 32 items, with eight items per domain. Six-point ordinal response scales range from 'very dissatisfied' or 'very unimportant' (1), to 'very satisfied' or 'very important' (6). Scoring is complicated and the developers recommend a computer programme. In summary, importance scores are used to weight satisfaction scores. Index or domain scores range from 0 to 30, where higher scores indicate better quality of life (Bowling, 1995, p54). The instrument has been self-completed by an older population.

The original instrument was developed and tested in patients receiving haemodyalysis, and several dialysis-specific items are available (Bowling, 1995). Factor analysis confirmed instrument construction. The QLI has been modified for use with cancer patients (Bowling, 1995).

There has been one evaluation of the QLI in an older population. This was a follow-up study of people discharged home from intensive care units in the USA (Kleinpell and Ferrans, 2002), as shown in Table 3.2.

Reliability

The index (0.96) and separate domains (range 0.79 for family to 0.94 for health and functioning) had a high level of internal consistency reliability, as shown in Table 4.2. There is no evidence for test-retest reliability.

Validity

Socio-demographic variables and health-service use

There was no statistically significant difference in QLI scores between middle-aged, young-old, or old-old people following recovery from a period of intensive care. Greater perceived health (and future health), greater social support, and hospital readmission explained 51% of the variance in higher QLI scores. A longer period of hospitalisation explained 48% of the variation in lower QLI scores, in accordance with hypotheses.

Acceptability

In a postal survey, self-completed questionnaires were returned by 52% of the population (n=164).

j) Quality of Well-Being Scale (formerly the Index of Well-Being) (Kaplan et al., 1976; Kaplan et al., 1984; Kaplan et al., 1993)

The Index of Well-Being was modified and renamed the Quality of Well-Being scale (QWB) to emphasize the focus on quality of life evaluation (Kaplan et al., 1993; McDowell and Newell, 1996).

The QWB uses a three-component model of health (Kaplan and Anderson, 1988, cited by McDowell and Newell, 1996) comprising: 1) functional assessment, 2) a value reflecting the utility or desirability of each functional level, and 3) an assessment of illness prognosis to anticipate future health-care need, which may describe positive health. The QWB is interview-administered.

Completion corresponds to the three-component model. First, three domains of self-reported function are assessed, namely mobility and confinement (MOB: three categories), physical activity (PAC: three categories), and social activity (SAC: five categories). Respondents select the most appropriate category to describe their perceived functional level. Domain categories give 45 possible combinations (3 x 3 x 5); with the inclusion of death, 46 function levels are defined for the second stage of completion (McDowell and Newell, 1996). In addition, respondents select from a list of 27 items symptoms or medical problems experienced over the previous eight days.

Social preference weights for each possible health state have been derived from empirical studies. At the second stage, the assignment of an appropriate weight, or utility, to a health state or functional level gives the QWB index score from 0 to 1, where 0 equates to death and 1 to complete well-being. A negative score may be generated, representing a state ‘worse than death’. QWB index scores can be converted into Quality-Adjusted Life-Years (QALYs), supporting their application in economic and policy analysis.

Stage three of the QWB addresses issues of prognosis to produce a well-life expectancy score (McDowell and Newell, 1996). This stage is not necessary for calculating the QWB index.

A self-administered version has been developed: the QWB-SA (Andresen et al., 1998b). Following a review of QWB items, five items were added to a mental health section and three self-rated health items were included. The QWB-SA has five domains: symptoms and problem complexes (58 acute and chronic items), self-care (two items), mobility, physical functioning (11 items for these two), and performance of usual activity (three items). For the first domain, respondents indicate the presence or absence (‘yes’ or ‘no’) of chronic (18), acute physical (25), and mental health symptoms (11) over the previous three days. The remaining four domains all use a three-day recall response option. The total number of items is inconsistent, ranging from 71 to 74. Symptom/problem weights for the QWB-SA are based on the original QWB weighting system. The focus of the original QWB is utility measurement and quality of life; the focus of the QWB-SA is symptoms and assessment of function. The QWB-SA has been recommended for self-completion by older adults (Andresen et al., 1998b).

There have been three evaluations of the QWB (Andresen et al., 1995; DeBon et al., 1995; Groessl et al., 2003) and one of the QWB-SA (Andresen et al., 1998b). These

studies include populations from a mixture of community-based older populations in the USA, as shown in Table 3.2. The results given below are derived from these studies.

Validity

(i) Socio-demographic variables and health-service use

Correlation between the QWB and age was –0.14 (Andresen et al., 1995), and between the QWB-SA and age –0.071 (Andresen et al., 1998b).

The QWB-SA discriminated between groups defined by self-reported health status, where worse health was associated with lower scores (Andresen et al., 1998b). Correlation between the QWB-SA and self-reported days spent in bed was –0.25, and with days of restricted activity was –0.34. The QWB-SA score did not discriminate by sex.

(ii) Construct validity: other instruments

Correlations between the QWB and Sickness Impact Profile (SIP) domain scores that had hypothesised associations ranged from –0.37 (SIP psychosocial domain) to –0.52 (SIP index) (Andresen et al., 1995), as shown in Table 4.3. Correlations between the QWB and three SF-36 domains (physical function [PF], role limitation-physical [RP], and general health [GH]) ranged from 0.36 (GH) to 0.39 (PF). Correlations between the QWB and other instruments that did not have hypothesised associations ranged from –0.09 (Chronic Disease Index) to –0.18 (Stress Scale).

The correlation between the QWB and specific functional activities and observed symptoms was assessed in residents of convalescent hospitals and senior centres (DeBon et al., 1995). Small to moderate correlations were found between the QWB and time taken to perform a range of functional activities, where better health was associated with reduced time taken. For example, correlation between QWB scores and time taken to walk 30 feet was –0.27. Correlation between the QWB and grip strength was 0.32. The QWB discriminated between groups defined by their need for assistance with walking and the presence or absence of depressive symptoms. A large correlation between self-reported and observed symptoms was also reported.

Correlations between the QWB-SA, SIP and SF-36 domains with hypothesised associations were 0.42 (SIP index), 0.47 (SF-36 physical component ssummary score) and 0.51 (SF-36 physical function) (Andresen et al., 1998b). Correlations between the QWB-SA, SIP and SF-36 domains that did not have hypothesised associations were 0.17 (SF-36 role-emotional), 0.22 (SF-36 mental component summary scores) and –0.40 (SIP psychosocial summary score). Correlation between the QWB-SA and SIP subscale scores for work and eating were –0.11 and –0.12, respectively.

Precision

When completed by community-dwelling respondents, the QWB had a limited score distribution (range: 0.50-0.90, mean: 0.72 [SD 0.08]) (Andresen et al., 1995). Score distribution for the QWB-SA approached normality, and covered the full range of possible scores (range: 0.25-1.00, mean 0.70 [SD 0.09]), without evidence of end effects (Andresen et al., 1998b).

Acceptability

High rates have been reported for telephone-interview completion of the QWB (186/200 patients, 98%) (Andresen et al., 1995) and self-completion of the QWB-SA (70%)

(Andresen et al., 1998b). Following completion of the QWB-SA, SF-36, and SIP, patient-rated satisfaction was least for the QWB-SA (60% very or somewhat satisfied); higher, and similar, levels of satisfaction were reported for the SF-36 (67%) and the SIP (69%) (Andresen et al., 1998b).

Time taken for QWB telephone-interview administration ranged from six to 30 minutes (mean 17.4 minutes) but, in comparison to the SIP and SF-36 (three domains), QWB administration met with more difficulties (Andresen et al., 1995). The QWB-SA had a mean completion time of 14.2 minutes; this was comparable to the SIP (19.3 minutes) but longer than that required for completing the SF-36 (12.5 minutes) (Andresen et al., 1998b).

Missing data has been reported for all items within the QWB-SA (missing items mean 4.7 [SD 9.3]) (Andresen et al., 1998b). Highest completion rates were found for items requiring a yes/no response (0.3% to 8.6% missing). 50% of respondents omitted at least one item from sections with a three-day recall response (missing items range: 3%-14.6%). The two self-care items were omitted by more than 11.6% of respondents, and between 12.3% and 16.9% omitted responses for mobility and physical function items.

k) SF-12: Medical Outcomes Study 12-item Short Form Health Survey (Ware et al., 1995)

In response to the need to produce a shorter instrument that could be completed more rapidly, the developers of the Medical Outcomes Study (MOS) 36-item Short Form Health Survey (SF-36) produced the 12-item Short Form Health Survey (SF-12) (Ware et al., 1995).

Using regression analysis, 12 items were selected that reproduced 90% of the variance in the overall Physical and Mental Health components of the SF-36 (Table 4.1). The same eight domains as the SF-36 are assessed and categorical response scales are used. A computer-based scoring algorithm is used to calculate scores: Physical Component Summary (PCS) and Mental (MCS) Component Summary scales are generated using norm-based methods. Scores are transformed to have a mean value of 50, standard deviation (SD) 10, where scores above or below 50 are above or below average physical or mental well-being, respectively. Completion by UK city-dwellers reporting the absence of health problems yielded a mean PCS score of 50.0 (SD 7.6) and MCS of 55.5 (SD 6.1) (Pettit et al., 2001). Although not recommended by the developers, Schofield and Mishra (1998) report eight domain scores and two summary scores. The SF-12 may be self-, interview-, or telephone-administered.

Several authors have proposed simplification of the scoring process and revision of the SF-12 summary score structure, where norm-based weighting is replaced by item summation to facilitate score interpretation and appropriateness for older people (Resnick and Nahm, 2001; Resnick and Parker, 2001). Additionally, factor analysis in two different populations of older people supported the inclusion of item 10 (‘Did you have a lot of energy?’) in the PCS rather than the MCS, and item 12 (‘How much of the time have your physical health or emotional problems interfered with your social activities?’) in the calculation of both the PCS and MCS. Evidence of high levels of internal consistency reliability as shown in Table 4.2 further supported these changes.

Having observed the difficulties experienced by older respondents in completing the SF-12, Iglesias et al. (2001) modified the response format for ‘stem and leaf’ items, in that, instead of a general phrase followed by several specific questions, a list of individual questions was provided. Following a pilot evaluation, this revised version of the questionnaire, the York SF-12, was evaluated in a randomised trial of hip protection in older women.

There have been seven evaluations of the SF-12. Two studies include community-based populations from the UK (Iglesias et al., 2001; Pettit et al., 2001), as shown in Table 3.2. The remaining studies are of older populations from various community settings in Switzerland (Theiler et al., 2002), Australia (Schofield and Mishra, 1998; Lim and Fisher, 1999), and the USA (Resnick and Parker, 2001; Resnick and Nahm, 2001) (Table 3.2). The results given below are derived from these studies.

Reliability

The developers have stated that assessment of internal consistency reliability is not appropriate for the SF-12 (Ware et al., 1995). However, this was evaluated for both the standard SF-12 and the York SF-12 following completion by a group of older women (Iglesias et al., 2001), as shown in Table 4.2. For both versions, Cronbach’s alpha exceeded 0.90 for the PCS, with slightly lower values for the MCS (standard SF-12:

0.88; York SF-12: 0.91). High levels of internal consistency reliability were found for the standard score calculation (MCS 0.70, PCS 0.84) and a revised score where item 10, Vitality, was included in the PCS and item 12, Social Time, included in both summary scores (MCS 0.70 and 0.80, PCS 0.89 and 0.87) (Resnick and Nahm, 2001; Resnick and Parker, 2001).

Validity

(i) Socio-demographic variables and health-service use

Both the SF-12 MCS and PCS discriminated between groups defined by the presence or absence of a range of health states, including health- and social care use, self-reported health problems, ADL limitation, depression, and impaired vision (Pettit et al., 2001), as shown in Table 4.3. The SF-12 MCS discriminated between groups with self-reported psychiatric problems and those without; the SF-12 PCS discriminated between those with and without dementia, and those with and without impaired hearing. The SF-12 did not discriminate between those with an organic brain syndrome and those without.

SF-12 and SF-36 profile and summary scores were compared following completion by a large group of Australian women stratified into three age-groups: young (18-22 years), middle-aged (45-49 years), and old (70-74 years) (Schofield and Mishra, 1998). Evidence supported an association between age and all eight domains, with an age-related decrease in PCS and increase in MCS scores (statistical significance not reported).

These women were also defined by self-reported physical health according to the number of symptoms experienced over the previous 12 months (Schofield and Mishra, 1998). In the older group, the SF-12 PCS discriminated between groups with normal or poor physical health. As hypothesised, higher summary scores were associated with fewer self-reported symptoms in older and younger women. In the older age-group, SF-12 role-physical was most discriminative when evaluated against self-reported physical health, followed by V, BP, and GH. All domains and both summary scales discriminated between groups defined by psychological distress (as defined by the General Health Questionnaire-12). In the older age-group, SF-12 MH was the most discriminating (in the distressed group, mean MH was 49.8 versus 78.8 in the non-distressed group), followed by mean scores for SF, BP, and PF.

As hypothesised, independent older people who exercised regularly had higher levels of physical and mental health than those who did not exercise regularly; statistical significance was achieved for physical health only (Resnick and Nahm, 2001; Resnick and Parker, 2001). Correlation between summary scores and the number of chronic illnesses was in accordance with hypotheses: –0.41 for PCS, –0.44 for MCS.

(ii) Construct validity: other instruments

As hypothesised, the SF-12 MCS explains greater variation in the SHORT-CARE depression scales than does the PCS (Gurland et al., 1984), while the PCS explains greater variation in ADL limitation (Pettit et al., 2001).

Correlation between the SF-12 PCS and MCS was 0.08, in accordance with hypotheses (Pettit et al., 2001).

(iii) Validity: other
Factor analysis gave the expected two-factor structure for both the standard SF-12 and the York SF-12 (Iglesias et al., 2001). Confirmatory factor analysis in a separate population supported a revised SF-12 model as detailed above (Resnick and Nahm, 2001).

Responsiveness
Older people who received drug therapy for moderate to severe osteoarthritis of the hip or knee completed both the SF-12 and the Western Ontario MacMaster Osteoarthritis (WOMAC) Questionnaire (Bellamy et al., 1988) at baseline and following three weeks of treatment (Theiler et al., 2002). Correlation between change scores for the SF-12 PCS and WOMAC domains of functional ability, pain, and stiffness were –0.64, –0.54, and –0.46, respectively. SF-12 PCS score improvement was statistically significant (mean change: 5.21) while MCS score change was not (mean change: 1.27).

Acceptability
Low survey response rates for the York SF-12 (29.5%) were justified by the study aim and target patient population, namely to recruit women at high risk of hip fracture to a randomized trial of hip protectors (Iglesias et al., 2001). Completion rates for the SF-12 and York SF-12 did not differ significantly; however, the York SF-12 had a statistically significantly lower item non-response rate (8.5% compared to 26.6%). A low participation rate for a pilot evaluation of the SF-12 and SF-36 was reported (18%) (Schofield and Mishra, 1998).

High completion rates have been reported for the SF-12, ranging from 78.2% (Lim and Fisher, 1999) to 94.5% (Pettit et al., 2001). When assessed by clinical diagnosis, lower completion rates were found for respondents with depression (91.4%) and dementia (65.3%) (Pettit et al., 2001). Not surprisingly, respondents with less severe dementia had higher completion rates (72.4%) than those requiring institutional care (55%). Two independent predictors for instrument completion have been proposed, namely the SHORT-CARE dementia score and first language. Lower completion rates were also found for females, older age-groups, and those with poorer health (Lim and Fisher, 1999).

Omission of one, two, and more than five items have been reported by 50%, 24%, and 13% of non-completers, respectively, with items spread across both mental and physical domains (Lim and Fisher, 1999). Item 7 (emotional impact on work and other activities) was most frequently omitted (10.7%) and item 1 (general health item) least frequently. For respondents who expressed an opinion, 88.1% found it ‘quite acceptable’ and 12.5% ‘very acceptable’; 2.7% found it ‘not at all acceptable’ (Pettit et al., 2001).

Feasibility

SF-12 scores varied depending on whether it was administered independently or embedded in the SF-36 (Schofield and Mishra, 1998). Across three age-groups, there were statistically significant differences in domain scores for the independently administered SF-12 and the SF-36, but no statistically significant differences in MCS and PCS. In addition, following completion in an older population-only sample, a statistically significant difference in SF-12 domain scores was found when it was completed independently compared to when it was embedded in the SF-36. There were no statistically significant differences in the SF-12 summary scores. Where a detailed domain profile is required, the authors recommend the SF-36.

l) SF-20: Medical Outcomes Study 20-item Short Form Health Survey (Stewart et al., 1988; Ware, Sherbourne and Davies, 1992)

The Medical Outcomes Study (MOS) 20-item Short Form Health Survey (SF-20) is a 20-item abbreviation of the same Rand instrument from which the SF-36 originates (Stewart et al., 1988; Ware et al., 1992; McDowell and Newell, 1996). The abridged instrument was intended to reduce respondent burden, whilst comprehensively addressing important issues in health status measurement.

The SF-20 assesses health across six domains, namely bodily pain (BP: one item), general health perception (GH: five items), physical function (PF: six items), mental health (MH: five items), social function (SF: one item), and role function (RF: two items), as shown in Table 4.1. Items have categorical response options (range: 3-6 options); several items have reversed scoring. Domain item summation scores are transformed into a scale from 0 to 100, where higher values denote better health. The instrument may be self-, interview-, or telephone-administered. Instrument self-administration takes approximately four minutes (McDowell and Newell, 1996), but longer completion times have been reported for older people (Siu et al., 1993a,b).

Three articles describe evaluations of the SF-20. Two describe the same residential-home population in the USA (Siu et al., 1993a,b), as shown in Table 3.2. The remaining study describes a community-based population in Canada (Carver et al., 1999): see Table 3.2. The results given below are derived from these studies.

Reliability

The results of reliability testing are reported in Table 4.2. Following interview administration, high levels of internal consistency reliability were reported for each multi-item domain in the range 0.76 (Physical Function) to 0.85 (Health Perception) (Carver et al., 1999). Telephone administration to a separate group with an average retest period of 22 days (range: 11-44 days) demonstrated a high level of reliability (0.96) (Carver et al., 1999).

Validity

(i) Socio-demographic variables and health-service use

When compared to the general population, older respondents scored lower on all SF-20 domains except mental health (Carver et al., 1999). With the exception of mental health, there were no statistically significant differences in domain scores between groups defined by sex.

The SF-20 was completed by old-old people on admission to residential care and again at a follow-up assessment (median 557 days) (Siu et al., 1993a,b). With the exception of BP, low scores on all SF-20 domains, and particularly MH, were predictive of future placement in nursing care. A low score on GH was predictive of future hospitalisation.

(ii) Construct validity: other instruments

Correlations between the SF-20 PF domain and other instruments that had hypothesised associations ranged from 0.51 (Spitzer Quality of Life Index [SQL]) through 0.63 (Barthel Index [BI]) and 0.65 (OMFAQ index) to 0.67 (OMFAQ-IADL) (Carver et al., 1999), as shown in Table 4.3. Correlation between the SF-20 RF domain and the same instruments ranged from 0.48 (BI) through 0.55 (SQL) and 0.56 (OMFAQ-ADL) to 0.59 (OMFAQ index). Correlations between domains that did not have hypothesised

associations ranged from 0.20 (SF-20 PF with the modified Mini-Mental State Examination [m-MMSE]) to 0.27 (SF-20 RF with m-MMSE). Correlations between the SF-20 mental health domain and WONCA/COOP charts ranged from –0.05 (Physical Fitness) to –0.71 (Emotional Condition) (Kempen et al., 1997).

(iii) Validity: other
The content validity of the SF-20 was assessed (Carver et al., 1999) against a theoretical description of quality of life in older people, comprising perceived quality of life, psychological well-being, and behavioural competence (Lawton, 1991). Health professionals working in geriatric rehabilitation highlighted the omission of important domains from the SF-20 when used for assessing older people, for example, memory, cognitive function, and self-administration of medication. In addition, several activities with different functional demands were combined inappropriately, for example, eating, dressing, bathing, and using the toilet.

Factor analysis described four factors (Carver et al., 1999). One GH item ('I have been feeling bad lately') grouped together with all MH domain items on one factor, and with the remaining GH items onto a second factor. PF and RF items loaded across two additional factors but did not describe domains that were entirely consistent with the SF-20.

Responsiveness
Responsiveness was assessed against the external criterion of improvement or deterioration in performance-based tests: gait, balance, or 50-foot walk time (Siu et al., 1993b). The SF-20 PF had moderate responsiveness (ES –0.43) where performance had deteriorated (n=43), but poor responsiveness (ES 0.10) where function had improved (n=32). The SF-20 MH and GH domains were not responsive to deterioration. When assessed against change in SF-20 score, changes in the performance tests were all in the hypothesised direction; however, only change in gait and balance achieved statistical significance.

In the same group, comparable levels of responsiveness were found for the SF-20 PF and COOP physical function chart. Change score correlations between SF-20 domains and COOP charts ranged from 0.05 (SF-20 PF with COOP physical function) to 0.74 (SF-20 BP with COOP bodily pain). The average correlation between change scores was 0.37.

Receiver Operating Characteristic curves were used to assess the sensitivity and specificity of the SF-20 PF and COOP physical function chart to change in performance-based tests (external criteria) (Siu et al., 1993b). Where the COOP physical function chart was unable to discriminate better than chance on any performance test, the SF-20 PF discriminated better than chance for deterioration in balance and gait.

Acceptability
Average SF-20 self-completion time by community-dwelling older people was five to seven minutes; completion rates exceeded 95% (Carver et al., 1999).

Precision

Ceiling effects have been reported for all domains (ranging from 8.1% for MH to 65.5% for SF). In four domains, namely PF, BP, RF, and SF, ceiling effects exceeded 20% (Carver et al., 1999). Floor effects exceeding 20% were reported for the RF domain (27.6%, as compared with 5.1% for SF and 13.3% for BP).

m) SF-36: Medical Outcomes Study 36-item Short Form Health Survey (Ware and Sherbourne, 1992; Ware et al. 1994; Ware, 1997)

The Medical Outcomes Study (MOS) Short Form 36-item Health Survey (SF-36) is derived from the work of the Rand Corporation during the 1970s (Ware and Sherbourne, 1992; Ware et al. 1994; Ware, 1997). It was published in 1990 after criticism that the SF-20 was too brief and insensitive. The SF-36 is intended for application in a wide range of conditions and with the general population. Ware et al. (1994; 1997) proposed that the instrument should capture both mental and physical aspects of health. International interest in this instrument is increasing, and it is by far the most widely evaluated measure of health status (Garratt et al., 2002a).

Items were derived from several sources, including extensive literature reviews and existing instruments (Ware and Sherbourne, 1992; Ware and Gandek, 1998; Jenkinson and McGee 1998). The original Rand MOS Questionnaire (245 items) was the primary source, and several items were retained from the SF-20. The 36 items assess health across eight domains (Ware, 1997), namely bodily pain (BP: two items), general health perceptions (GH: five items), mental health (MH: five items), physical functioning (PF: ten items), role limitations due to emotional health problems (RE : three items), role limitations due to physical health problems (RP: four items), social functioning (SF: two items), and vitality (V: four items), as shown in Table 4.1. An additional health transition item, not included in the final score, assesses change in health. All items use categorical response options (range: 2-6 options). Scoring uses a weighted scoring algorithm and a computer-based programme is recommended. Eight domain scores give a health profile; scores are transformed into a scale from 0 to 100 scale, where 100 denotes the best health. Scores can be calculated when up to half of the items are omitted. Two component summary scores for physical and mental health (MPS and MCS, respectively) can also be calculated.

The SF-36 can be self-, interview-, or telephone-administered.

67 articles describe evaluations of the SF-36 in an older population, as shown in Table 3.3. Most of the studies describe populations from North America across a range of hospital and community settings; two of the studies describe evaluations in Japanese populations (Suzuki et al., 2002; Seki et al., 2003). 20 studies describe populations in the UK. The results given below are derived from these articles.

Reliability

The results of reliability testing for the SF-36 are shown in Table 4.2. Numerous authors have reported moderate to high levels of internal consistency reliability for all domains ranging from 0.49 (SF) to 0.96 (PF). Low to high levels of test-retest reliability have been reported for SF-36 domains ranging from 0.24 (SF) to 0.87 (GH). Where most domains have high levels of reliability (greater than 0.70), the SF and RE domains have consistently lower levels of reliability (less than 0.70): see Table 4.2.

Several studies report item-total correlation for the SF-36, as shown in Table 4m.i. Low levels (less than 0.40) have been found for the GH (range 0.32 to 0.38) (McHorney et al., 1994a; Wood Dauphinee et al., 1997) and MH domains (0.38) (Beusterien et al., 1996).

Item discriminant validity, that is, where an item correlates more highly with its proposed domain than with other domains, was found across all domains in four studies (McHorney et al., 1994a; Beusterien et al., 1996; Wood Dauphinee et al., 1997; Stadnyk et al., 1998). The results from three of the studies are shown in Table 4m.ii.

Scaling success rates, where the percentage of scaling successes (positive correlation with hypothesised domains) is reported relative to the total number of scaling tests with other domains, were provided by three authors (McHorney et al., 1990, cited by McHorney, 1996; McHorney et al., 1994a; Beusterien et al., 1996) as shown in Table 4m.iii. Rates were generally high for all domains, the lowest being for GH. When completed by the old-old (McHorney et al., 1994a) and people with cognitive impairment (McHorney et al., 1990, cited by McHorney, 1996), scaling success rates were lower, particularly for the GH and V domains.

The Response Consistency Index (RCI), proposed by the instrument developers as an internal consistency check on 15 item pairs (Ware, 1997), was assessed (Beusterien et al., 1996; Stadnyk et al., 1998). In a group of young-old people with depression who self-completed the instrument, a high level of response consistency (91.7% at baseline, 95.1% at six weeks), consistent with results with a representative US population, was found (Beusterien et al., 1996). Following interview administration to a group of frail old-old people, the RCI could not be applied for 23% of item pairs due to missing data (Stadnyk et al., 1998). Where assessment was possible, response consistency was found in 74% of respondents; in 21% there was one inconsistency, in 4.5% two or more. Most inconsistencies were found for the SF domain (7.4%), one pairing from the PF domain (walk one block with moderate activities: 5.7%), and one from the V domain (energy with feeling worn out: 5.1%).

Validity

(i) Socio-demographic variables and health-service use

The majority of studies reviewed supported the ability of specific SF-36 domains to discriminate between different socio-demographic features or health-related variables when completed by groups of older people, as shown in Table 4.3.

Age

Several studies investigated whether levels of health as measured by the SF-36 varied with age. The strongest evidence of a decline in health with age was found for SF-36 PF (Mangione et al., 1993; McHorney et al., 1994b; Hayes et al., 1995; Dexter et al., 1996; Brazier et al., 1996; Anderson et al., 1996; Schofield and Mishra, 1998; Walters et al., 2001; Inaba et al., 2003), Vitality (Mangione et al., 1993; Hayes et al., 1995; Brazier et al., 1996), and RP domain scores (McHorney et al., 1994b; Walters et al., 2001). With the exception of the PF domain, when compared to both the general population and young-old age-groups, statistically significantly better health scores have been reported for older age-groups across all domains (Dexter et al., 1996; Baldassarre et al., 2002).

Table 4m.i Item-total correlation

		SF-36 domains							
Author	***Age (yrs)***	***PF***	***RP***	***BP***	***GH***	***V***	***SF***	***RE***	***MH***
McHorney et al. (1994a)	65-74	0.49-0.78	0.67-0.72	0.74	0.38-0.69	0.66-0.77	0.72	0.62-0.69	0.64-0.77
McHorney et al. (1994a)	>75	0.45-0.78	0.63-0.74	0.68	0.34-0.77	0.60-0.71	0.71	0.61-0.73	0.57-0.77
Beusterien et al. (1996)	67	0.40-0.77	0.57-0.73	0.72	-	0.60-0.72	0.40-0.49	0.51-0.54	0.38-0.62
Wood Dauphinee et al. (1997)	70.1	0.41-0.86	0.70-0.77	0.81	0.32-0.67	0.58-0.69	0.74	0.73-0.88	0.46-0.61
Stadnyk et al. (1998)	>80	0.56-0.82	0.56-0.70	0.69	0.40-0.65	0.53-0.67	0.56	0.74-0.81	0.53-0.70

Table 4m.ii Item discriminant validity

	SF-36 domains							
Author	***PF***	***RP***	***BP***	***GH***	***V***	***SF***	***RE***	***MH***
Beusterien et al. (1996)	0.02-0.53	0.03-0.48	0.06-0.42	NR	0.15-0.48	0.19-0.43	0.07-0.43	0.07-0.44
Wood Dauphinee et al. (1997)	–0.03-0.60	0.21-0.62	0.38-0.69	0.03-0.71	0.24-0.62	0.22-0.64	0.07-0.62	0.10-0.52
Stadnyk et al. (1998)	–0.25-0.43	–0.08-0.46	0.26-0.46	–0.24-0.42	–0.23-0.42	–0.10-0.45	–0.10-0.29	–0.23-0.49

Table 4m.iii Scaling success rates (%)

		SF-36 domains							
Author	***Age (yrs)***	***PF***	***RP***	***BP***	***GH***	***V***	***SF***	***RE***	***MH***
McHorney et al. (1994a)	65-74	96	100	100	85	100	100	100	100
McHorney et al. (1994a)	>75	94	97	69	63	66	100	100	100
Beusterien et al. (1996)	60-86	100	100	100	NR	100	100	100	100
McHorney et al. (1990)*	>60	97.5	97	100	65	72	94	100	100
McHorney et al. (1990)	>60	100	100	100	100	100	100	100	100

Key:

PF physical function ***RP*** role-physical ***BP*** bodily pain ***GH*** general health ***V*** vitality ***SF*** social function ***RE*** role-emotional ***MH*** mental health

* includes people with cognitive impairment

In addition, despite a significant decline in health for the older population as reported by health transition items or external diagnostic criteria, GH (Mangione et al., 1993; Hayes et al., 1995; Doraiswamy et al., 2002) and SF scores (Doraiswamy et al., 2002) have been reported as not differing between age-groups in the general older population. Despite a trend towards poorer health among older people with Parkinson's disease, statistically significant differences between the domains scores of young-old and old-old respondents were not found for any domain (Hobson and Meara, 1997).

Further studies have reported constant MH scores (McHorney et al., 1994b; Walter et al., 2001) or improved MH scores (McHorney et al., 1994b; Schofield and Mishra, 1998; Baldassarre et al., 2002) with advancing age. This has also been reported in older people diagnosed with moderate or severe depression, where people aged over 70 years scored more highly on SF-36 MCS and Vitality domains than people aged between 60 and 70 years (Doraiswamy et al., 2002).

Following breast reconstruction surgery for cancer, women aged over 65 years scored more highly on SF-36 domains relating to mental health (SF, RE, MH) than women aged under 65 years, but younger women had better physical health (Girotto et al., 2003). Similarly, at three months after surgery for coronary heart disease, young-old women had better physical health scores, but poorer emotional health scores, than older women (Baldassarre et al., 2002).

When defined by age (younger or older than 65 years), four SF-36 domains (PF, RP, BP, GH) discriminated between patients with end-stage renal disease receiving haemodialysis, with statistically significantly higher domain scores in the older group (Rebello et al., 2001). With the exception of RP and V domains, all domains discriminated between patients defined by age who received a renal transplant, with statistically significantly higher domain scores in the older group. Similarly, the SF-36 PCS discriminated between patients defined by age for those receiving haemodialysis and transplant, with significantly higher PF scores in the older group.

Sex

In a community-based population, after adjusting for age, women had poorer health scores than men across all SF-36 domains (Walters et al., 2001; Doraiswamy et al., 2002; Inaba et al., 2003). Similarly, in a community-based survey across six European settings, women reported worse GH scores than men (Heslin et al., 2001). In a Scottish population, women had worse scores for PF, SF, and MH domains (Lyons et al., 1994). At one year post-stroke, women had better RE, MH, and SF scores than men (Anderson et al., 1996).

Health status

Several authors have reported the ability of all SF-36 domains to discriminate between groups with and without identified health problems (Anderson et al., 1996; Jenkinson et al., 1995; Ho et al., 2001; Overcash et al., 2001; Baldassarre et al., 2002; Doraiswamy et al., 2002; Ekman et al., 2002) or long-standing disabilities (Lyons et al., 1994) and, with the exception of MH, to discriminate between groups with and without long-standing illness (Brazier et al., 1996). All domains discriminated between groups defined at one year post-stroke by level of independence in physical function (Barthel Index) or mental health (General Health Questionnaire-28) (Anderson et al., 1996). The PF, RP, SF, and GH domains discriminated between levels of disease severity in older people with Parkinson's disease (Hobson and Meara, 1997). As hypothesised, patients using a

community-based continence service had lower scores for the SF-36 PF, SF, and GH domains, whilst patients using mental health services had lower MH scores (Hill and Harries, 1994; Hill et al., 1996).

The SF-36 MH and SF domains discriminated between groups with and without anxiety or depression, and between people with possible and probable morbidity (Sharples et al., 2000). The hypothesised correlation between SF-36 scores and self-reported morbidity and symptoms was reported. For example, respondents with self-reported chest or non-chest pain, or arthritis or rheumatism, reported higher BP scores than those without self-reported pain, and more highly than those reporting heart problems. Similarly, those reporting nervous or emotional problems reported higher MH scores than those not reporting these symptoms, and more highly than those reporting hypertension.

For patients receiving knee-replacement surgery for osteoarthritis, the SF-36 discriminated between groups differing in co-morbidity and self-reported health (Bombardier et al., 1995). However, the Western Ontario and McMaster Universities Osteoarthritis Index (WOMAC), a disease-specific instrument, provided better discrimination between patients differing as to outcome of, and satisfaction with, knee surgery. With the exception of BP, SF-36 domains discriminated between trauma survivors aged 65 years and over (mean of 2.8 years post-injury) and age-adjusted normal values for an uninjured population (Inaba et al., 2003).

Both the SF-36 PCS and MCS scores discriminated between older women from the general population (normative values for those aged 65 years and above) and those with coronary heart disease prior to surgical intervention; there was no statistically significant between-group difference at three months post-surgery (Baldassarre et al., 2002). In a separate study, all SF-36 domains discriminated between groups defined by the diagnosis of chronic heart failure (Ekman et al., 2002). Both the SF-36 MCS and PCS discriminated between groups with and without dyspnoea (Ho et al., 2001).

The SF-36 did not discriminate between older people defined by their level of hearing impairment (Morgan et al., 2002). Hearing impairment explained 4.4% of score variation in the SF-36 MH domain. When defined by scores on the Hearing Handicap Inventory for the Elderly, a statistically significant correlation between hearing disability and three SF-36 domains was found: hearing disability explained 11%, 9%, and 4.4% of score variation in the MH, PH, and BP domains, respectively.

The SF-36 RP and SF domains discriminated between groups defined by fear of falling (Suzuki et al., 2002). The PF, GH, and V domains discriminated only between females defined by their fear of falling. The SF-36 PF and RP domains and PCS discriminated between groups defined by the experience of extremity fractures in the previous ten years, but statistical significance was not reported (Wildner et al., 2002).

Hospital-service use
As hypothesised, older people consulting their doctor during the preceding two weeks had worse scores on several domains (PF, MH: Lyons et al., 1994; Brazier et al., 1996) (RE: Lyons et al., 1994) or all domains (Walters et al., 2001). Outpatient attendance in the previous three months was associated with worse scores on all domains (Lyons et al., 1994) or most domains, the exceptions being PF and V (Brazier et al., 1996). Casualty attendance over the same period was associated with worse scores on four

domains, namely BP, PH, RP, and SF (Brazier et al., 1996). With the exception of MH (Lyons et al., 1994), and RE and V domain scores (Brazier et al., 1996), hospital inpatient stay was associated with worse scores for most domains. After adjusting for age and mode of administration, hospital outpatients had worse scores on the PF and RP domains when compared to general practice patients (Hayes et al., 1995). With the exception of BP, a trend towards higher scores across seven domains was found for a community-based low-care group when compared to high-care groups; group difference reached statistical significance for PF scores (Murray et al., 1998). The SF-36 GH domain discriminated between community-dwelling older people and those living in sheltered accommodation, sheltered housing, or health-care institutions; community-dwelling older people reported better health (Heslin et al., 2001). With the exception of GH, older people living alone had lower scores on all domains than those living with others (Walters et al., 2001).

The SF-36 discriminated between older people defined by social needs according to a social work-specific questionnaire; lower SF-36 scores were strongly associated with the need for social needs assessment (Berkman et al., 1999).

(ii) Construct validity: other instruments
The association between the SF-36 and numerous other instruments has been evaluated, as shown in Table 4.3.

Physical function
Correlation of the SF-36 PF domains with interview-administered measures of ADL were as follows: with Katz-ADL: 0.30, with OMFAQ-IADL: 0.36 (Reuben et al., 1995). Correlation of the SF-36 with a modified Katz-ADL ranged from 0.30 (RP) to 0.79 (PF) (Sharples et al., 2000).

Correlation of the SF-36 PF domains with performance tests of functional ability were as follows: with the Physical Performance Test 0.59, with the National Institute on Aging Battery 0.65 (Sherman and Reuben, 1998), with the Timed Up and Go test –0.26 (Jaglal et al., 2000). Correlation of the SF-36 with the Guralnik Performance Test (Guralnik et al., 1994) ranged from 0.34 (RP) to 0.74 (PF) (Sharples et al., 2000). With the exception of MH and RE domains, all SF-36 domains discriminated between groups defined by a performance-based assessment (Sherman and Reuben, 1998).

Correlation of the SF-36 PF domains with patient-reported measures of ADL were as follows: with the Functional Status Questionnaire (FSQ)-ADL 0.51 (Reuben et al., 1995) and 0.56 (Sherman and Reuben, 1998), and with the FSQ-IADL 0.76 (Reuben et al., 1995) and 0.78 (Sherman and Reuben, 1998). Correlation of the SF-36 with ADL assessment ranged from –0.37 (PF) to –0.43 (RP) (Andresen et al., 1999). Correlations between three physical activity instruments and the SF-36 MH and BP domains ranged from 0.17 to 28; correlation of these instruments with the SF-36 PF and GH domains ranged from 0.26 to 0.42, and were in accordance with hypotheses (Harada et al., 2001).

Correlation between the SF-36 PF and Functional Independence Measure (FIM) when completed by cognitively intact or cognitively impaired groups of older people was 0.53 and 0.33, respectively, in accordance with hypotheses (Seymour et al., 2001). Correlation between all SF-36 domains and the FIM ranged from –0.029 (RE) to 0.22 (MH), and was smaller than hypothesised for both groups; group difference reached statistical significance for the PF domain. Correlation between the SF-36 RP, PF, and

PCS and the Functional Disability Index ranged from –0.56 (RP) to 0.77 (PF) (Wildner et al., 2002). Correlation among the SF-36 RP, PF, and PCS scores ranged from 0.68 (RP with PF) to 0.86 (PF with PCS).

In patients having undergone knee replacement surgery, correlation between SF-36 and the WOMAC domains ranged from 0.15 (SF-36 MH with WOMAC stiffness) to –0.55 (SF-36 BP with WOMAC pain) (Bombardier et al., 1995). Correlation between the SF-36 BP and PF domains and related WOMAC domains were –0.55 and –0.50, respectively, and were smaller than hypothesised.

Mental health

Nursing-home residents without cognitive impairment completed the SF-36 and several other instruments by interview (Andresen et al., 1999). Correlations between SF-36 domains and the Geriatric Depression Scale (GDS) with hypothesised associations ranged from 0.34 (RE) to –0.71 (MH). Correlations between SF-36 domains and the GDS that did not have hypothesised associations ranged from –0.25 (PF) to –0.44 (PF). Correlations between the SF-36 and the mini-Mental State Examination were less than 0.18 (Stadnyk et al., 1998; Andresen et al., 1999).

Following completion by older people with depression, correlations between the SF-36 and the clinician-assessed Hamilton Rating Scale for Depression (HAMD) ranged from –0.12 (PF) to –0.57 (MH and MCS) (Beusterien et al., 1996). Correlations between SF-36 and HAMD domains with hypothesised associations were –0.43 (RE), –0.45 (SF), and –0.57 (MH and MCS). In a different population of older people with depression, correlations between specific SF-36 domains and the HAMD were –0.20 (MH), –0.26 (RP), and –0.32 (PF) (Doraiswamy et al., 2002). Correlation between the SF-36 MCS and the Quality of Life Depression Scale was –0.69. Correlation between the SF-36 and the depression-specific Clinician's Global Impression of Severity and Improvement ranged from –0.08 (PH) to –0.53 (MH and MCS) (Beusterien et al., 1996).

Interview administration to severely ill older people with chronic heart failure showed correlations between the SF-36 and Sense of Coherence scale ranging from 0.10 (BP) to 0.46 (MH); for age- and sex-matched healthy controls, correlations ranged from 0.00 (BP) to 0.39 (RE) (Ekman et al., 2002).

Quality of life

Correlations between the SF-36 and the generic Assessment of Quality of Life instrument (AQoL) ranged from 0.04 (BP and RP with AQoL physical senses) to 0.64 (PF with AQoL independent living) (Osborne et al., 2003). Correlations between the SF-36 MCS and PCS and the AQoL utility score were 0.41 and 0.37, respectively. In patients with coronary heart disease awaiting surgery, correlations between the SF-36 MCS and PCS and the utility Feeling Thermometer were 0.30 and 0.50, respectively, and in accordance with hypotheses (Baldassarre et al., 2002).

The strongest correlations between SF-36 and NHP domains ranged from –0.52 (MCS with NHP emotional reaction) to –0.88 (V with NHP energy) (Crockett et al., 1996). Smaller correlations ranged from 0.00 (GH with NHP pain, RP with NHP sleep) to –0.49 (SF with NHP energy). As hypothesised, the SF-36 PCS was most strongly correlated with SF-36 physical function domains, and the MCS with mental health and social function domains. GH had a larger correlation with PCS than with MCS (values not reported).

In patients with a range of medical conditions, correlations between the SF-36 and NHP domains that had hypothesised associations ranged from –0.25 (SF with NHP social isolation) to –0.76 (PF and NHP physical mobility) (Stadnyk et al., 1998). Correlations between domains without hypothesised associations ranged from 0.02 (RP with NHP emotional reaction) to –0.41 (GH with NHP emotional reaction). As hypothesised, correlations in excess of 0.60 were found between the SF-36 PF domain and related instruments, namely the Barthel Index, NHP physical mobility, OARS-IADL, and the Spitzer Quality of Life Index.

In community-based older males, correlations between the SF-36 and Sickness Impact Profile (SIP) that had hypothesised associations ranged from 0.67 (SF) to 0.78 (PF) (Weinberger et al., 1991). Correlations between three SF-36 domains and the SIP ranged from –0.21 (GH with SIP psychosocial function) to –0.47 (PF with SIP physical function) (Andresen et al., 1995). In a subsequent study, correlations between the SF-36 and SIP domains ranged from 0.02 (SF with SIP work) to –0.86 (PF with SIP physical function) (Andresen et al., 1998b). Correlations between the SF-36 and SIP index ranged from –0.41 (MH) to –0.85 (PF).

Correlations between the SF-36 and Functional Profile Inventory (FPI) domains ranged from –0.03 (SF with FPI spiritual activity) to –0.69 (PF with FPI total and physical exercise) (Larson et al., 1998). Following completion by cancer patients, correlations between the SF-36 and Functional Assessment of Cancer Therapy Scale (FACT-G) ranged from 0.02 (RP with FACT-G relationship with doctor) to 0.61 (SF with FACT-G functional well-being) (Overcash et al., 2001). Both SF-36 summary scores had a correlation of 0.53 with the FACT-G total score.

(iii) Validity: other

Predictive validity

The predictive validity of the SF-36 in a population of chronically ill older people has been assessed in terms of mortality at four years, use of inpatient resources at two years, and visits to the General Practitioner (GP) at two years (McHorney, 1996). The GH and PF domains were most predictive of mortality whilst PF, RP, and BP were most predictive of hospitalisation. The BP, GH and V were most predictive of visits to the GP. MH and RE domains were least predictive of all scenarios.

Internal validity

Analysis of the SF-36 in outpatients with various medical conditions and aged over 50 years suggested that those aged over 70 years were more optimistic and reported being better off in terms of their pain, mental and general health, and physical role (BP, MH, GH, RP) than those aged between 50 and 70 years (Wolinsky and Stump, 1996). Factor analysis supported the eight-factor solution proposed by the developers. A subsequent confirmatory analysis gave a nine-factor model; the additional factor ‘health optimism’ included two general health items, namely ‘getting ill’ (item 11a) and ‘getting worse’ (item 11c). Factor analysis following completion by groups of young-old (Dexter et al., 1996) and frail old-old (Stadnyk et al., 1998), supported the two-factor solution of mental and physical health, and the eight-domain structure proposed by the instrument developers.

Proxy completion

Moderate to high levels of agreement between cognitively intact patients and known lay proxies were found for the more observable health domains, for example, physical

function, role-physical function, and general health. Moderate levels of agreement were found for the remaining SF-36 domains (Pierre et al., 1998; Yip et al., 2001). Agreement was lower when evaluated at the item level (Yip et al., 2001).

SF-36 completion by cognitively intact older people was compared with completion by health professional proxies and lay proxies (Ball et al., 2001). Agreement was closer between patients and health professional proxies, ranging from 0.32 (RE) to 0.69 (BP), than between patient and lay proxies: range 0.10 (RP) to 0.50 (BP). Health professional proxies scored lower than patients on all but the BP and MH domains, whereas lay proxies scored lower than patients on all domains. Difference in agreement between professional and lay proxies was statistically significant for PF, BP, and MH. As hypothesised, while strong levels of correlation were found between the Functional Independence Measure (FIM) and the SF-36 (PF) when completed by patients (0.53) and health professionals (0.66), correlation with lay proxies was smaller (0.38), and suggests that informed health professionals are better able to interpret a patient's health status than patient-nominated lay proxies.

Responsiveness

Following completion by community-dwelling older females, Brazier et al. (1996) evaluated the ability of the SF-36 to discriminate between hypothetical health states, improvement in health state when assessed against change in health-service use, and change in long-standing illness and by age-group. A hypothetical improvement in health status defined by having or not having a long-standing illness, or by hospitalisation in the previous 12 months, were associated with small to large effect sizes (ES) for SF-36 domains: ES for long-standing illness ranged from 0.31 (MH) to 0.96 (PF) and for hospitalisation ranged from –0.03 (MH) to 0.81 (PF). Other hypothetical improvements in health status, namely GP consultation in the previous two weeks, outpatient attendance in the previous three months, and moving from the over 85-year age-group to the 75 to 79-year age-group, were associated with low to moderate ES in the range 0.05 (RE) for outpatient attendance to 0.57 (RP) for GP consultation.

Rehabilitation/care coordination

Minimal SF-36 score change after three months was reported following community-based continence or mental health-care programmes (Hill and Harries, 1994; Hill et al., 1996). Other assessment instruments were lacking, but in-depth interviews suggested that the SF-36 did not address areas of health that participants considered important, for example, positive change in mood, feelings, and outlook.

With the exception of GH and RE domain scores, statistically significant mean score change was found for all domains following a rehabilitation programme for frail older people. ES were small and ranged from 0.00 (for RE) to 0.35 (for SF) (Stadnyk et al., 1998). The NHP demonstrated similar low levels of responsiveness, but larger ES were found for the Spitzer Quality of Life Index (0.73), the Barthel Index (0.68), and the OMFAQ-IADL (0.61), which may more appropriately reflect the goals of the rehabilitation programme.

Following a home-care nursing programme for patients with acute or chronic illness, all SF-36 domains, with the exception of GH, were more responsive to change in health than the older people-specific Quality of Life Profile-Senior Version (QOLPSV), with statistically significant improvement in four out of nine QOLPSV domain scores (Irvine

et al., 2000). The SF-36 also discriminated between groups defined by the number of nurse visits.

In patients with chronic debilitating illness, SF-36 score improvement in three or fewer domains over 12 months was frequently associated with deterioration in a similar number of domains (Wolinsky et al., 1998). Consistency in specific domain score change between patients was lacking. The majority of respondents remained unchanged on between three and five domains, and baseline characteristics were generally unrelated to change.

With the exception of PF, mean score improvement across all SF-36 domains was greater in patients whose depression improved over six weeks than in those for whom it remained unchanged. MH, RE, and V had the greatest score improvement, in accordance with hypotheses (Beursterien et al., 1996).

With the exception of GH and SF scores, all SF-36 domains and clinical observations of day-hospital patients showed improvement in health at six months (Fowler et al., 2000). Scores for the Barthel Index (BI), Philadelphia Geriatric Center Morale Scale, and Geriatric Depression Scale (GDS) suggested a decline in health. Due to a change from interview administration at baseline to self- and/or interview administration at six months, score change must be interpreted with caution. Score-change correlation between SF-36 domains (except RP, RE) and other instruments ranged from 0.08 (GH with BI) to –0.55 (GH with GDS).

A statistically significant improvement in all SF-36 domains was found at ten weeks for a community-based exercise group; change was not statistically significant for the control group of usual activity (Cochrane et al., 1998). ES for the exercise group ranged from 0.27 (PF) to 0.93 (RP).

Participants in a trial of cardiac rehabilitation had statistically significant score improvements for BP, GH, MH, and V domains at six months. Mean changes ranged from 7.0 for V to 13.0 for BP; ES statistics ranged from 0.34 for V to 0.60 for BP (Seki et al., 2003). GH had the largest ES (0.85). Change was not statistically significant for the usual care group.

Institutionalisation at 18 months was defined as an external criterion for health deterioration following the assessment of care coordination versus usual care in community-dwelling older people (Osborne et al., 2003). High levels of responsiveness were found for the SF-36 PCS, and PF and GH domains when evaluated using both Relative Efficiency and Receiver Operating Characteristic curves. Three domains, namely PF, BP, and V, discriminated between baseline differences in people who remained community-dwelling or were living in an institution at 18 months.

Drug therapy

Following four weeks of drug treatment for congestive heart failure, small to moderate ES statistics were found for SF-36 domains ranging from –0.01 (BP) to 0.31 (RP) (Jenkinson et al., 1997). There was little enhancement in ES when recalculated to include patients reporting improvement global health (43%). It is suggested that standardized instruments may not usefully reflect change in health status of importance to patients.

Data pooling from three placebo-controlled drug trials for osteoarthritis indicated a statistically significant mean score improvement across all SF-36 domains after two weeks for those receiving the active drug, ranging from 1.7 (GH) to 19.6 (RP), and for both summary scores: 2.4 for MCS and 4.8 for PCS (Lisse et al., 2001). At 12 weeks, with the exception of GH and RE, a statistically significant score improvement was found for all domains and ranged from 2.2 (MH) to 17.9 (RP).

Diabetics with poor glycaemic control who received insulin treatment as part of a clinical trial reported statistically significant improvements after four weeks in the MH, RE, RP, and V domains; mean score changes ranged from 9.0 (ES 0.47) for MH to 14.0 (ES 0.64) for V (Reza et al., 2002). Improvement was sustained at 12 weeks for the MH, SF, and V domains; mean score changes ranged from 11 (ES 0.58) for MH to 16.0 (ES 0.73) for V. For those who continued with oral medication, there was no statistically significant improvement in health status.

Surgical intervention
The SF-36 PF domain was responsive to change at six weeks and six months post-hip fracture repair, and for the difference between these two assessment periods (ES range 0.70 to 1.3) (Jaglal et al., 2000). PF, RP, and BP discriminated between groups defined by their status pre-fracture versus six weeks post-surgery, pre-fracture versus six months post-surgery, and status at six weeks versus status at six months. At six months, GH had reached, and MH exceeded, the pre-morbid level. SF and RE discriminated between status pre-fracture and status at six weeks only.

The SF-36 BP and PF domains were responsive to change post-hip fracture repair in 45 patients reporting a poorer outcome at four months. ES ranged from 0.82 (BP) to 0.88 (PF); standardised response means ranged from 0.68 (BP) to 0.77 (PF) (Tidermark et al., 2003a). With the exception of RE, the domains discriminated between groups defined by the external criterion ‘good versus less good clinical outcome’. Change score correlation between the SF-36 and EuroQol index ranged from 0.03 to 0.45; the largest correlations were with the BP, V, and PF domains.

Following surgical intervention for coronary artery disease, statistically significant improvements of 7.70 (ES 1.32) and 7.26 (ES 1.23) were found at three months for the SF-36 PCS and MCS, respectively (Baldassarre et al., 2002).

Precision
The developers hypothesised that data quality and scaling may be weaker when the instrument is completed by older people, and that older people may have more sickness than the general population and hence score at the floor of a scale (Ware, 1997). Floor effects in excess of 20% for the RP and RE domains were reported in 12 studies (see Table 3.3). Ceiling effects in excess of 20% for the RP, RE, and SF domains were reported in 16 studies (Table 3.3).

Acceptability
Instrument completion rates varied by mode of administration and were generally higher with interview administration than with self-completion (for example, Hayes et al., 1995; Mallinson, 1998; Parker et al., 1998). Age and mode of administration were found to have independent and statistically significant associations with completion rates (p less than 0.001) (McHorney et al., 1994b; Hayes et al., 1995). The oldest respondents and those with physical or cognitive impairment experienced greater

difficulties with completion than younger respondents and those in better health (Hayes et al., 1995; Parker et al., 1998).

Several authors have suggested that self-completion by the old-old (Lyons et al., 1994; Parker et al., 1998) and in older people following a stroke (O'Mahony et al., 1998) is inappropriate, but this may be true of most patient-reported health instruments (Hayes et al., 1995). For example, in three old-old populations, self-completion by hospital inpatients was associated with lower completion rates (46%) than self-completion by ambulatory hospital outpatients (71%) or general practice clinic patients (93%) (Parker et al., 1998). Completion rates were higher (77%) for interview administration to a sample of hospital inpatients.

Where reported, SF-36 completion rates for interview administration ranged from 73% (Anderson et al., 1996) to 98% (Crockett et al., 1996) in the young-old, and between 77% (Parker et al., 1998) and 100% (Murray et al., 1998; Jaglal et al., 2000; Seymour et al., 2001) in older populations. The exclusion of patients with cognitive impairment from the majority of studies may artificially inflate completion rates (Lyons et al., 1997; Andresen et al., 1999).

Self-completion rates ranged between 56% (Cochrane et al., 1998; Hayes et al., 1995) and 100% (Hayes et al., 1995; Beusterien et al., 1996) in young-old groups, and between 46% (Parker et al., 1998) and 93% (Parker et al., 1998) in older populations. Patients aged over 80 years who had undergone knee arthroplasty were no less likely to respond than younger patients when asked to self-complete a postal questionnaire that included the SF-36 (p equals 0.061) (Bombardier et al., 1995). However, responders to baseline and follow-up questionnaires have been found to be both younger and healthier (according to SF-36 scores) than those who responded to baseline questionnaires only (Andresen et al., 1996). Lower response rates are generally found in studies comprising older populations in poorer health, and from hospital, institutional, or residential-care settings (Hayes et al., 1995; Parker et al., 1998; Walters et al., 2001). In contrast, non-responders to general population surveys are more likely to be younger and male (McHorney et al., 1994b).

Although 91% of UK community-based and hospital outpatient respondents indicated that all or most items were clear, intelligible, and applicable, 43% of these respondents were unable to self-complete the instrument due to physical impairments or unfamiliarity with questionnaire completion; the majority of these respondents were aged over 75 years (Hayes et al., 1995). 14% experienced some difficulty with the multi-choice items. 61% and 12% of respondents omitted one or more items following self- or interview administration, respectively. 20% of community-based respondents in Canada experienced some confusion regarding one or more items during self-administration and telephone-interview administration (Wood Dauphinee et al., 1997). This was particularly associated with items having long question 'stems' which may be difficult to retain. In a smaller, UK-based population, 71% of respondents experienced some difficulty in self-completion for some or all items; 64% required help from a relative or friend (Mallinson, 1998). Unsolicited comments on questionnaires from this population highlighted three areas of completion difficulty, namely item relevance, misunderstanding, and item formatting. For example, the 'double-barrelled' nature of several items caused some confusion.

Several authors have found problems with item applicability or relevance (for example, Hayes et al., 1995; Dexter et al., 1996; Wood Dauphinee et al., 1997; Andresen et al., 1998a; Mallinson, 1998; O'Mahony et al., 1998; Parker et al., 1998; Fowler et al., 2000). These issues arise mainly in response to items related to work and physical activity, particularly vigorous activities: respondents often indicate that they are 'retired' from work or too old to perform such activities. Repetition of an activity was a reason for not completing several items referring to distance walked, namely walking more than one mile, walking several blocks, and walking one block. Several general health perception items were frequently omitted. Where health is assessed in the context of a respondent's own age-group, item 11b 'I am as healthy as anybody I know' was considered to be ambiguous (Mallinson, 1998). Item 11c 'I expect my health to get worse' was viewed as unnecessarily negative (Hayes et al., 1995). Anecdotal reports further confirmed the unpopularity of this item; but this did not affect the reliability of the GH domain (Sharples et al., 2000).

Modifications to items frequently omitted by older respondents have been recommended (Hayes et al., 1995; Hobson and Meara, 1997) but the impact on the status of the SF-36 as a generic instrument or a new older people-specific version should be considered (Hayes et al., 1995). Item completion rates from selected evaluations are shown in Table 4m.iv.

Domain scores may be calculated where half or fewer of the items are omitted; a person-specific estimate, the mean value of the non-omitted items, is substituted (score proration) (McHorney et al., 1994a). Where more than half of the domain items are omitted, domain scores are not calculated. In particular, the omission of items relating to general health, work, and vigorous or physical prevented the calculation of GH, PH, V, and role limitation domains in several studies (for example, Brazier et al., 1996; O'Mahony et al., 1998; Parker et al., 1998). Cautious interpretation of the role limitation and SF domains has been advised due to the lack of participation in certain activities expressed by many older respondents and associated difficulty in answering the vigorous activities items (Fowler et al., 2000). Domain completion rates from selected evaluations are shown in Table 4m.v. As hypothesised, data completion, and hence domain score calculation, was lower with older populations and those in poorer health (McHorney et al., 1994a; Parker et al., 1998).

Older respondents from several studies highlighted the absence of items from the SF-36 addressing issues of confidence and positive change in mood, self-control, carer burden, feelings, and future outlook (Hill and Harries, 1994; Hill et al., 1996). The limited relevance to nursing-home residents of the content of the SF-36 has been highlighted: six items are work-related and nine describe activities generally not undertaken by nursing-home residents (Andresen et al., 1999).

Median interview completion time by patients with a range of medical or orthopaedic conditions was 20 minutes (Stadnyk et al., 1998). A shorter average completion time of 12.5 (5.5) minutes was reported for supervised self-completion in those attending hospital outpatient appointments (Wood Dauphinee et al., 1997). The majority of patients from hospital outpatient or GP clinics completed the SF-36 in ten minutes using self- or interview administration (median 8 minutes, range: 4 to 30 minutes, Hayes et al., 1995; range: 10-45 minutes, Hamilton et al., 1996). Completion times of between 13 and 14 minutes were reported for interview administration in geriatric and general medical clinics, respectively (Weinberger et al., 1991).

Table 4m.iv SF-36 item completion rates (%)

			SF-36 Domains							
Author	***Age (yrs)***	***Administration***	***PF***	***RP***	***BP***	***GH***	***V***	***SF***	***RE***	***MH***
McHorney et al. (1994a)	<65	Self	91	96	94	94	95	95	97	92
McHorney et al. (1994a)	65-74	Self	83	90	90	91	91	92	93	88
McHorney et al. (1994a)	>75	Self	76	86	87	84	85	91	88	84

Table 4m.v SF-36 domain completion rates (%); domain score calculation for complete data or after score proration (greater than 50% of data available)

			SF-36 Domains							
Author	***Age (yrs)***	***Administration***	***PF***	***RP***	***BP***	***GH***	***V***	***SF***	***RE***	***MH***
Brazier et al. (1996)*	mean 80.1	Self	68.1	86.5	94.3	80.5	81.4	88.9	83.8	83.8
Brazier et al. (1996)	mean 80.1	Self	93.0	91.1	95.4	84.9	90.3	95.7	86.5	91.1
McHorney et al. (1994a)	<65	Self	97	97	100	98	98	99	97	99
McHorney et al. (1994a)	65-74	Self	95	95	99	96	96	99	95	98
McHorney et al. (1994a)	>75	Self	94	91	99	94	94	100	91	99
Parker et al. (1998)	77.0	Self (GP out-patient)	91	97	100	94	97	100	97	94
Parker et al. (1998)	80.0	Self (outpatient)	91	84	86	79	87	89	86	86
Parker et al. (1998)	76.0	Self (inpatient)	73	77	87	80	85	92	73	62
Parker et al. (1998)	76.0	Interview (in-patient)	95	96	85	96	95	94	95	75
Stadnyk et al. (1998)	>80	Interview	100	98.6	94.5	99.3	99.3	92.5	97.2	99.3

Key:
PF physical function ***RP*** role-physical ***BP*** bodily pain ***GH*** general health ***V*** vitality ***SF*** social function ***RE*** role-emotional ***MH*** mental health
* Domain score calculation before recommended score proration (>50% of data available)

n) Sickness Impact Profile (Bergner et al., 1976; revised: Bergner et al., 1981)

The Sickness Impact Profile (SIP) was developed in the USA to provide a broad measure of self-reported health-related behaviour (Bergner et al., 1976; Bergner et al., 1981). It was intended for a variety of applications, including programme-planning and assessment of patients, and to inform policy decision-making (Bergner et al., 1976; Bergner et al., 1981; McDowell and Newell, 1996).

Instrument content was informed by the concept of 'sickness', which was defined as reflecting the change in an individual's activities of daily life, emotional status, and attitude as a result of ill-health (McDowell and Newell, 1996). Item derivation was based on literature reviews and statements from health professionals, carers, patient groups, and healthy subjects describing change in behaviour as a result of illness. The SIP has 136 items across 12 domains: alertness behaviour (AB: ten items), ambulation (A: 12 items), body care and movement (BCM: 23 items), communication (C: nine items), eating (E: nine items), emotional behaviour (EB: nine items), home management (HM: ten items), mobility (M: ten items), recreation and pastimes (RP: eight items), sleep and rest (SR: seven items), social interaction (SI: 20 items) and work (W: nine items).

Each item is a statement. Statements that best describe a respondent's perceived health state on the day the instrument is completed are ticked. Items are weighted, with higher weights representing increased impairment. The SIP percentage score can be calculated for the total SIP (index) or for each domain, where 0 is better health and 100 is worse health. Two summary scores are calculated: Physical function (SIP-PhysF), a summation of A, BCM, and M, and psychosocial function (SIP-PsychF), a summation of AB, C, EB, and SI. The five remaining categories are scored independently. The instrument may be self- or interview-administered.

The Functional Limitation Profile (FLP) is an anglicized version of the SIP (McDowell and Newell, 1996), but evaluations in older people have not been identified. Several abbreviated versions of the SIP have been developed, including a 68-item version (De Bruin et al., 1992), which has been applied with older people (Jannink-Nijlant et al., 1999).

12 articles describe evaluations of the SIP. One study describes a community-based population in Australia (Liddle et al., 1996) and one a hospital-based population in Canada (Page et al., 1995), as shown in Table 3.2. One study evaluates the mobility scale of a 68-item SIP in a community-based population in the Netherlands (Jannink-Nijlant et al., 1999): see Table 3.2. The remaining studies describe various community-based populations in the USA (Table 3.2). The results given below are derived from these studies.

Reliability

Moderate to high levels of internal consistency reliability have been reported for SIP domains, ranging from 0.59 (E) to 0.84 (BCM) (Rothman et al., 1989; Andresen et al., 1998; Jannink-Nijlant et al., 1999), as shown in Table 4.2. High levels of internal consistency reliability have been reported for the SIP index (0.95) (Andresen et al., 1998). There is no evidence of test-retest reliability in an older population.

Following completion by cognitively intact older people, the scores for telephone and face-to-face administrations of the SIP A, M, and BCM domains were compared (Morishita et al., 1995). Correlation between domains in the two modes of administration ranged from 0.89 (A) to 0.97 (M), and was 0.96 for the SIP-PhysF summary score. Mode of administration was not significantly associated with respondents' scores for any category.

Validity

(i) Socio-demographic variables and health-service use

In patients about to undergo heart surgery, the SIP-PsychF summary score discriminated between patients with and without severe co-morbidity (worse health in those with severe co-morbidity) (Page et al., 1995), as shown in Table 4.3. Post-operatively, the SIP-PhysF discriminated between patients with and without severe co-morbidity.

The SIP domain and summary scores discriminated between groups defined by nursing-home residential status or diagnosis of chronic obstructive pulmonary disease (Rothman et al., 1989). As hypothesised, nursing-home residents had higher levels of impairment across all domains and both summary scores. Following completion by respondents who had received intensive care, the SIP domains of BCM, M, A, SR, and HM discriminated between old-old respondents and young-old and middle-aged respondents, with higher levels of impairment in the first group (Kleinpell and Ferrans, 2002).

Completion by community-dwelling older people demonstrated a small correlation between the mean SIP index and summary scores, and the number of reported bed-days and restricted activity days (Andresen et al., 1998a).

(ii) Construct validity: other instruments

Correlations between the SIP-PhysF and both the Barthel Index (BI) and an Index of Activities of Daily Life was 0.74, in accordance with hypotheses (Rothman et al., 1989), as shown in Table 4.3. Correlations between the SIP-PsychF and the Philadelphia Geriatric Center Morale Scale and the Life Satisfaction Index (LSI) were –0.40 and –0.31, respectively, and were weaker than hypothesised. Correlations between the SIP domains and instruments that did not have hypothesised associations ranged from –0.16 (SIP-PhysF with LSI) to 0.47 (SIP-PsychF with BI). Correlation between the two SIP summary scores was 0.67.

Correlations between the SIP index and summary scores and the Functional Profile Inventory (FPI) ranged from –0.08 (index with FPI spiritual activity) to –0.64 (SIP-PhysF summary with FPI body care) (Larson et al., 1998).

Correlations between the SIP index and the Quality of Well-being Scale (QWB) was –0.52; correlation between the SIP summary scores and the QWB ranged from –0.37 (SIP-PsychF) to –0.49 (SIP-PhysF), in accordance with hypotheses (Andresen et al., 1995). Correlations between the SIP index and three SF-36 domains ranged from –0.33 (general health) to –0.47 (physical function). Correlations between SIP summary scores and the SF-36 ranged from
–0.21 (PsychF with SF-36 general health) to –0.40 (PhysF with SF-36 physical function). Correlations between the SIP and variables or instruments that did not have hypothesised associations ranged from 0.14 (Chronic Disease Index) to 0.31 (positive affect).

In a subsequent study, correlations between the SIP index and SF-36 domains ranged from –0.41 (mental health) to –0.85 (physical function), in accordance with hypotheses (Andresen et al., 1998b). Correlations between SIP domains and the SF-36 ranged from 0.02 (W with SF-36 social function) to –0.86 (SIP-PhysF with SF-36 physical function). Correlations between the SIP-PsychF summary score and the SF-36 physical and general health domains ranged from –0.21 (general health) to –0.28 (SF-36 physical function).

Following completion by older males, correlations between SIP and SF-36 domains that had hypothesised associations ranged from 0.67 to 0.78 for social and physical function, respectively (Weinberger et al., 1991).

(iii) Validity: other
The ability of the SIP (68-item) M domain to screen an older population for mobility difficulties and falls was assessed against the Guralnik Performance Test (Guralnik et al., 1994, cited by Jannink-Nijlant et al., 1999). The SIP-M had high sensitivity for poor function (91%) but low specificity for good function (58%). Additionally, the SIP-M discriminated between older people defined as recurrent fallers and non-fallers, and was able to identify people at risk of recurrent falling.

Responsiveness
Following assessment by an occupational therapist, community-dwelling older people were randomly allocated to receive or not to receive recommended modifications to their home-setting; a third group was not assessed (Liddle et al., 1996). Participants completed a package of outcome measures including the SIP at baseline and after six months. Although there was a small mean change in SIP index and physical health scores, which was greater for the intervention group, this did not reach clinical or statistical significance over time or between groups.

Following surgery for coronary heart disease, statistically significant improvement in SIP index and summary scores at six months post-surgery was found (Page et al., 1995).

Precision
Several studies have reported ceiling effects (Weinberger, 1991; Andresen et al., 1995; Andresen et al., 1998a,b). Although end effects were not reported for the SIP index (7.8% scored 0), SIP summary scores had floor effects of 27.3% and 22.0% for SIP-PhysF and SIP-PsychF, respectively. Several domain scores also had floor effects, ranging from 30.5% (for SI) to 86.9% (for W).

Acceptability
Following physician instruction in the use of the SIP and receipt of a patient's completed SIP prior to appointment, some agreement between patients and physicians concerning the presence of disability was found (37%) (Goldsmith and Brodwick, 1989). Agreement was less in a control group where physicians did not receive instructions or review the patient's SIP (22%). Agreement concerning the absence of disability was comparable between groups. The results suggest that the SIP may increase physicians' awareness of functional status, and therefore increase agreement with the patient regarding the presence of functional disability. Where 64% (20 of 31) of physicians indicated that the SIP was helpful and 84% (26 of 31) supported its use for older people with chronic disease, particularly those with complex problems, 63% (25 of 35) had not discussed the SIP with patients and most considered it too long for a

clinical setting. Although 94% (33 of 35) of physicians felt that the SIP could be useful in patient care and 52% (15 of 29) thought it useful in considering management alternatives, a six-month pre-post audit of practice indicated no impact on patient management. The majority of patients (33 of 35, or 94.5%) thought that inclusion of the SIP would be beneficial to care.

Response rates were variable. Self-administration showed response rates ranging from 43% (SIP-M only) (Jannink-Nijlant et al., 1999) to 75.6% for a postal survey containing the SIP and the SF-36 (Andresen et al., 1998a). In a separate study with only a 68.2% study response rate, 100% SIP completion was reported (Andresen et al., 1995).

Response rates of 93% have been reported for interview administration (Weinberger et al., 1991). Interview completion time ranged from 20 to 65 minutes (mean 35 minutes). Longer completion times were associated with speech difficulties (Rothman et al., 1989). Interview completion times for the SIP and SF-36 in specific settings were a) geriatric clinic: 33 minutes versus 15 minutes, b) general medical clinic: 21 minutes versus 14 minutes (Weinberger et al., 1991). The SIP PF domain was found to have a mean telephone completion time of 11.5 minutes (Morishita et al., 1995). In interview administration in both community and residential nursing-home settings, the length of the interview was not considered to be problematic (Rothman et al., 1989).

Longer self-completion times have been reported for the SIP (mean 19.7 minutes) relative to the SF-36 (mean 12.6 minutes) (Andresen et al., 1998a). The SIP mobility domain required only a few minutes for self-completion; despite low completion rates (43%), the authors reported that the instrument was easily understood (Jannink-Nijlant et al., 1999).

The original SIP does not define health within the questionnaire and several respondents were observed to discount their own functional impairments as part of old age (Rothman et al., 1989). Feedback from interviewers and respondents supported inclusion of a definition of health to support greater consistency and simplification of the process of responding to each item.

Approximately 66% of respondents have reported similarly high levels of satisfaction with both the SIP and SF-36 (Andresen et al., 1998a,b). Frequently omitted items were related to sexual activity, with missing responses of between 9.9% and 11.7% (Andresen et al., 1998a and 1998b, respectively), and interaction with children, with missing responses of 8.5% (Andresen et al., 1998a).

o) Spitzer Quality of Life Index (Spitzer et al., 1981)

The Spitzer Quality of Life Index (SQL) was developed during the 1970s for use as a brief evaluative instrument for medical interventions and as a global assessment specifically for patients with terminal cancer or other serious chronic ill-health (Spitzer et al., 1981; McDowell and Newell, 1996). Although not recommended for application in healthy populations, there are reference standards for healthy populations and respondents with various disease states (Spitzer et al., 1981; McDowell and Newell, 1996).

The five items were derived from literature searches, and from a survey of the opinions of patients having a variety of chronic ailments, their relatives, and health professionals with an interest in quality of life assessment. A clear conceptual background guided item selection. Following pilot-testing, content validity was further checked through discussion with patients, clinicians, and researchers. Interview- and self-administered versions of the instrument exist. Both have five domains, namely activity level (AL: what is your main activity?), activities of daily living (ADL: ability to look after yourself), feelings of healthiness (H: what is your state of health?), quality of social support (SS: what support do you receive from others?), and psychological outlook (O: how do you feel about your life?), as shown in Table 4.1. For each domain, respondents select the most appropriate statement from a choice of three that apply to the previous week. Items sum to give an index score of between 0 and 10, where 0 is the worst health and 10 the best quality of life.

Modifications have been made to support the application of the SQL with older people (Stolee et al., 1996; Stadnyk et al., 1998). Terminology was modified to enhance applicability with this population. Domains to address cognition and personal environment were added, and the ADL domain was modified to reflect the needs of geriatric assessment. These modifications may support consideration of this instrument as older people-specific.

There have been three evaluations of the modified SQL. These included various community- and hospital-based populations in Canada (Stadnyk et al., 1998; Simpson, 2002; Carver et al., 1999), as shown in Table 3.2. The results given below are derived from these studies.

Validity

Construct validity: other instruments

Completion by patients with a range of chronic medical conditions demonstrated correlations between SQL and SF-36 domains ranging from 0.45 (O with SF-36 mental health) to 0.60 (ADL with SF-36 physical function) (Stadnyk et al., 1998), as shown in Table 4.3. Correlations between SQL and SF-36 domains ranged from 0.02 (SS with SF-36 social function) to 0.45 (O with SF-36 mental health).

Correlations between the SQL index and a range of instruments were 0.41 (modified-Mini-Mental State Examination), 0.44 (Barthel Index), 0.51 (SF-20 physical function), 0.55 (SF-20 role function), 0.61 (OMFAQ index), and 0.76 (OMFAQ-IADL) (Carver et al., 1999).

Responsiveness

Large ES were found for the SQL (ES 0.73), Barthel Index (BI) (ES 0.68), and OMFAQ-IADL (ES 0.61) following a rehabilitation programme for frail older people with mainly medical conditions (Stadnyk et al., 1998). The SF-36 and NHP had low levels of responsiveness (ES less than 0.35). The authors suggest that the SQL, BI, and OMFAQ reflected the goals of the rehabilitation programme more appropriately.

Participants in a rehabilitation programme post-hip fracture had a slight reduction in SQL score at four weeks following injury (Simpson, 2002). Post-injury scores, the statistical significance of score changes, and differences between the two rehabilitation programmes were not reported.

Table 4.1 Generic patient-reported health instruments

Instrument (no. items)	***Domains (no. items)***	***Response options***	***Score***	***Completion (time in minutes)***
Assessment of Quality of Life instrument (AQoL) (12-15)	Illness *(not in Utility calculation)* (3), Independent living (IL) (3), Physical ability (PA) (3), Psychological well-being (PWB) (3), Social relations (SR) (3)	Categorical: 3 options (0-3) Current health	Summation Domain profile (0-9, 9 worst HRQL) Index (0-45, 45 worst HRQL) Utility (–0.04 to 1.00)	Self (5-7)
COOP Charts for Primary Care Practice (COOP) (8+1)	Bodily pain (BP) (1), Daily activities (ADL) (1), Emotional condition (EC) (1), Physical fitness (PF) (1), Quality of life (QL) (1), Social activities (SA) (1), Social support (SS) (1), Overall health perception (OH) (1), Change in health status (1)	Categorical: 1-5 (illustrated) 2-week recall	Chart profile (1-5, 5 no limitations)	Interview or self
WONCA/COOP (6+1)	Bodily pain (BP) (1), Daily activities (ADL) (1), Emotional condition (EC) (1), Overall health perception (OH) (1), Physical fitness: walking (PF) (1), Social activities (SA) (1), Change in health status (1)	Categorical: 1-5 (illustrated) 2-week recall	Chart profile (1-5, 5 no limitations)	Interview or self Interview (mean 49.0, range: 29-65)
European Quality of Life Questionnaire (EuroQol) (5+1)	*EQ-5D* Anxiety/depression (1), Mobility (1), Pain/discomfort (1), Self-care (1), Usual activities (1) *EQ-thermometer* Global health (1)	*EQ-5D* Categorical: 3 options *EQ-thermometer* VAS Current health	*EQ-5D* Summation: domain profile Utility index (–0.59 to 1.00) *Thermometer* VAS (0-100)	Interview or self
Functional Status Questionnaire (FSQ) (34)	*6 core domains:* Activities of daily living (ADL) (3), Instrumental ADL (IADL) (6), Psychological function (PsychF) (5), Social function (SF) (3), Work performance (WP) (6), Quality of social interaction (SI) (5) *6 single items:* Bed disability days, Reduced usual activities, Social interactions (frequency), Overall health, Sexual relationships (satisfaction), Work status	Categorical: 4, 5, or 6 response options 4-week recall	Summation Domain profile (6 domains 0-100, 100 best function) Plus 6 single items scores	Interview (15)
Goteborg Quality of Life instrument (GQL) (15)	*Part I: GQL instrumen*t Social well-being (4: economy, family, housing, work), Physical well-being (6: appetite, fitness, health, hearing, memory, vision), Mental well-being (5: mood, energy, endurance, self-esteem, sleeping) *Part II: Symptom profile*	Adjectival responses 1-7	Summation Domain profile (1-7, 7 best health) Index (7-105, 105 best HRQL)	Self
Health Status Questionnaire-12 (HSQ-12) (12)	Bodily pain (BP) (1), Energy/fatigue (E) (1), Mental health (MH) (3), Physical functioning (PF) (3), Perceived Health (PH) (1), Role limitation: mental (RM), (1)Role limitation: physical (RP) (1), Social functioning (SF) (1)	Categorical: 3-6 options Recall 4 weeks	Algorithm Domain profile Summary: physical and mental health (0-100, 100 best health)	Interview
Index of Health-related Quality of life (IHQL) (44)	Disability: dependence, dysfunction Discomfort: pain, symptoms Distress: dysphoria, disharmony, fulfilment	Categorical: 5 options	Algorithm 5-level classification across 3 domains Index (0-1, 1 no impairment)	Interview

Nottingham Health Profile (NHP) (38)	Bodily pain (BP) (8), Emotional reactions (ER) (9), Energy (E) (3), Physical mobility (PM) (8), Sleep (S) (5), Social isolation (SI) (5)	Yes/no; positive responses weighted Recall 'general' health	Algorithm Domain profile 0-100, 100 is maximum limitation	Interview Self (10-15)
Quality of Life Index (QLI) (64)	*Satisfaction (S) and Importance (I) of each domain:* Family (S 8, I 8) Health and functioning (S 8, I 8) Psychological / spiritual (S 8, I 8) Social and economic (S 8, I 8)	Likert scale 1-6 for satisfaction, importance	Algorithm: satisfaction score weighted by importance score Domain profile (0-30, 30 best HRQL) Index (0-30)	Self
Quality of Well-being Scale (QWB) (30)	Mobility and confinement (MOB) (3 categories) Physical activity (PAC) (3 categories) Social activity (SAC) (5 categories) Symptoms and medical problems (27)	Categorical: yes/no Recall 6 days Symptoms 8 days	Algorithm Index 0-1, 1 complete well-being	Interview Telephone (mean 17.4, range 6-30)
Quality of Well-being - Self-administered (QWB-SA) (71-74)	Mobility and Physical functioning (11) Self-care (2), Usual activity (3) Symptoms (58): acute physical (25), chronic (18), mental health (11)	Categorical: yes/no Recall 3 days	Algorithm Index 0-1, 1 complete well-being)	Self (mean 14.2)
SF-12: MOS 12-item Short Form Health Survey (12)	Bodily pain (BP) (1), Energy/Vitality (V) (1), General health (GH) (1), Mental health (MH) (2), Physical functioning (PF) (2), Role limitation-emotional (RE) (2), Role limitation-physical (RP) (2), Social functioning (SF) (1)	Categorical: 2-6 options Recall: standard 4 weeks, acute 1 week	Algorithm Domain profile (0-100, 100 best health) Summary: Physical (PCS), Mental (MCS) (mean 50, sd 10)	Interview or self
SF-20: MOS 20-item Short Form Health Survey (20)	Bodily pain (BP) (1), General health (GH) (5) Mental health (MH) (5), Physical functioning (PF) (6) Role functioning (RF), Social functioning (SF) (1)	Categorical: 3-6 options Recall: standard 4 weeks, acute 1 week	Algorithm Summation Domain profile (0-100, 100 best health)	Self (5-7)
SF-36: MOS 36-item Short Form Health Survey (36)	Bodily pain (BP) (2), General health (GH) (5) Mental health (MH) (5), Physical functioning (PF) (10) Role limitation-emotional (RE) (3), Role limitation-physical (RP) (4), Social functioning (SF) (2), Vitality (V) (4)	Categorical: 2-6 options Recall: standard 4 weeks, acute 1 week	Algorithm Domain profile (0-100, 100 best health) Summary: Physical (PCS), Mental (MCS) (mean 50, sd 10)	Interview (mean values 14-15) Self (mean 12.6)
Sickness Impact Profile (136)	Alertness behaviour (AB) (10), Ambulation (A) (12) Body care and movement (BCM) (23), Communication (C) (9) Eating (E) (9), Emotional behaviour (EB) (9) Home management (HM) (10), Mobility (M) (10) Recreation and pastimes (RP) (8), Sleep and rest (SR) (7) Social interaction (SI) (20), Work (W) (9)	Check applicable statements. Items weighted: higher weights indicate increased impairment Recall current health	Algorithm Domain profile (0-100%, 100 worst health); Index (0-100%) Summary: Physical (A, BCM, M), Psychosocial function (AB, C, EB, SI)	Interview (range: 21-33) Telephone: PF only (11.5) Self (19.7)
Spitzer Quality of Life (5)	Activity level (AL) (1), Activities of daily living (ADL) (1), Feelings of healthiness (H) (1), Quality of social support (SS) (1), Psychological outlook (O) (1)	Check applicable statement (3 options, score 0-2) Recall previous week	Summation Index (0-10, 10 best health)	Interview

Table 4.2 Reliability of generic instruments *(references defined in Table 4.3)*

Instrument	*Cronbach's alpha*	*Test-retest correlation [retest period]*
AQoL	utility: 0.73; profile: 0.43 (PS), 0.52 (PW), 0.52 (SR), 0.76 (IL)[1]	-
COOP	*not applicable*	average 0.93, range: 0.78-0.98 [1 hour][2]
EuroQol	*not applicable*	EQ-5D index: 0.67, EQ thermometer: 0.53 [6 months][3]
FSQ	0.42 (quality of social interaction) to 0.90 IADL[4]; 0.74 (psychological function) to 0.91 (IADL, social function)[5]; 0.80 (ADL) & 0.81 (IADL)[6]; 0.63 (ADL) & 0.79 (IADL)[7]	-
GQL	GQL instrument 0.84[8]; symptom profile 0.82[9]	-
NHP	agreement between domains 0.82[10] 0.52 (social isolation) to 0.80 (physical mobility, emotional reactions)[11]	0.81 (social isolation) to 0.97 (physical mobility)[10]
QLI	index: 0.96; domains: 0.79 (family), 0.83 (social, economic), 0.92 (psychological, spiritual), 0.94 (health, functioning)[12]	-
SF-12	MCS 0.81, PCS 0.84[13] 0.94 (physical function [PF] York SF-12) & 0.90 (PF original SF-12); 0.91 (mental health [MH] York SF-12) & 0.88 (MH original SF-12)[14]	-
Revised scoring	MCS 0.80, PCS 0.87[15]; MCS 0.72, PCS 0.81[16]	MCS 0.73, PCS 0.86 [2-4 weeks][15,16]
SF-20	0.76 (physical function) & 0.85 (general health)[17]	0.96[17]
SF-36	0.60 (vitality) to 0.90 (physical function, bodily pain, role-emotional)[18] 0.64 (social function) to 0.91 (physical function)[19] 0.76 (general health) to 0.93 (physical function)[20] 0.74 (general health) to 0.90 (physical function)[21] 0.78 (general health) to 0.93 (physical function)[22] 0.49 (social function) to 0.96 (role-emotional)[23] 0.84 (general health) to 0.92 (social function)[6] 0.67 (general health) to 0.90 (physical function)[24] 0.72 (social function) to 0.91 (physical function) (all domains >0.70)[25] 0.80 (social function) to 0.92 (physical function) (all domains >0.70)[26] 0.56 (social function) to 0.91 (physical function >0.90) (2 domains <0.70, inc. general health: 0.66)[3] 0.73 (social function) to 0.93 (physical function) (all domains >0.70, bodily pain 0.90, 2 domains >0.90)[36] 0.90 (physical function);[7] 0.79 (social function);[28] 0.82 to 0.94[27] 0.82 to 0.96 (for groups 50-70yrs & >70yrs) - domains not specified; 0.94 (physical function in >70yrs)[29] Cognitively impaired: 0.41 (social function) to 0.93 (bodily pain)[30] Cognitively impaired: 0.69 (general health), 0.71 (vitality), 0.93 (physical function, mental health)[31] Cognitively intact: 0.78 (general health), 0.86 (vitality), 0.91 (physical function)[31] 0.50 (social function - telephone administration) to 0.89 (mental health - interview)[32] 0.64 to 0.86 (patient); 0.69 to 0.90 (proxy) - domains not specified[33]	0.65 (role-emotional) to 0.87 (general health), 7 domains >0.70 [1 month][26] 0.28 (social function) to 0.70 (vitality), 4 domains (MH,PF,V,BP) >0.60 [6 months][3] 0.52 (social function) to 0.80 (mental health), 4 domains (MH,PF,RP,V) >0.70 [1 week][34] 0.24 (role-emotional) to 0.80 (bodily pain), 7 domains >0.60, 5 domains >0.70 [median 6 days, range: 2-21 days][25] 0.25 (role-emotional) to 0.97 (physical function), 6 domains >0.70 [1 month])[10] 0.75 to 0.96 (patient), 0.75 to 0.94 (proxy); domains not specified[35] Summary Scores MCS 0.79, PCS 0.82 [1 week][34]
SIP	0.59 (sleep and rest, eating) and 0.95 (SIP index)[36] 0.60 (sleep and rest, eating) and 0.84 (body care and movement);[37] 0.84 (mobility)[38]	-

Table 4.3 Validity of generic instruments

Instrument	*Socio-demographic variables and health-service use*	*Patient-reported health instruments*
AQoL	Lower scores predictive of increased health-care at 18 months[1]	*Utility* [h] with SF-36: range 0.19 (bodily pain) to 0.62 (physical function); with SF-36 summary scores: MCS 0.41, PCS 0.37 [h] with OMFAQ: range –0.20 (social resources) to –0.68 (self-care)[1] *Domains* [h] with SF-36: range 0.04 (PA with bodily pain, role-physical) to 0.64 (IL with physical function) [h] with OMFAQ: range 0.03 (IL with self-care) to –0.82 (IL with self-care)[1]
COOP	Chronic illness: strongest correlation with COOP PF and DA[2] COOP EC predictive of nursing care, OH predictive of future hospitalisation Change scores not associated with placement in skilled care[39] General population *vs* post-hip fracture (PF and DA charts)* [40]	[h] with RAND scales: range 0.01 (PF with emotional status [ES]) to 0.69 (ES with ES)[2] EC with depression scales: range 0.70 to 0.74[41] BP with pain scales:, range 0.75 to 0.89[42] [h] with NHP: range 0.09 (OH with Sleep) to 0.75 (DA with physical mobility)[11] [h] with Rehabilitation Activities Profile: range 0.08 (OH with relationship) to 0.79 (DA with mobility and personal care)[11] [h] with Barthel Index: range 0.18 (BP) to 0.75 (DA)[11]
WONCA/ COOP (W/C)	-	range 0.51 (PF-no illustration with Activity Scale) to 0.71 (EC with or without illustration) with SF-20 mental health)[43] [h] with EQ-5D index: range –0.35 (change in health) to 0.59 (DA)[44] [h] with EQ-thermometer: range –0.29 (PF, SA) to –0.65 (OH)[44] [h] with EQ-5D items: range 0.13 (W/C SA with EQ pain) to 0.74 (W/C P with EQ pain)[44]
EuroQoL	*EQ-5D index and EQ thermometer* Community-dwelling: number of GP visits,* hospital inpatient stay,** limiting long-term illness,*** high/low levels of disability*[3] A&E attendance *n/ss*; age-group *n/ss*;[3] age* [45] *EQ-5D index only* GP visit previous 3 months* [44] Post-hip fracture: levels of pain, healing, mobility, self-care at 4 and 17 months* [45] Initial fracture severity at 12 and 24 months* [46] Healed undisplaced *vs* healed displaced fractures* [46] General population *vs* patients post-acetabular fracture[47] *EQ thermometer only* Disability level* [44]	*EQ-5D index* [h] with WONCA/COOP charts: range –0.35 (change in health) to 0.59 (daily activities)[44] *EQ thermometer* [h] with WONCA/COOP charts: range –0.29 (physical fitness; social activities) to –0.65 (overall health)[45] with Barthel Index* [44] *EQ-5D items* [h] with WONCA/COOP: range 0.13 (P/D with social activities) to 0.74 (P/D with body pain)[44] [h] with Barthel Index (BI): items M, SC and EQ-index with BI mobility domain;* item SC with BI-dressing and BI index* [44]
FSQ	Age *n/ss*[4] With mortality rates: range –0.06 (PsychF) to –0.25 (SF)[5]	with Physical Performance Test (PPT) and OMFAQ: range 0.45 (IADL with PPT) to 0.70 (ADL with OMFAQ-IADL)[6] [h] FSQ IADL with FSQ ADL 0.73[6] with SF-36: range 0.33 (ADL with role emotional) to 0.76 (IADL with physical function)[6]

		[h] Cardiac patients - IADL with cardiac symptoms: range 0.48 (angina) to 0.72 (specific activity)[5] [h] Psychological function (PsychF) with symptoms: range 0.32 (angina) to 0.73 (specific activity)[5] [h] IADL with SF-12: MCS 0.21, PCS 0.75; PsychF with SF-12 MCS 0.82, PCS 0.32[5]
GQL	-	with Hearing Coping Assessment 0.34, Life Orientation Test –0.50, Beck Depression Inventory 0.61[9]
HSQ-12	Age (not MH, SF),* long-standing illness** [48] Self-reported health problems, depression, limited ADL*** [49] Psychiatric problems (HP,MH,RM);* impaired hearing: (PF,RP,SF,E);* impaired vision (not RM);* dementia (not MH, BP)* [49] Health-service use (not RM)* [49]	[h] HSQ-12 domains range 0.19 (PF with RM) to 0.72 (PF with RP)[49]
IHQL	-	[h] with SHORT-CARE (SC): range 0.08 (discomfort with SC depression) to 0.14 (disability with SC somatic symptoms)[50]
NHP	Social class, living status, age (<70 yrs *vs* >70 yrs) *n/ss* [51] [h] GP consultation* [51] Mental status and sex (SI)* [51] [h]Anxiety and depression (SI, E)[10] Sex (group with low back pain [LBP]): E,S,SI* [52] Age (group with osteoarthritis [OA] of hip): PM,E* [52] OA hip *vs* LBP;* OA hip or LBP *vs* fit elderly* [52] [h] Levels of fitness, well-being, chronic illness* [51] Fit elderly with self-reported musculoskeletal problems: RM,BP* [52] [h] Various domains: possible *vs* probable morbidity [10] Chronic obstructive airways disease *vs* general population (PM,E,ER,SI)* [53] General population *vs* post-hip fracture (PM,SI,S,E)* [40]	[h] with Physical Performance Score 0.53 (E), 0.54 (P), 0.70 (PM)[10] [h] with Katz ADL scale 0.51 (P), 0.58 (E), 0.74 (PM)[10] [h] with SF-36: range 0.02 (ER with role physical) to –0.76 (PM with physical function)[25] with SF-36 range 0.00 (BP with general health; S with role physical) to –0.80 (E with vitality)[53] [h] with WONCA/COOP: range 0.01 (S with overall health) to 0.75 (PM with daily activities)[11] [h] with Rehabilitation Activities Profile (RAP) range 0.01 (S with relationships) to 0.79 (PM with mobility and personal care)[11] [h] with Barthel Index range 0.04 (S) to 0.79 (PM)[11]
QLI	Age *n/ss*[12] Better health, social support, hospital readmission 51% of score variance for higher QLI; longer hospitalisation 48% of score variation for lower QLI[12]	-
QWB	Age *n/ss*[54] Assistance walking,** prehensile tasks,* depression* [55]	[h] with SIP summary and index –0.37 (psychosocial summary), –0.49 (physical summary), –0.52 (index)[54] [h] with SF-36 three domains 0.36 (general health), 0.37 (role limitation), 0.39 (physical function)[54] [h] with Chronic Disease Index –0.09, Stress Scale –0.18[54] with functional activities (time): range –0.20 to –0.45[55]
QWB-SA	Sex *n/ss*[56] Self-reported health,*** Bed days –0.25, Restricted activity days –0.34[56]	[h] with SIP work –0.11, eating –0.12, psychosocial –0.40, physical 0.42 [56] [h] with SF-36 role-emotional 0.17, MCS 0.22, PCS 0.47, physical function 0.51 [56]
SF-12	Age: domain scores,*** summary scores *n/ss*[57] Age: PCS decreased, MCS constant[13] Sex: better health in men* [13] Depression,*** impaired vision,*** self-reported health,**	MCS explains greater variation in SHORT-CARE depression[58] PCS explains greater variation in ADL limitation[49] PCS with MSC 0.08[49]

	ADL limitation,*** Use of health or social services* [49] Length of hospital stay, emergency care, health status in older people *vs* general population* [13] Self-reported physical health: RP most discriminative* [57] Psychological distress (GHQ-12): MH most discriminative* [57] MCS: self-reported psychiatric problems*** [49] PCS: dementia,* impaired hearing** [49] Number of chronic illnesses: PCS –0.41, MCS –0.44 [15,16] Exercise* [15,16]	
SF-20	Age: general population *vs* older people* (not MH)[17] Sex *n/ss* (MH only*) [17] GH predictive of hospitalisation and MH of skilled nursing care placement; change scores not associated with placement in skilled care [39, 60]	MH with WONCA/COOP charts: range –0.05 (physical fitness) to –0.71 (emotional condition)[43] [h] PF with Spitzer Quality of Life (SQL) 0.51, Barthel Index 0.63, OMFAQ index 0.65, OMFAQ-IADL 0.67 [17] [h] RF with Barthel Index 0.48, SQL 0.55, OMFAQ-IADL 0.56, OMFAQ index 0.59 [17] [h] PF with modified mini-Mental State Examination (m-MMSE) 0.20, RF with m-MMSE 0.27 [17]
SF-36	*Age* Health decline with age: PF;[3,18,21,28,57,60,61] Vitality;[3,60] RP[28] Health decline with age: all domains except MH[62] Advancing age: MH constant;[28,62] MH improved[57,62,63] Young-old (better health) *vs* general population* (not PF)[21,63] General older population (across age-groups)* (not GH or SF)[60,64] Patients diagnosed with moderate depression: MCS higher in group aged >70 yrs than in group aged <70 yrs* [64] Post-operative women: SF,RE,MH higher (PF lower) in group aged >65 yrs than in group aged <65 yrs [63,65] Haemodialysis patients (PCS,PF,RP,BP,GH domains) and renal transplant patients (PCS, 6 domains excl. RP,V): higher scores in group aged >65 yrs than in younger group* [66] *Sex* All domains: worse health in women* [28,61,64] GH: worse health in women* [67] PF,SF,MH: worse health in women* [27] 1 yr post-stroke RE,ME,SF: better health in women* [18] *Health status* All domains: diagnosed health problems,* [18,63,64,68,69,70] long-standing disability* [3,7] Post-stroke PF,* MH* [18] Cancer (not SF,BP,MH,MCS)* [71] Chronic obstructive airways disease: older *vs* general population: PM,E,ER,SI* [53] Dyspnoea MCS, PCS*** [69] Parkinson's disease severity: PF,RP,SF,GH* [72]	*Physical function: performance tests* [h] PF with Physical Performance Test 0.59, National Institute on Aging Battery 0.65[7] [h] Guralnik Performance Test: range 0.34 (RP) to 0.74 (PF);[10] Timed Up and Go test –0.26 [79] [h]all domains (except MH, RE) discriminate between performance-based assessments* [7] *ADL (patient-reported)* PF with Katz ADL 0.30, OMFAQ-IADL 0.36[6] [h] PF with modified Katz ADL: range 0.30 (RP) to 0.79 (PF)[10] PF with FSQ: ADL 0.51, IADL 0.76[6] [h] PF with ADL 0.56, IADL 0.78[7] [h] with ADL assessment: range –0.37 (PF) to –0.43 (RP)[34] [h] with Physical activity instruments: range MH, BP (0.17 to 0.28) to PF, GH (0.26 to 0.42)[80] [h] PF with Functional Independence Measure (FIM): cognitively intact 0.53, impaired 0.33 (PF scores discriminate between groups defined by cognitive status);* other correlations smaller than hypothesised (range –0.029 RE to 0.22 MH)[30] [h]RP,PF,PCS with Functional Disability Index: range –0.56 (RP) to 0.77 (PF)[75] [h] with WOMAC: range 0.15 (MH with WOMAC stiffness) to –0.55 (BP with WOMAC pain); –0.50 (PF with WOMAC physical function); convergent correlations smaller than hypothesised[73] [h] PF with other domains: range 0.07 (MH) to 0.48 (RP) (V 0.34, SF 0.34)[25] SF-36 domains (RP,PF,PCS): range 0.68 RP with PF, to 0.86 PF with PCS[75] [h] Patient *vs* proxy completion, PF with FIM: patients 0.53, professionals 0.66, lay proxy 0.38[81] *Mental health* [h] with Geriatric Depression Scale: range PF –0.25 to MH –0.71[34] with m-MSE <0.18 [25,34] [h] with Hamilton Rating Scale-Depression (HAMD): range PF –0.12 to MH and MCS –0.57 (SF –0.45, RE –0.43)[19] [h] with Clinician's global impression of depression (severity, improvement): range PCS –0.07, PF –0.08

	Post-knee replacement co-morbidity and self-reported health: all domains* [73] Trauma survivors *vs* uninjured population (not BP)[61] Heart disease* [63] Chronic heart failure* [70] Fear of falling: RP,SF (males & females);* PF,GH,V (females only)* [74] Fracture in previous 10 yrs: PF, RP, PCS[75] [h] Service-users – incontinence: PF,SF,GH* mental health: MH* [76] *Health-service use* GP visits;* [28] not PF,MH;[3,27] not RE[27] Outpatient visits* [27] not PF,V[3] A&E visits: not RE,MH,V,GH* [3] Hospital inpatient* not MH[3,27] not RE,V[3] Hospital outpatients (worst health) *vs* GP patients: PF, RP* [60] Low-care- *vs* high-care dependent older people: all domains (PF*) (not BP)[77] Community (best GH) *vs* sheltered housing or institutions: GH* [67] Living alone (worst health) *vs* living with others: all domains except GH* [28] *Predictors in chronically ill patients* Mortality at 4 yrs: 7 domains - mostly GH,PF not MH Hospitalisation at 2 yrs: PF,RP,BP not MH,RE GP visit or inpatient status at 2 yrs: BP,GH,V not MH,RE[78]	to MH and MCS –0.53[19] [h] HAMD with MH –0.20, RP –0.26, PF –0.32; Quality of Life Depression Scale with MCS –0.69, PCS –0.14[64] [h] MH with other domains: range PF 0.07 to V 0.50 (GH 0.46)[25] [h] MH, SF discriminate between groups defined by anxiety or depression, possible or probable morbidity[10] [h] SF-36 and range of self-reported morbidity or symptoms: expected correlations reported[10] with Sense of Coherence scale - elders with chronic heart failure: range 0.10 (BP) to 0.46 (MH); healthy controls 0.00 (BP) to 0.39 (RE)[70] *Quality of life* [h] with Assessment of Quality of Life (AQoL): range 0.04 (BP and RP with AQoL physical senses) to 0.64 (PF with AQoL independent living)[1] with AQoL utility: MCS 0.41, PCS 0.37[1] [h] Feeling Thermometer with MCS 0.30, PCS 0.50[63] [h] with NHP: range –0.02 (RP with NHP emotional reaction) to –0.76 (PF with NHP physical mobility)[25] with NHP: range 0.00 (GH with NHP pain) to –0.80 (V with NHP energy)[53] [h] PF with Barthel Index, NHP mobility, OMFAQ-IADL, Spitzer Quality of Life Index - all >0.60[25] [h] with SIP: range 0.67 (SF) to 0.78 (PF)[82] with Functional Assessment of Cancer General Scale (FACT-G): range 0.02 (RP with FACT-G relationship with doctor) to 0.61 (SF with FACT-G functional well-being); MCS and PCS 0.52[71] [h] with Functional Profile Inventory (FPI): range –0.03 (SF with spiritual activity) to –0.69 (PF with FPI total and physical exercise)[83]
SIP	Severe *vs* non-severe co-morbidity: before heart surgery PsychF;* following heart surgery PhysF* [84] [h] Older nursing-home residents (greater impairment) *vs* older people with chronic airways disease (domain, summary)*** [37] Following intensive care: old-old (greater impairment) *vs* young-old and middle-aged (BCM,M,A,SR,HM)* [12] SIP-68 M: sensitivity for poor function high (91%), specificity for good function low (58%) (external criterion Guralnik Performance Test)[38] SIP-68 M: recurrent fallers *vs* non-fallers,* predictive of risk of recurrent falling [38]	*Summary* [h] with SF-36 –0.21 (PsychF with GH) to –0.40 (PhysF with PF), Quality of Well-being Scale (QWB) –0.37 (PsychF) to –0.49 (PhysF);[54] [h] Summary scores 0.67;[37] [h] PhysF with Barthel Index (BI) 0.74, Index of ADL 0.74, Life Satisfaction Index (LSI) -0.16[37] Bed days 0.14, Restricted activity days (RAD) 0.12[56] [h] Functional Profile Inventory (FP): range –0.13 (spiritual activity) to –0.64 (body care)[83] [h] PsychF with LSI –0.31, PGC Morale Scale –0.40, BI 0.47[37] with Bed days 0.22, RAD 0.22[56] [h] with FPI range –0.06 (spiritual activity) to –0.53 (body care)[83] *Index* [h] with Chronic Disease Index 0.14, Positive Effect 0.31[54]; with Bed days 0.28, RAD 0.24[56] [h] with QWB –0.52[54] [h] with SF-36: range –0.41 (MH) to –0.85 (PF)[56] [h] 0.67 (SF) to 0.78 (PF)[82] with three SF-36 domains: –0.33 (GH) to –0.47 (PF)[54] [h] with FPI: range –0.08 (spiritual activity) to –0.62 (body care)[83]

		Domains [h] with SF-36: range 0.02 (W with social function) to –0.86 (PhysF with physical function)[54]
SQL	-	[h]with modified-Mini-Mental State Examination 0.41, Barthel Index 0.44, SF-20 physical function 0.51, SF-20 role function 0.55, OMFAQ-ADL 0.61, OMFAQ-IADL 0.76[17] [h] modified SQL with SF-36: range 0.45 (O with mental health) to 0.60 (ADL with physical function)[25] [n] modified SQL SS with SF-36 social function 0.02[25]

Key:
[h] = hypothesis supported by correlation
levels of statistical significance: * = p <0.05; ** = p <0.01; *** = p <0.001
n/ss = not statistically signigicant

References

1 Osborne et al. (2003)
2 Nelson et al. (1990)
3 Brazier et al. (1996)
4 Yarnold et al. (1995)
5 Cleary and Jette (2000)
6 Reuben et al. (1995)
7 Sherman and Reuben (1998)
8 Nyrgren et al. (2001)
9 Andersson et al. (1995)
10 Sharples et al. (2000)
11 Van Balen et al. (2001)
12 Kleinpell and Ferrans (2002)
13 Lim and Fisher (1999)
14 Iglesias et al. (2001)
15 Resnick and Nahm (2001)
16 Resnick and Palmer (2001)
17 Carver et al. (1999)
18 Anderson et al. (1996)
19 Beusterien et al. (1996)
20 Wood Dauphinee et al. (1997)
21 Dexter et al. (1996)
22 McHorney et al. (1994a)
23 Murray et al. (1998)
24 Wolinsky et al. (1998)
25 Stadnyk et al. (1998)
26 Andresen et al. (1996)
27 Lyons et al. (1994)
28 Walters et al. (2001)
29 Mangione et al. (1993)
30 Seymour et al. (2001)
31 McHorney et al. (1990)
32 Weinberger et al. (1994)
33 Pierre et al. (1998)
34 Andresen et al. (1999)
35 Yip et al. (2001)
36 Andresen et al. (1998a)
37 Jannick-Nijlant et al. (1999)
39 Siu et al. (1993a)
40 Van Balen et al. (2003)
41 Doetch et al. (1994)
42 Manz et al. (2000)
43 Kempen et al. (1997)
44 Coast et al. (1998)
45 Tidermark et al. (2002a)
46 Tidermark et al. (2002b)
47 Tidermark et al. (2003)
48 Bowling and Windsor (1997)
49 Petitt et al. (2001)
50 Livingstone et al. (1998)
51 Hunt et al. (1980)
52 Thorsen et al. (1995)
53 Crockett et al. (1996)
54 Andresen et al. (1995)
55 DeBon et al. (1995)
56 Andresen et al. (1998b)
57 Schofield and Mishra (1998)
58 Gurland et al. (1984)
59 Siu et al. (1993b)
60 Hayes et al. (1995)
61 Inaba et al. (2003)
62 McHorney et al. (1994b)
63 Baldassarre et al. (2002)
64 Doraiswamy et al. (2002)
65 Girotto et al. (2003)
66 Rebello et al. (2001)
67 Heslin et al. (2001)
68 Jenkinson et al. (1995)
69 Ho et al. (2001)
70 Ekman et al. (2002)
71 Overcash et al. (2001)
72 Hobson and Meara (1997)
73 Bombardier et al. (1995)
74 Suzuki et al. (2003
75 Wildner et al. (2002)
76 Hill et al. (1996)
77 Murray et al. (1998)
78 McHorney (1996)
79 Jaglal et al. (2000)
80 Harada et al. (2001)
81 Ball et al. (2001)
82 Weinberger et al. (1991)
83 Larson et al. (1998)
84 Page et al. (1995)

Chapter 5: INSTRUMENT REVIEWS - OLDER PEOPLE-SPECIFIC INSTRUMENTS

a) Brief Screening Questionnaire (Smeeth et al., 2001)

The Brief Screening Questionnaire (BSQ) was proposed as a primary care screening assessment for individuals aged 75 years and over (Smeeth et al., 2001). The questionnaire was devised in response to recommendations from the NSF-OP (DH, 2001) that older people should receive a single annual assessment appropriate to their individual circumstances.

Items were selected to represent health-related issues specified in the DH General Practice document (General Practice in the National Health Service. A new contract. DH, 1989). The BSQ contains 26 screening items addressing issues related to activities of daily living (ADL), cognitive impairment, economic status, mental health, social contact, symptoms, visual impairment, and hearing impairment (see Table 5.1). Three additional items about smoking, alcoholic intake, and physical activity were included as recommended by the NSF-OP (DH, 2001).

Items use dichotomous or categorical response options. Where appropriate, respondents indicate the frequency of a problem. Scoring is not given, but appears to be based on summation. The instrument may be self-completed or interview-administered.

There has been one evaluation of the BSQ. This included a community-based older population in the UK (Smeeth et al., 2001), as shown in Table 3.4.

Validity

Validity: other

Participants were randomised to self-complete the instrument, or receive interview administration by lay interviewer or nurse. High specificity (greater than 90% for all domains) across all forms of administration supported the diagnostic accuracy of the BSQ. However, the low sensitivity (less than 50% for all domains) suggests the BSQ should be used with caution when screening for poor vision, hearing impairment, depression, and cognitive problems.

Acceptability

High overall completion rates were reported (78% response rate). Postal self-completion showed the highest response rates (83.5%), compared with nurse-led interview (75.9%) and lay-person interview (73.9%). Some 21% of postal respondents reported needing help to complete the questionnaire. The proportion of missing or invalid responses was higher for postal self-completion (mean 4.1%) than for interview administration (compare 1.1% for nurse-led interview and 0.6% for lay-person interview). The oldest age-groups had lower response rates with postal self-completion. Men were more likely to respond than women (80.5% versus 76.7%), and responders were slightly younger than non-responders (80.3 versus 81.0 years) (both p less than 0.001).

Self-completion yielded more self-reported problems (22 of 26 items) than interview administration; nurse-led interviews yielded fewer self-reported problems than lay-person interviews.

b) Comprehensive Assessment and Referral Evaluation (Gurland et al., 1977; revised 1984)

The Comprehensive Assessment and Referral Evaluation (CARE) was developed in the UK and the USA for evaluating health and social concerns in older people (aged 65 years and over) (Gurland et al., 1977,1984). The instrument addresses medical, psychiatric, nutritional, economic, and social issues, and is recommended by the authors for use with both patients and the older general population.

Instrument content was derived from existing instruments. The original version comprises 1500 items and requires administration by a trained interviewer (Gurland et al., 1977) (Table 5.1). The schedule includes interviewee self-report and test items, where the interviewee is invited to recall facts, or carry out a simple calculation or activity. Interviewer-rated items include observation and global items, based on a review of information gathered during the interview. Most items use a categorical response scale, often with two- or three-point options. Completion of the CARE gives a global overview and narrative summary of an individual's health status (Gurland et al., 1977). Alternatively, domain scores may be calculated. Although not clear in the original publication, this appears to be based on item summation, where higher scores indicate worse health.

Shortened versions of the instrument have been developed. The CORE-CARE (Golden et al., 1984) and SHORT-CARE (Gurland et al., 1984) comprise 329 and 143 items, respectively, across six domains, namely depression, dementia, disability (activity limitation), memory, sleep, and somatic symptoms (see Table 5.1). Narrative summaries, global judgements, and domain or scale scores are produced.

Within the CORE-CARE four summary scores, namely psychiatric (48 items), medical/physical/nutritional (191 items), environmental/social problems (75 items), and service needs (15 items) summary scores, were defined from 22 'indicator scales' (total 329 items) (Golden et al., 1984). The indicator scales comprise the following domains: cognitive impairment (CI: ten items), depression (DP: 29 items), subjective memory problems (SM: nine items), somatic symptoms (SS: 34 items), heart disorder (HT: 15 items), stroke effects (SE: nine items), cancer (CA: six items), respiratory symptoms (RS: six items), arthritis (AR: nine items), leg problems (LP: nine items), sleep disorder (SL: eight items), hearing disorder (HE: 14 items), vision disorder (VD: 11 items), hypertension (HP: four items), ambulation problems (AM: 27 items), activity limitation (AL: 39 items), financial hardship (FH: eight items), dissatisfaction with neighbourhood (ND: eight items), fear of crime (FC: 18 items), social isolation (SI: 34 items), retirement dissatisfaction (RD: seven items), and service utility (SU: 15 items). These items were selected using psychometric analysis and following expert opinion (Golden et al., 1984; McDowell and Newell, 1996).

The SHORT-CARE has additional diagnostic scales for depression, dementia, and disability (Gurland et al., 1984; McDowell and Newell, 1996). Although highly sensitive for the diagnosis of depression (84%) and dementia (91%), in 'non-case' community-dwelling older people, specificity is low (35% for depression and 30% for dementia), which suggests that the scales should be used with caution (Gurland et al., 1984). A misclassification rate of 16% when the scales are used separately is reduced to 2% when the scales are used together.

The 30-item CARE 2000 (Gurland, BJ et al.) is listed on the Quality of Life Instruments Database (www.qolid.org/public/CARE.html) and is designed for the comprehensive assessment of function and other aspects of quality of life in older people. An evaluation of this instrument was not found through electronic searches.

Five articles describe the original development and evaluation of the CARE (Gurland et al., 1977), and subsequent development and evaluation of the CORE-CARE (Golden et al., 1984; Teresi et al., 1984a,b) and SHORT-CARE (Gurland et al., 1984a). All these articles describe the same community-based older populations from the USA and UK, as shown in Table 3.4. A further article describes an evaluation of the SHORT-CARE in a UK population (Petit et al., 2001). The results given below are derived from these articles.

Reliability

Evidence for internal consistency reliability is shown in Table 5.2. There is no published evidence for test-retest reliability. Internal reliability for the SHORT-CARE diagnostic scales ranged from 0.64 (dementia) to 0.84 (disability) (Gurland et al., 1984a). Similarly high levels of internal reliability have been reported for the CORE-CARE indicator scales (range 0.28 to 0.95) (Golden et al., 1984; Teresi et al., 1984b).

Due to low inter-rater reliability, a training manual and more explicit rules for administration were produced (Gurland et al., 1977). High inter-rater reliability was reported following the simultaneous rating of 30 interviewees (indicator scales range from 0.70 to 0.80) (Golden et al., 1984). High inter-rater reliability was also reported for the SHORT-CARE diagnostic items ranging from 0.78 (disability) (Teresi et al., 1984a) to 0.94 (depression) (Gurland et al., 1984).

Validity

(i) Construct validity: socio-demographic variables and health-service use

The validity of the CARE activity limitation (AL) and cognitive impairment (CI) domains, and their utility as screening tools, was assessed using measures of family inconvenience and decision to institutionalise older relatives as criterion variables (Teresi et al., 1984a). As hypothesised, a low AL score correctly predicted 84% of those families reporting that they were inconvenienced and 58% of those not inconvenienced. Similarly, a high score correctly predicted 32% of those reporting inconvenience and 99% of those not inconvenienced; a medium score predicted 60% of those reporting inconvenience and 93% of those not inconvenienced. Furthermore, a low AL score was associated with 77% of families planning institutional care and 57% of those not planning care. A high score correctly predicted 23% of those planning care and 98% of those not planning care; a medium score predicted 43% of those planning care and 90% of those not planning care.

Low CI scores correctly predicted 60% of those families reporting inconvenience and 69% of those not inconvenienced. A high score correctly predicted 40% of families reporting inconvenience and 100% reporting no inconvenience. A low CI score was associated with 25% of families reporting inconvenience and 93% of those not inconvenienced.

(ii) Construct validity: other instruments
Following detailed interviews with the sample population, psychiatrists and social scientists completed diagnostic and global ratings (Global Diagnostic Rating [GDR]). Family members of participants were also interviewed (Family Informant Scale [FIS]) (Teresi et al., 1984b). Correlation between the CARE psychiatric domain and instrument domains that had hypothesised associations ranged from 0.33 (FIS depression with CARE depression) to 0.75 (GDR depression with CARE depression), as shown in Table 5.3. Correlations between domains that did not have hypothesised associations ranged from 0.01 (between several domains including FIS cognitive impairment with CARE depression) to 0.41 (FIS age with CARE cognitive impairment). A larger than hypothesised correlation was reported between CARE sleep and depression items (0.55).

Correlation between CARE (medical conditions) and the FIS that had hypothesised associations ranged from 0.45 (arthritis; hypertension) to 0.70 (activities of daily living). With the exception of the correlation between CARE activities of daily living and FIS total somatic complaints (0.59), correlations between domains that did not have hypothesised associations were less than 0.29.

A range of correlations between the CARE domains and the GDR and FIS domains were in accordance with hypotheses, as shown in Table 5.3. Correlation between the CARE (medical conditions) and FIS ranged from 0.45 (arthritis, hypertension) to 0.70 (ADL). Correlations between the CARE (service needs) and FIS and GDR ranged from 0.62 (FIS ambulation domains) to 0.70 (GDR and FIS activity limitation domains). Correlation between the CARE (social needs) and FIS social isolation was 0.41, and with the GDR ranged from 0.61 (crime) to 0.64 (finances).

Correlation between the CARE indicator scales and the GDR ranged from 0.40 (CARE service needs: total service utilization) to 0.75 (psychiatric domain: depression). Correlations between the CARE cognitive impairment and activity limitation domains and the GDR were 0.71 and 0.70, respectively. Correlations between CARE indicator scales and the FIS ranged from 0.30 (service needs: family service provision) to 0.70 (service needs: activity limitation). Correlations between the FIS and several CARE domains ranged from 0.33 (depression) to 0.41 (social isolation).

(iii) Validity: other
A range of correlations between the CORE-CARE indicator scales were in accordance with hypotheses (Golden et al., 1984) (see Table 5.3). A correlation of 0.78 was found between activity limitation (AL) and ambulation (AM); correlations of 0.58 and 0.60 were found between total service utilisation (TSU) and AL and AM, respectively. A correlation of 0.40 was found between arthritis (AR) and AM scales. Correlations of 0.54 and 0.51 were found between AM and somatic symptoms (SS) and respiratory symptoms (RS), respectively. Depression had correlations of 0.50 with AL, 0.54 with SS, and 0.55 with sleep disorder (SL). Correlations of less than 0.50 were found between several scales, namely cognitive impairment with RS, subjective memory problems with hypertension, AR with hearing disorder, and TSU with dissatisfaction with neighbourhood.

Using the GDR and FIS as external criteria, the predictive validity of the CORE-CARE at one year was assessed (Teresi et al., 1984a). Based upon the presence of a condition

at baseline, correct classifications for the prediction of conditions at one year ranged from 77% (arthritis and ambulation problems, sleep disorder) to 98% (cancer). The odds ratios between CORE-CARE reported items and death at one year ranged from 0.40 (hearing problems and dissatisfaction with the neighbourhood) to 3.1 (heart problems). The most important predictors for death at one year in community-dwelling people were cognitive and functional impairment, older age, and male sex. Activity limitation, cognitive impairment, and age were the strongest predictors for service utilization.

Clinician diagnosis was compared with SHORT-CARE depression and dementia scales in a sample of 26 people (Gurland et al., 1984). There was agreement for ten individuals where a depressive disorder was not reported. However, whereas the SHORT-CARE identified psychiatric problems in 16 individuals, clinician diagnosis identified 12 individuals. Correlation between clinician diagnosis and independent informant reports were 0.33 (pervasive depression), 0.66 (personal time dependency), and 0.69 (pervasive dementia). In additional, individuals diagnosed with pervasive dementia using the SHORT-CARE scales had observed outcomes at one year consistent with dementia.

Responsiveness

Following completion of the SHORT-CARE and HSQ-12 at baseline and 18 months, regression analysis of change scores indicated that change in the SHORT-CARE activities of daily life (ADL) domain was predicted by the baseline ADL score and by change in HSQ-12 physical function, role physical, and social function scores, explaining 56% of the variance in change score (Petit et al., 2001). Change in the SHORT-CARE depression domain score was predicted by the baseline depression score and change in HSQ-12 mental health and role mental scores, explaining 41% of the variance in change score.

Acceptability

There is some flexibility in the CARE administration schedule, but approximately 1.5 hours is required for the original instrument (range: 45 minutes to 2.5 hours) (Gurland et al., 1977; McDowell and Newell, 1996). Although the developers suggest that the interview schedule is suitable for use in household surveys, community health-service settings, and clinical research, the burden on interviewer and respondent should be considered.

Both the CORE-CARE and SHORT-CARE also require administration by a trained interviewer, but are less time-consuming; the SHORT-CARE takes approximately 30 minutes to complete (Gurland et al., 1984).

c) Epic/Elderly Assessment System (EASY-Care) (Philp, 1997; modified 2000)

The Epic, revised to Elderly Assessment System (EASY-Care) was developed across Europe, including the UK, during the 1990s to provide a holistic and standardised approach to comprehensive geriatric assessment (CGA) (Philp, 1997; 2000). It was developed for use in primary and community health-care settings, and is recommended for use as an annual medical and social assessment procedure for older people (aged over 75 years) (Bath et al., 2000).

Development was supported by a grant from the European Regional Office of the World Health Organisation (WHO), involving cross-cultural adaptation and testing across several European countries. Instrument content was derived from several existing questionnaires, including the WHO-11 Countries Survey Instrument (Heikkinen et al., 1983), OMFAQ IADL scales (Fillenbaum, 1988), Barthel Index (Mahoney and Barthel, 1965), and the SF-36 (Ware et al., 1997). Additional items were derived from the consensus agreement of experts experienced in the health- and social care of older people.

The broad domains cover physical, mental, and social functioning; the original EASY-Care comprises activities of daily living and instrumental activities of daily living (ADL and IADL), cognitive impairment, continence, depression, economic status, global health, hearing, loneliness, mobility, and vision in what is described by the developers as a comprehensive geriatric assessment (Philp, 1997; 2000). 24 core items assess IADL and ADL. However, there is inconsistency regarding the number of items, reported variously in published articles as 24 (Philp, 1997; Bath et al., 2000), 26 (Philp et al., 2002), and 38 (Philp, 2000). The original instrument was criticised for its lack of items addressing social activities, social support systems, and sleep (Richardson, 2001).

The new English version of the EASY-Care (EASY-Care Information Sheet, 2003) comprises 85 items across six domains (see Table 5.1), namely general health (19 items), physical abilities (disability) (17 items), memory (six items from the Cognitive Impairment Test), home, safety, and support (14 items), health-care services received (22 items), and looking after your health (seven items). Additional information about perceived needs, goal-setting, and satisfaction with care may also be gathered.

The general health domain includes depression (four items from the Geriatric Depression Scale, one from WHO-11) and single items addressing a range of issues such as chewing, hearing, loneliness, and vision (from WHO-11), global health (from the SF-36), and communication. Physical abilities comprises IADL (six items from the OMFAQ IADL scale) and ADL (11 items, including nine from the Barthel Index). The home, safety, and support domain comprises items related to accommodation (7), access to public services (1), family and friends (2), finance (1), and safety (3).

Items have categorical response options, which sum to produce domain scores. Scores for the disability domain are weighted and produce a score from 0 to 100, where 100 is best health (Philp, 2000).

The developers indicate that instrument modification ensured the incorporation of domains considered important in the assessment of older people and defined by the NSF-OP (EASY-Care Information Sheet, 2003; EASY-Care Training Pack, 2003). The

most recent version contributes to the Single Assessment Process for older people and includes contact and overview assessments; it is intended to serve as a foundation for a more detailed specialist or comprehensive assessment, if required (The Single Assessment Process and EASY-Care as the Contact and Overview Assessment tool, 2003: Appendix 2, p26). A detailed training pack is available from the Sheffield Institute for Studies on Ageing, UK.

There have been three evaluations of the EASY-Care in different community-based older populations in the UK (Philp et al., 2001, 2002; Bath et al., 2000), as shown in Table 3.4. The results given below are derived from these studies.

Reliability
The results of test-retest reliability are shown in Table 5.2. There is no evidence for internal consistency reliability.

High levels of reliability were found for the disability score (0.87) and seven single items (>0.70) following a two-week retest completion by patients in a day rehabilitation unit (Philp et al., 2002): see Table 5.2. A trained nurse assisted completion. Remaining items had lower levels of reliability (<0.40 for communication, telephone, feeding, and cognitive impairment).

Validity
Construct validity: socio-demographic variables and health-service use
The EASY-Care discriminated between groups defined by levels of deprivation (Bath et al., 2000). As hypothesised, those experiencing greater deprivation had poorer levels of health and functional status.

Acceptability
Interview administration to individuals living in less deprived areas yielded slightly higher response rates (79% versus 75% in deprived areas) (Bath et al., 2000). Average completion time was 39 minutes (range 18 to 50 minutes) in comparison to 49 minutes (range 29 to 65 minutes) for the WONCA-COOP charts (Philp et al., 2001).

Feasibility
Due to the improved level of information in relation to patient need, general practitioners selected the EASY-Care in preference to the WONCA-COOP charts (Philp et al., 2001). High levels of patient and nursing staff satisfaction were also reported, although this may be related to personal contact time with nursing staff.

Consultation costs with the EASY-Care were estimated at £2,233 per 500 patients (£4.47 per patient), with an associated cost saving from decreased consultation times of £304 per 500 patients assessed (Philp et al., 2001).

d) Functional Assessment Inventory (Pfeiffer et al., 1981)

The Functional Assessment Inventory (FAI) was designed for the assessment and screening of functional status in older people (Pfeiffer et al., 1981). It is an abbreviated version of the Older American Resources and Services (OARS) Multidimensional Functional Assessment Questionnaire (OMFAQ) (Pfeiffer et al., 1981). The omission of OMFAQ items relating to medical services restricts the assessment of service utilisation. The addition of items relating to life satisfaction and self-esteem broadens the assessment of health status.

The FAI has eleven sections across five domains, as shown in Table 5.1, namely impairment of activities of daily living (ADL), economic resources, mental health, physical health, and social resources. Mental health is assessed using the Short Portable Mental Status Questionnaire (SPMSQ) (ten items), the Short Psychiatric Evaluation Schedule (SPES) (15 items), and two further items about life satisfaction and self-esteem. The results from the SPMSQ determine whether self- or proxy completion is appropriate. Additionally, the availability, use of, and perceived need for, social and medical services in the previous six months is assessed. Finally, interviewer-perceived level of impairment is assessed across the five domains. The number of items within each domain is not clear, but the total is reported to be 90 fewer than the OMFAQ (120 items) (Pfeiffer et al., 1981).

Most items have multiple response options. Several items require written answers. Interviewers rate impairment across each domain using a six-point scale. A coding scheme modified from the OMFAQ is used to produce five domain scores (McDowell and Newell, 1996). The rating process compares patient status against standard descriptive phrases.

There have been four evaluations of the FAI (Pfeiffer et al., 1981, 1989; Cairl et al., 1983; Robinson et al., 1986): see Table 3.4. All studies describe a range of care and community settings in US populations. The results given below are derived from these studies.

Reliability

The results of test-retest reliability are shown in Table 5.2. There is no evidence for internal consistency reliability.

Interview administration with a retest interval of three to five weeks demonstrated low to high inter-observer reliability (n=2) (Cairl et al., 1983). The highest level of reliability was reported for the Short Portable Mental Status Questionnaire (0.83). Very low levels of inter-observer reliability were reported for the economic resources section (0.16).

Retest after a one-week interval showed high levels of agreement between interviewer ratings and clinical assessment by a home-care team for all domains: weighted kappa 0.53 (mental health) to 0.78 (social resources) except ADL, where interviewer-rated levels of impairment were higher than the clinical assessment (Robinson et al., 1986).

Validity

(i) Construct validity: socio-demographic variables and health-service use

The FAI discriminated between groups across four settings, with ADL as the strongest predictor and economic resources the weakest predictor of impairment levels (Pfeiffer et al., 1981). Respondents resident in nursing homes were the most impaired across all domains. Respondents from adult congregate living facilities had high levels of impairment for ADL, mental health, and social resources; those from day-care and senior centres were generally less impaired across all domains. Regarding the latter, respondents from day-care centres were most impaired in ADL, mental health and physical health domains, and those from senior centres were most impaired in the social resources domain.

The discriminative ability of the FAI was supported in a subsequent replicate study (Pfeiffer et al., 1989). As hypothesised, institutionalised respondents living in a mental health facility or nursing home were more impaired than respondents who visited a senior centre or a control group of well older people. Respondents who used a visiting-nurse service had levels of impairment that fell between these two extremes.

(ii) Construct validity: other instruments

The relationship between the FAI and OMFAQ was assessed following completion by community-dwelling older people and nursing-home residents (Cairl et al., 1983): see Table 5.3. When the FAI was completed, first correlations ranged from 0.27 to 0.86 for the economic and social resources domains, and Short Psychiatric Evaluation Schedule, respectively. When the OMFAQ was completed, first correlations ranged from 0.06 to 0.74 for the economic resources and ADL domains, respectively.

(iii) Other types of validity assessment

Correlations between FAI domains ranged from 0.32 (mental health with physical health) to 0.58 (mental health with ADL) (Pfeiffer et al., 1989). Although the small correlation between mental and physical health was expected, a larger correlation between physical health and ADL was hypothesised.

Acceptability

An overall refusal rate of 15% was reported following completion by respondents in various settings (Pfeiffer et al., 1989). The highest refusal rate was for those in receipt of a visiting-nurse service. Time for interview administration ranged from 30.6 minutes (Cairl et al., 1983) to 40 minutes (Pfeiffer et al., 1989). Shorter completion times for the FAI as compared with the OMFAQ were reported following completion by community-dwelling older people (30.6 minutes for the FAI versus 44.6 minutes for the OMFAQ) and nursing-home residents (37.0 versus 47.3 minutes) (Cairl et al., 1983). Response or completion rates were not clearly reported. However, in another study, interview administration to community-dwelling older people and nursing-home residents resulted in a completion rate of 87.9% (Cairl et al., 1983).

Completion difficulties have been reported for the economic resources domain (Pfeiffer et al., 1989).

e) Geriatric Postal Screening Survey (Alessi et al., 2003)

The Geriatric Postal Screening Survey (GPSS) was proposed as a screening tool for the identification of older people who would benefit from a comprehensive geriatric assessment, and associated health- and social-care services (Alessi et al., 2003). The questionnaire was devised in response to the need effectively to target older people who would benefit most from further detailed assessment and subsequent management, specifically those at risk of frailty or functional decline.

Item selection was informed by the literature, existing screening instruments, and expert opinion. Five conditions common in older people with evidence to support their contribution to functional decline were included: falls or problems with balance, functional impairment, depression, cognitive impairment, and urinary incontinence. Additional items included the assessment of pain, weight loss, use of assistive devices, medication use, and level of social support. The initial instrument contained 38 items. Following testing, ten items were selected: five representing health conditions and five general indicators of health status, including health perception, medications, pain, and weight loss (see Table 5.1).

Items use dichotomous or categorical response options. Item summation produces a 'risk score' of 0 to 10, where 10 is the worst health. Scores above 4 indicate high risk, scores lower than 4 indicate low risk. 39% of responders in the development survey were defined as high risk. Telephone confirmation found a low true negative rate (11% had no care needs).

There has been one evaluation of the GPSS. This was a community-based older population in the USA (Alessi et al., 2001), as shown in Table 3.4.

Reliability

The three-week test-retest reliability of the GPSS was 0.86. Individual items (not listed) ranged from 0.48 to 0.92 (see Table 5.2). Kappa agreement between risk ratings was 0.76 (88.5% agreement). There is no evidence for internal consistency reliability.

Validity

(i) Construct validity: socio-demographic variables and health-service use

Health-service use by a random sample of those who responded to the original development survey was assessed over 12 months. Groups defined as high-risk had significantly greater levels of health-service use than their low-risk counterparts. This was corroborated in a subsequent survey where people at high risk had more hospital admissions, more hospital days, and were more likely to be admitted to a nursing home than the low-risk group.

(ii) Construct validity: other instruments

Several instruments discriminated between groups defined by the GPSS as high- or low-risk, with scores for the high-risk groups suggesting greater levels of impairment. The high-risk group also had greater levels of co-morbidity.

(iii) Other types of validity assessment

When assessed against a structured telephone interview and a clinical assessment, the GPSS had high sensitivity and specificity in identifying three health conditions (risk of

falling, depression, and urinary incontinence), but limited accuracy in identifying functional impairment (ADL) and cognitive impairment. Where the GPSS functional impairment addressed ADL only, the clinical assessment also included items related to IADL.

Acceptability

The GPSS has large print (14 font). High response rates were reported for both the development (88%) and main survey (90%). 11% of respondents indicated that they required assistance in completing the form. For postal non-responders contacted by telephone, there was no statistically significant difference compared with responders in terms of age, percentage classified as high risk (42%), or mean risk score. However, the high level of non-responders classified as high risk led the authors to suggest that persuasive methods to increase response rates, for example, telephone contact and home visits, may be required.

f) Geriatric Quality of Life Questionnaire (Guyatt et al., 1993b)

The Geriatric Quality of Life Questionnaire (GQLQ) is a partly individualised health status instrument designed for the assessment of health status in frail older people (Guyatt et al., 1993). Three domains of health are addressed, namely activities of daily living (ADL), symptoms, and emotional function.

Instrument content was derived from the literature, existing instruments, and interviews with medical professionals and patients (n=100, mean age 78.5 years). It has 25 items over three domains. An individualised approach to assessment is used for the ADL and symptom domains. Respondents are invited to identify problems and to specify the degree of difficulty or distress experienced; two lists of 24 items are provided. Respondents choose up to eight items that bother them most in their daily lives. Frequency is rated for each item using a seven-point categorical scale. Items identified at baseline are rated at subsequent evaluations. The nine standardised items of the emotional function domain use a seven-point scale. Items within each domain sum to produce three scores: ADL and symptoms range from 7 to 49, where 7 is the worst and 49 the best health; emotional function ranges from 9 to 63, where 9 is the worst and 63 the best health.

There has been one evaluation of the GQLQ. This was a hospital-based older population in Canada (Guyatt et al., 1993b), as shown in Table 3.4. Although demonstrating good content validity, high levels of comparative responsiveness and validity were not demonstrated in relation to simpler instruments (Guyatt et al., 1993b).

Responsiveness
Following a 12-month trial of day-care, Modified Standardised Response Means were small for the GQLQ ADL (0.26) and symptom domains (0.30), the Rand physical function domain (0.29), and the Barthel Index (0.20). Greater levels of responsiveness were found for the emotional function domains of the GQLQ (0.50) and Rand (0.63).

Trial participants completed several instruments at baseline and 12 months. Correlations between change scores for the GQLQ ADL and instruments that had hypothesised associations ranged from 0.27 (global physical function) to 0.41 (Barthel Index); correlation with the Rand emotional function was 0.06. Correlations between change scores for the GQLQ emotional function domain and instruments that had hypothesised associations ranged from 0.44 (global ratings of emotional function) to 0.61 (Rand emotional function); correlation with the Barthel Index was 0.17.

The change score correlation between the GQLQ and global overall health was 0.18. The authors suggest that change in global rating may have limited validity. Change score correlations between GQLQ domains ranged from 0.04 (ADL with emotional function) to 0.31 (symptoms with emotional function).

Acceptability
The mean interview administration time was 30 minutes (range 20 to 60 minutes). Due to the potential impact on respondents of emotional function items, the recommended order for domain completion was ADL, followed by emotional function, and, lastly, symptoms.

Due to impaired cognitive ability, 33% of study participants were considered inappropriate as respondents. Low levels of missing data were reported (ADL 0.19%, emotional function 3.6%, symptoms 2.3%).

g) Geriatric Screening Questionnaire (23-item) (Fernandez Buergo et al., 2002)

The Geriatric Screening Questionnaire (GSQ) was proposed as a simple, primary care screening tool for identifying older people (aged over 65 years) at risk of functional decline and who would benefit from a comprehensive geriatric assessment (Fernandez Buergo et al., 2002).

Item selection was informed by a survey of risk factors in older people. The initial instrument contained 23 items addressing issues of cognitive impairment, daily activities, economic status, general health status, mental health, and social support: see Table 5.1. Items use dichotomous response options. Item summation gives a score from 0 to 23, where 0 is better health and 23 indicates worst health and greater risk of functional decline. Following initial testing, two reduced questionnaires with five and six items, respectively, were produced. The developers recommend the six-item questionnaire as a valid screening instrument to support the implementation of a comprehensive geriatric assessment in a primary-care setting.

The instrument is designed for interview administration in a primary-care or home setting.

There has been one evaluation of the GSQ, in a community-based older population in Spain (Fernandez Buergo et al., 2002), as shown in Table 3.4.

Reliability

Two-week test-retest reliability at item level ranged from 0.60 to 0.86 (Table 5.2). There is no evidence for internal consistency reliability.

Validity

Validity: other

Following definition of groups using scores on a comprehensive geriatric assessment as having positive or negative health, the 23-item GSQ had a sensitivity of 50% and a specificity of 89% when used as a confirmation test for functional decline. When used as an exclusion test for functional decline, sensitivity was 88% and specificity 40%. The six-item questionnaire had a sensitivity of 58% and a specificity of 89% when used as a confirmation test, and a sensitivity of 81% and a specificity of 56% when used as an exclusion test. Similar levels were reported for the five-item questionnaire (including age).

Acceptability

High completion rates were reported (91.2%).

h) Iowa Self-Assessment Inventory (Morris and Buckwalter, 1988; revised 1990)

The IOWA Self-Assessment Inventory (ISAI) was designed to encompass multiple physical, mental, and social functions in older people (Morris and Buckwalter, 1988). It was developed for use in needs assessment and assessment of individuals, and to inform screening procedures for admission to residential facilities. The authors state that it may be useful for large-scale community surveys to assess service needs, or for planning purposes such as housing.

Instrument content was derived from the literature and existing instruments, with particular reference to the OMFAQ (Fillenbaum and Smyer, 1981, cited by Morris and Buckwalter, 1988). The preliminary ISAI has six domains, namely activities of daily living (ADL), cognitive status, economic resources, mental health, physical health, and social resources. Each domain has 20 items with four-point response scales. Items sum to produce a score from 20 to 80, where 80 is better health. The instrument was designed to be self-completed by relatively well older people, or interview-administered. The expert opinion of health professionals confirmed content validity, clarity, and readability. Piloting of the instrument involved completion by groups of older people who were housebound and receiving home-delivered meals, and participants in a congregate meals programme. As hypothesised, scores discriminated between groups, with members of the congregate meals programme reporting better levels of health.

Factor analysis of the preliminary ISAI gave a seven-factor solution with four factors contributing 52% of the variance: these were cognitive status, economic resources, mobility (ability to get around, ADL, and social activities), and physical health (Morris et al., 1990). The ISAI was revised to seven domains: anxiety/depression, alienation, cognitive status, economic resources, mobility, physical health, and social support. Items with the highest loadings were retained for each of the seven factors, resulting in eight items per domain. Four-point response options were retained to produce domain scores of 8 to 56, where 8 is worst and 56 is best health.

There have been two evaluations of the preliminary ISAI (Morris and Buckwalter, 1988; Morris et al., 1989) and a single evaluation of the revised ISAI (Morris et al., 1990). All evaluations referred to older people from various community settings within the USA; two evaluations referred to the same population group (Table 3.4). The results given below are derived from these studies.

Reliability

High levels of internal consistency reliability have been reported for all domains of the preliminary and revised ISAI (Table 5.2). There is no evidence for test-retest reliability.

Validity

(i) Construct validity: socio-demographic variables and health-service use

The correlation between preliminary ISAI domains and several demographic variables was assessed (Morris et al., 1989). Correlation between ISAI economic resources and income was 0.36, between ISAI ADL and age –0.32, and between three ISAI domains (economic resources, mental health, and social resources) and education level ranged from 0.21 to 0.27 (see Table 5.3). The remaining correlations were all very small. Domains discriminated between groups defined by income level (economic resources),

age (ADL, cognitive status, economic resources), education (ADL, cognitive status, economic and social resources, mental and physical health) and living arrangements (mental health).

As hypothesised, the preliminary ISAI ADL, physical health, and social resources domains discriminated between groups defined as relatively fit and attending a meal programme, and those who were housebound and receiving home-delivered meals; cognitive status and mental health did not discriminate between groups (Morris and Buckwalter, 1988). Although not hypothesised, economic resources discriminated between groups, with lower scores for housebound elderly people. Sex, age, educational level, and type of living arrangement did not differ significantly between groups.

(ii) Validity: other
When the preliminary ISAI was completed by well elderly, correlations were in accordance with hypotheses and ranged from 0.50 (economic resources with mental health) to 0.63 (mental health with physical health) (Morris and Buckwalter, 1988): see Table 5.3. When completed by housebound elderly people, correlations were also in accordance with hypotheses and ranged from 0.55 (cognitive status with ADL) to 0.71 (cognitive status with mental health). Further evaluation of inter-domain correlation for the preliminary ISAI ranged from 0.19 (cognitive status with economic resources) to 0.59 (physical health with ADL) (Morris et al., 1989).

Following completion of an experimental nine-domain initial revision to the ISAI by 420 community-dwelling older people, inter-domain correlation ranged from 0.04 (alienation with mobility) to 0.89 (anxiety with mental health) where, as hypothesised, related domains had the largest correlation (Morris et al., 1990). Further revision resulted in the revised 56-item, seven-domain instrument, where inter-domain correlations ranged from 0.16 (physical health and cognitive status) to 0.40 (physical health and mobility).

Acceptability
Median self-completion time for the revised ISAI was 15 minutes compared with 30-45 minutes for the preliminary version and more than one hour for the OMFAQ (Morris et al., 1990).

i) LEIPAD Quality of Life questionnaire (De Leo et al., 1998)

The LEIPAD was developed under the auspices of the World Health Organisation European office for use as a comprehensive evaluative instrument suitable for the assessment of quality of life in older people (De Leo et al., 1998). The cross-cultural development of the instrument provided a basis for the name 'LEIPAD', an acronym derived from the participating countries (Leiden, the Netherlands, and Padua, Italy). Translations are available in several languages, including English, Dutch, Finnish, and Italian.

Instrument content was derived from existing instruments and the opinion of psycho-geriatricians. Several versions of the instrument were assessed before the LEIPAD was defined. The LEIPAD is self-reported and comprises 49 items, 31 of which measure seven domains: cognitive functioning (five items), depression/anxiety (four items), life satisfaction (six items), physical function (five items), self-care (six items), social functioning (three items), and sexual functioning (two items): see Table 5.1. Each item uses a four-point categorical scale. Items sum to give domain scores or a global score from 0 to 93, where 93 is maximum impairment. Factor analysis of the core items gave two factors: psychosocial function (life satisfaction, depression/anxiety, cognitive functioning) and physical function (self-care, physical function).

The additional 18 items serve as moderators for assessing the influence of social desirability factors and personality characteristics on the seven domain scores. These 18 items cover five domains and are taken from available instruments, namely Perceived Personality Disorder Scale, Anger Scale, Social Desirability Scale, Self-esteem Scale, and the Trust in God Scale. When completed by older people, the moderator scales did not influence self-reported health status and the authors indicate that they will be omitted from a revised, short version of the LEIPAD. There is as yet no publication relating to the revised version.

The LEIPAD has been evaluated in one further published study (Condello et al., 2003) since the original developmental article (De Leo et al., 1998), as shown in Table 3.4. Both studies used European community-dwelling older populations. The results given below are derived from these studies.

Reliability

The two-week test-retest reliability of the LEIPAD was assessed following completion by 50 Italians and produced a high coefficient of 0.81 (De Leo et al., 1998) (Table 5.2). High levels of internal consistency reliability were reported for four of the seven domains in the same population (ranging from 0.43 sexual function to 0.79 cognitive functioning).

Validity

(i) Construct validity: other instruments.

Correlation between LEIPAD and the Rotterdam Questionnaire (RQ) domain scores that had hypothesised associations ranged from 0.70 (LEIPAD physical function with RQ physical distress) to 0.71 (LEIPAD depression/anxiety with RQ psychological distress) and 0.79 (LEIPAD self-care with RQ ADL) (De Leo et al., 1998): see Table 5.3. Correlations between domains that did not have hypothesised associations ranged

from 0.10 and 0.13 (LEIPAD sexual function with RQ psychological distress and RQ physical distress, respectively) to 0.14 (LEIPAD social function with RQ ADL).

(ii) Validity: other
The LEIPAD domains and index discriminated between groups with and without personality disorders; the presence of personality disorders explained 47.5% of variance in the index score (Condello et al., 2003). Correlation between the LEIPAD index and different personality disorders (diagnostic scale) ranged from 0.35 (passive-aggressive) to 0.68 (depressive).

Acceptability
Completion rates ranged from 80.4% to 100%. The proportion of non-responders varied by country of administration (Finland 19.6%, Italy 8.3%, Holland 0%), and may have been influenced by different recruitment methods (De Leo et al., 1998). Self-administration takes approximately 15-20 minutes.

j) OARS Multidimensional Functional Assessment Questionnaire (OMFAQ) (Pfeiffer, 1975; revised: George and Fillenbaum, 1985).

The Older Americans Resources and Services (OARS) Multidimensional Functional Assessment Questionnaire (OMFAQ) was developed in the USA during the 1970s for use as a screening and evaluative instrument of functional status in and service use by adults, specifically older people (Pfeiffer, 1975; George and Fillenbaum, 1985). The developers suggest that it may also inform resource allocation. The instrument has provided the foundation for many subsequent instruments which aim to assess functional ability in older people, for example, the FAI (Pfeiffer et al., 1989).

Instrument content was informed by the expert opinion of medical and social care professionals, relevant literature, and existing instruments. Developmental versions of the instrument were piloted with patient groups. The instrument developers defined a three-element OARS assessment model, comprising: i) individual functional status, ii) health and social service use, and iii) a transition matrix to describe service use according to functional level (George and Fillenbaum, 1985; McDowell and Newell, 1996). The OMFAQ parts A and B represent the first two elements of this model. Administration by a trained interviewer is required. Before administering the OMFAQ, respondents complete a ten-item mental status questionnaire to determine whether or not proxy completion is necessary. The OMFAQ has been translated into a number of languages.

Part A, the OARS Multidimensional Functional Assessment Questionnaire (OMFAQ), assesses function across five domains, namely activities of daily living (ADL) - both instrumental activities of daily living (IADL) and basic ADL, economic resources (income, reserves, and assets), mental health (cognitive functioning, life satisfaction, psychiatric status, self-evaluation of mental health), physical health (medication use, illness/chronic conditions, self-evaluation of health status), and social resources (amount/adequacy of social interaction, availability of help). Although the developers advise against applying domains separately, evidence suggests that the ADL and IADL are frequently used as separate scales (see below). There are inconsistencies in the number of reported items, both within domains (Liang et al., 1989) and for the total instrument. The original article describes a total of 104 items: 70 answered by the respondent, 10 by an informant, 14 by the interviewer, and 10 items within the mental status questionnaire (George and Fillenbaum, 1985). McDowell and Newell (1996) report 120 items including sub-parts within several questions. The Quality of Life database (www.qolid.org/) reports 101 items including subparts.

The final section of the OMFAQ (interview section) requires interviewers to rate each domain on a six-point scale ranging from best function to complete impairment. Five domain scores are produced that may be summed to give a cumulative impairment score (CIS) ranging from 5 to 30, where 5 is excellent function and 30 total impairment (McDowell and Newell, 1996). Alternatively, domain scores may be dichotomised into impaired versus not impaired. Further guidance for scoring is provided in a comprehensive user's manual (Fillenbaum, 1988, cited by McDowell and Newell, 1996). A computer-coding programme may be used which incorporates clinical judgement to weight individual items (George and Fillenbaum, 1985; McDowell and Newell, 1996).

Part B, the Services Assessment Questionnaire (SAQ), assesses the respondent's need for health and social services across 24 categories. The frequency of use in the previous six months, provision of service, and perceived need for a service are assessed. Parts A and B can be applied separately. The transition matrix links service use to functional status. This procedure requires information about functional status collected at two times and the service packages used in the interval between test dates.

Three additional sections within the OMFAQ, comprising basic demographic information (11 items), informant assessment (ten items), and interviewer rating (19 items) are also completed. The informant assessment section elicits information from a knowledgeable informant in relation to the five OMFAQ core domains. The interviewer section requires the interviewer to estimate the reliability of responses (four items) and to rate the five domains (15 items).

Exploratory factor analysis supports five multi-item functional domains: ADL (two scales), economic resources (one scale), mental health (four scales), physical health (one scale), and social resources (three scales) (George and Fillenbaum, 1985). Factor analysis of the social resources domain yielded four factors: perceived resource adequacy, resource availability, social attachments, and interaction (Harel and Deimling, 1984).

A five-item screening instrument based on instrumental activities of daily living has been recommended for the speedy identification of older community residents with impaired functional capacity (Fillenbaum, 1985; McDowell and Newell, 1996).

12 articles describe the evaluation of the OMFAQ, as shown in Table 3.4. With the exception of one evaluation in a community-dwelling Australian population (Osborne et al., 2003), all studies describe populations from North America across a range of hospital and community settings. The results given below are derived from these articles.

Reliability

Evidence of reliability for the OMFAQ is shown in Table 5.2. High levels of internal consistency reliability have been reported for the IADL items within the ADL domain, ranging from 0.68 (Reuben et al., 1995) to 0.92 (Breithaupt and McDowell, 2001). High levels of item-total correlation were found for all items within the ADL domain (range 0.68 to 0.84) (Breithaupt and McDowell, 2001).

Following a five-week retest period, 91% of OMFAQ responses were reported to be identical (Fillenbaum, 1978, cited by George and Fillenbaum, 1985).

Multiple raters across different disciplines and geographical regions rated the five domains of 30 completed OMFAQ interviews (Fillenbaum and Smyer, 1981). High levels of inter-rater reliability were reported for all domains: 0.66 for physical health, 0.78 for economic resources, 0.80 for mental health, 0.82 for social resources, and 0.86 for ADL. Raters agreed on 74% of the ratings.

Validity

(i) Construct validity: socio-demographic variables and health-service use

An inverse relationship between IADL scores and survival status at one year was reported for community-dwelling older people (Fillenbaum, 1985). The overall death rate was 5%. The death rate for those unable to perform any of the five IADL activities unaided was 27%. Only 2% of respondents who could perform all activities died within the year.

(ii) Construct validity: other instruments

Correlations between the ADL domain and OMFAQ domains that had hypothesised associations ranged from 0.54 (mental health) to 0.60 (physical health). Correlation between the ADL domain and OMFAQ domains that did not have hypothesised associations ranged from 0.11 (social resources) to 0.19 (economic resources) (Fillenbaum, 1985): see Table 5.3. A similar relationship was reported at one-year follow-up.

Correlation between the OMFAQ social resources domain and several mental health instruments was assessed (Harel and Deimling, 1984). Correlation with the Minnesota Multiphasic Personality Inventory (MMPI) ranged from –0.05 (talks to people) to –0.46 (doesn't feel lonely), with self-rated mental health ranged from 0.08 (has emergency support) to 0.31 (doesn't feel lonely), and with interviewer-rated mental health ranged from 0.04 (has confidant) to 0.40 (informal assistance). Social attachments and social interaction explained a limited amount of variance within each of the three mental health instruments (ranging from 13% for self-rated mental health to 31% for MMPI and interviewer-rated mental health).

Correlations between the OMFAQ IADL items and instruments that had hypothesised associations ranged from 0.56 (Physical Performance Test) to 0.70 (Functional Status Questionnaire ADL) (Reuben et al., 1995). Correlation between ADL and SF-36 domain scores ranged from 0.36 (physical functioning) to 0.49 (role-physical).

Correlation between the OMFAQ index score and the Functional Autonomy Measurement System (SMAF), a clinical measure of disability, was 0.80, in accordance with hypotheses (McCusker et al., 1999). Correlation between instrument domains that had hypothesised associations ranged from 0.31 (ADL with SMAF communication) to 0.77 (IADL with SMAF IADL).

Correlations between the OMFAQ IADL and SF-20 physical and role function domains were in accordance with hypotheses, and ranged from 0.56 (role function) to 0.67 (physical function) (Carver et al., 1999). Correlation between the OARS IADL and SF-36 physical function domain was greater than 0.60, also in accordance with hypotheses (Stadnyk et al., 1998).

Correlations between the OMFAQ self-care domain and the AQoL index (utility) and independent living domain were –0.68 and 0.82, respectively, in accordance with hypotheses (Osborne et al., 2003). Correlations ranged from 0.03 (OMFAQ independent living with AQoL social resources) to –0.40 (OMFAQ social relationships with AQoL self-care), in accordance with hypotheses (see Table 5.3).

(iii) Validity: other
Interviewer-rated domain summary scores were compared with clinical criteria for 33 community-dwelling older people (Fillenbaum and Smyer, 1981). Mental and physical health were assessed by a gero-psychiatrist and a physician, respectively. Home-based, self-care capacity was assessed by a physical therapist, and economic resources by comparison with an objective six-point economic scale. The mean time between OMFAQ administration and assessment ranged from nine days (gero-psychiatrist and physician) to 35 days (physical therapist; change in ADL capacity was recorded). Large correlations were found between interviewer summary ratings and external criteria, as shown in Table 5j below.

Table 5j Correlation between OMFAQ domains and clinically assessed criteria

Domains	***n***	***Spearman correlations***
Economic	49	0.68
Mental health	31	0.67
Physical health	31	0.82
Self-care capacity	30	0.89

Several authors have explored the OMFAQ factor structure. Three-factor (Fillenbaum, 1985) and four-factor solutions (Harel and Deimling, 1984) were found for the social resources domain, comprising attachments, social interaction, social support, and adequacy of social resources. Four factors were found for mental health, namely alienation, cognitive deficit, life satisfaction, and psychosomatic symptomatology (Liang et al., 1989). Items describing 'affect' were lacking. Two broad factors, IADL and ADL, constitute the physical health domain (Fillenbaum, 1985; McDowell and Newell, 1996). Five items were retained in the IADL factor, namely travel, shopping, meal preparation, housework, and handling personal finances (Fillenbaum, 1985). The IADL items are Guttman-scaled, shopping being the easiest activity and housework the most difficult (Fillenbaum, 1985; McDowell and Newell, 1996).

When stratified for age, IADL scores had predictive validity for both mental and physical health at one year; low scores were predictive of mortality (Fillenbaum, 1985; McDowell and Newell, 1996).

The properties of the ADL domain (seven ADL items, seven IADL items) were evaluated using Item Response Theory (IRT) (Breithaupt and McDowell, 2001). With completion by elderly caregivers, the most highly discriminating ADL items were 'getting out of bed', 'toilet transfer', and 'dressing'; the most highly discriminating IADL items were 'shopping', 'getting places', and 'preparing meals'. The two sub-scales measure most precisely at different functional levels: ADL is most precise at lower functional levels, IADL is most precise at higher levels. Despite strong inter-domain correlation (0.79), the items were described by a two-dimensional IRT analysis. The difference in severity and type of activity covered by ADL and IADL, respectively, supports the independent use of the two sections.

Responsiveness
Although moderate to good levels of responsiveness were reported for the OMFAQ physical health (PH) domain at six weeks and six months following surgical repair of hip fracture (ES 0.80 and 0.50, respectively), it was less responsive than the SF-36 and a

condition-specific measure (Jaglal et al., 2000). The OMFAQ PH domain discriminated between groups defined by their pre-fracture versus six-week post-operative scores, and six-week versus six-month post-operative scores.

Following the assessment of community care co-ordination versus usual care, an external criterion of health deterioration was defined as admission to institutional care after 18 months (Osborne et al., 2003). Low levels of responsiveness were found for the OMFAQ when assessed by Relative Efficiency and Receiver Operating Characteristic curves. However, in contrast to two generic measures of HRQL, the baseline scores for the OMFAQ self-care and social resources domains discriminated between people who remained community-dwelling and those requiring institutionalised care at 18 months.

Precision

Skewed response distributions with associated ceiling effects have been reported for the ADL domain, with a large percentage of respondents rated as independent in activities (67% Reuben et al., 1995; range: 25-75% , with greater ceiling effects for IADL) Breithaupt and McDowell, 2001). Floor effects have not been reported.

Acceptability

Part A of the OMFAQ takes approximately 30 minutes to complete (McDowell and Newell, 1996). The whole interview takes approximately 45 minutes (Fillenbaum and Smyer et al., 1981; McDowell and Newell, 1996).

High participation rates have been reported for both proxy (81%) (Breithaupt and McDowell, 1996) and respondent completion (Reuben et al., 1995).

k) Perceived Well-being Scale (Reker and Wong, 1984)

The Perceived Well-being Scale (PWB) was developed in Canada during the 1980s for assessing the psychological, physical, and general well-being of community-dwelling older people and those living in institutions (Reker and Wong, 1984).

Item content was derived from the literature, existing instruments, and consultation with psychology students. Factor analysis produced two factors and supported item reduction from 32 to 14 items. Items are rated on seven-point Likert scales (1: strongly agree, to 7: strongly disagree). Psychological well-being (six items) is scored from 6 to 42, where 42 is better health. Physical well-being (eight items) is scored from 8 to 56, where 56 is better health. The total index is scored from 14 to 98, where 98 is best general well-being. The method of administration was not reported.

The original publication describes instrument evaluation in both community-dwelling older Canadians and those living in institutions (Reker and Wong, 1984), as shown in Table 3.4. Further evaluation included community-dwelling older women (Cousins, 1997). The results given below are derived from these articles.

Correspondence with the instrument developers described a revised instrument with two additional items in the psychological well-being domain (peace of mind, afraid of many things) (Reker, 1995: unpublished manuscript). A published evaluation of this instrument was not found through electronic searches.

Reliability

High levels of internal consistency reliability have been reported for both domains: 0.82 for psychological well-being and 0.78 for physical well-being, and for the total index: 0.91 (Reker and Wong, 1984), as shown in Table 5.2.

Moderate levels of reliability were reported following a two-year retest period, namely 0.79 for psychological well-being, 0.65 for physical well-being, and 0.78 for general well-being (Reker and Wong, 1984): see Table 5.2. Moderate test-retest reliability (0.60) was found following a four-week retest period (Cousins, 1997).

Validity

(i) Construct validity: socio-demographic variables and health-service use

As hypothesised, the PWB discriminated between community-dwelling older people, who rated well-being more highly, and those living in institutions (Reker and Wong, 1984): see Table 5.3.

As hypothesised, the PWB discriminated between groups of older women defined by age and level of exercise; older groups reported worse health irrespective of additional health symptoms, and more active women reported better health (Cousins, 1997).

(ii) Construct validity: other instruments

Correlations between the PWB general well-being index score and several new and established instruments with hypothesised associations included the Personal Optimism Scale (0.40), Commitment to Life Events Survey (0.42), Beck Depression Scale (–0.54), and the Memorial University of Newfoundland Scale of Happiness (MUNSH) (0.70) (Reker and Wong, 1984): see Table 5.3. Correlation between the PWB index score and

self-rated physical symptoms was –0.25, which was smaller than hypothesised. Correlation between PWB psychological and physical well-being domains and self-reported physical health were 0.06 and –0.40, respectively, in accordance with hypotheses. Correlations between PWB physical and psychological well-being domain scores and the MUNSH were 0.52 and 0.69, respectively.

Correlation between the PWB index score and several single item assessments ranged from –0.39 (for medication intake and self-rated global health) to –0.51 (symptoms) (Cousins, 1997).

l) Philadelphia Geriatric Center Multilevel Assessment Instrument (Lawton et al., 1982)

The Philadelphia Geriatric Center Multilevel Assessment Instrument (PGCMAI) was designed for the multidimensional assessment of community-dwelling older people (Lawton et al., 1982). It is recommended for use in research and service-based assessment. Informed by the OMFAQ and other instruments, the PGCMAI describes four core components of a 'good life', namely behavioural competence, psychological well-being, quality of life, and quality of the environment (Lawton et al., 1982; Wissing and Unosson, 2002). An activity hierarchy is defined within each domain.

The PGCMAI has seven domains, namely activities of daily living (ADL): physical self-maintenance and instrumental activities of daily living (IADL); cognition: mental and cognitive status; perceived environment: housing, neighbourhood, and personal security; personal adjustment: morale, psychiatric symptoms; physical health: self-rated health, health behaviour, health conditions; social interaction with friends and family; and time use: ways of spending time, for example, hobbies. The original 135 items were revised to 147 items across the seven domains (14 sub-scales) (McDowell and Newell, 1996), as shown in Table 5.1. Mid-length (68 items) and short (24 items) versions have been described.

The respondent answers selected items only; an informant may provide additional information. Although response options are not clarified in the published literature, checked items within each domain and sub-domain are summed to produce seven domain scores (McDowell and Newell, 1996). As with the OMFAQ, interviewers use five-point scales to provide summary assessments of interviewees across the seven domains.

There has been one evaluation of the PGCMAI in a mixture of population settings in the USA (Lawton et al., 1982), and two evaluations in a community-dwelling population in Sweden (Wissing and Unosson, 2001; 2002), as shown in Table 3.4. The results given below are derived from these studies.

Reliability

Internal consistency reliability was assessed for the full, mid-length, and short versions (Lawton et al., 1982), as shown in Table 5.2. Higher levels of internal reliability were reported for all domains in the longer version (range: 0.71 to 0.93). Four domains in the mid-length version (range: 0.29 for social interaction to 0.66 for physical health) and all but one domain (ADL) in the short version (range: 0.04 for social interaction to 0.63 for cognition) had very low levels of internal reliability.

Good levels of reliability were reported across all domains following a three-week retest period (range: 0.73 for social interaction to 0.95 for physical health) (Lawton et al., 1982): see Table 5.2.

Validity

(i) Construct validity: socio-demographic variables and health-service use

The 'criterion group' variable represents residential status: independent versus dependent living (Lawton et al., 1982). Correlations with the PGCMAI ranged from 0.05 for perceived environment to 0.54 for ADL.

(ii) Construct validity: other instruments
Correlations between PGCMAI respondent scores and interviewer and clinician ratings were in accordance with hypotheses, ranging from 0.36 (perceived environment) to 0.87 (ADL) for interviewer ratings, and 0.23 (cognition) to 0.65 (physical health) for clinician ratings, respectively (Lawton et al., 1982): see Table 5.3. Correlation between sub-domain items and summary domain ratings ranged from 0.19 (perceived environment) to 0.87 (ADL); correlation between sub-domain items ranged from 0.09 (cognitive symptoms) to 0.78 (psychiatric symptoms).

(iii) Validity: other
The mid-length PGCMAI discriminated between groups defined by the presence or absence of leg ulcers; those without leg ulcers reported better health across several domains (Wissing and Unosson, 2002).

Responsiveness
Patients with open ulcers had worse health scores over four years for mobility and ADL domains (Wissing and Unosson, 2001). Those with healed ulcers showed improved scores for subjective housing and neighbourhood. The social domain index discriminated between patients with healed ulcers and those with unhealed ulcers after four years.

Acceptability
The instrument is interview-administered; the full-length version takes approximately 50 minutes to complete (Lawton et al., 1982). In a sample of 615 respondents completing a 216-item schedule, 55 (8.9%) required assistance and the results from 25 respondents (4.0%) were unusable due to missing data.

m) Quality of Life Cards (QLC)(Rai et al., 1995)

The Quality of Life Cards (QLC) were developed in Holland to evaluate the impact of old age on an individual's quality of life (Rai et al., 1995). Instrument content was informed by a literature review which identify multiple domains contributing to the concept of quality of life.

80 items or 'cards' assess three core domains: affect, life experience, and satisfaction/happiness. 20 cards contain words or statements describing positive or negative affect. 20 cards describe positive or negative life experiences. 40 cards assess the level of satisfaction or happiness in four key areas: family life, health or function, personal life, and religion. Respondents pick cards containing a word or statement that best applies to them. The score of 1 is given for a card describing a 'positive' affect, life experience, or level of satisfaction/happiness' and –1 is given for cards depicting a 'negative' attribute. Items sum to give scores ranging from –80 to 80, where 80 is the best quality of life.

There has been one evaluation of the QLC. This was a community-based older population in Holland (Rai et al., 1995), as shown in Table 3.4.

Reliability
A three-day retest completion by 11 people showed a very high level of reliability (0.99), as shown in Table 5.2.

Validity
(i) Construct validity: other instruments
Correlation between the QLC and the Delighted-Terrible scale was –0.96, and with a visual analogue scale was 0.93 (see Table 5.3).

(ii) Other types of validity assessment
Correlation between QLC total score and scores for the affect, life experience, and satisfaction/happiness domains ranged from 0.90 to 0.97.

n) Quality of Life Profile - Seniors Version (Raphael et al., 1995a,b)

The Quality of Life Profile - Seniors Version (QOLPSV) was designed to evaluate the quality of life of community-dwelling older people (Raphael et al., 1995a, 1997). The developers suggest that the QOLPSV may be used to assess the impact of medical and social interventions on quality of life, to assess service needs, and to identify areas where health promotion is indicated.

Relevant literature and group meetings with community-dwelling older people and service-providers informed instrument content. Further modifications were made following completion by two groups of older people.

The instrument is self-completed, and has 111 items over three domains and nine sub-domains (see Table 5.3). The Being domain comprises physical, psychological, and spiritual sub-domains (36 items); Belonging comprises physical, social, and community sub-domains (36 items); and Becoming comprises practical, leisure, and growth sub-domains (39 items). Completion is in two stages: first, respondents rate the relative importance and enjoyment for each item using a five-point scale. The importance scores "serve as a weight for converting enjoyment scores into quality of life (QOL) scores [QOL = (importance score/3) x (enjoyment score –3)]" (Raphael et al., 1995, p162). QOL scores range from –3.43 (not at all satisfied with important issues) to +3.43 (very satisfied with important issues). Where an activity is enjoyed, items rated as important produce high QOL scores. Conversely, where an activity is not enjoyed, items rated as important produce low QOL scores. Importance, enjoyment, and QOL scores may be calculated for each domain and sub-domain.

The second stage asks respondents to rate the degree of control, or how much opportunity they have for improving or maintaining control, for the nine sub-domains using a five-point scale (1 - worst, to 5 - best). The result helps with QOL score interpretation.

Discussions with health professionals produced both short (54-item) and brief (27-item) versions of the QOLPSV (Raphael et al., 1995b). The full version is recommended for exhaustive diagnostic surveys. The short version is recommended for research purposes and where less extensive detail is required. The brief version is recommended for screening purposes.

Three articles describe the original development and evaluation of the QOLPSV (Raphael et al., 1995a,b, 1997). All describe the same community-dwelling Canadian population, as shown in Table 3.4. Irvine et al. (2000) evaluated the enjoyment subscale of the brief QLPSV only. They also describe a simplified scoring format where the enjoyment (and importance) of each item is rated on a five-point scale (1 - not satisfied). Items within each sub-scale are summed. The results given below are derived from these articles.

Reliability

High levels of internal consistency reliability have been reported for the three versions of the QOLPSV (Raphael et al., 1995a,b, 1997), as shown in Table 5.2. Moderate to high levels of internal consistency reliability have been reported for sub-scales of the QOLPSV brief version ranging from 0.47 (Belonging-community) to 0.82 (Becoming-

leisure), where seven domains had an alpha greater than 0.70 (Irvine et al., 2000). However, these levels were generally lower than the SF-36 when completed in the same population (domain range 0.76 to 0.94). There is no evidence for test-retest reliability.

Validity

(i) Construct validity: socio-demographic variables and health-service use

Recognising the importance of environment to quality of life, instrument developers hypothesised that socio-demographic variables such as income, education, and age would be good indicators of environmental quality (Raphael et al. 1995b, 1997). Consequently, the small correlation between quality of life scores (all versions) and these variables was not expected (numerical values not reported).

As hypothesised, the QOLPSV brief version discriminated between groups defined by level of nursing care required (Irvine et al., 2000). Low scores for several domains were correlated with more intensive levels of nursing care: Becoming-practical (–0.40), Being-physical (–0.43), Being-spiritual (–0.36), Belonging-physical (–0.46), Belonging-social (–0.50).

(ii) Construct validity: other instruments

Correlations between the QOLPSV domains and self-reported health status ranged from 0.37 (Belonging-social) to 0.57 (Being-physical) (Raphael et al. 1995b, 1997).

Correlations between the QOLPSV and several patient-reported measures of health status were in accordance with hypotheses and ranged from 0.11 (Being-psychological with National Council on Aging Activity Questionnaire [NCAAQ]) to 0.62 (Belonging-community with Memorial University of Newfoundland Scale of Happiness, and Becoming-leisure with NCAAQ) (Raphael et al. 1995a, 1995b, 1997): see Table 5.3. Correlations between the QOLPSV and the Life Satisfaction Scale ranged from 0.19 (Being-physical) to 0.37 (Belonging-social). Correlations were similar across the three versions of the QOLPSV.

(iii) Other types of validity assessment

Correlation between the three versions of the QOLPSV ranged from 0.95 to 0.99 (full length version with short version) and 0.88 to 0.98 (full length version with brief version), and were in accordance with hypotheses (Raphael et al. 1995a,b,1997).

Responsiveness

The relationship between change in instrument score and aspects of nursing care in patients with acute or chronic illness was assessed against several hypotheses: first, that health scores for acute care patients would improve more than those for palliative or chronic care patients; second, that patients receiving care from one provider would experience greater score improvement than patients receiving care from multiple nurse providers; and finally, that the proportion of visits made by registered nurses would be positively associated with score improvement (Irvine et al., 2000).

A statistically significant improvement in instrument score was found for four out of nine sub-scales, namely Being-physical, Being-psychological, Becoming-practical, and Becoming-growth. However, the QOLPSV was less responsive than the SF-36 (except for general health). Moreover, score change did not discriminate between acute, chronic, or palliative care patients, and continuity of care was not associated with

greater improvement in health status. Unlike score change with the SF-36, QOLPSV score change did not discriminate between groups defined by the number of nurse visits.

Acceptability
Interview administration of the QOLPSV takes up to one hour.

Instrument developers reported a 67% response rate (Raphael et al. 1995a,b, 1997). Although a 51% response rate was reported for a test-retest completion, 100% correct completion was reported for the QOLPSV-brief version (Irvine et al., 2000). Missing values for the QOLPSV and SF-36 were similar.

o) Quality of Life - Well-being, Meaning, and Value (Sarvimäki and Stenbock-Hult, 2000)

The Quality of Life - Well-being, Meaning, and Value (QLWMV) represents a battery of instruments for assessing quality of life in older people (Sarvimäki and Stenbock-Hult, 2000).

Five domains of quality of life are defined, namely well-being (satisfaction with living area, economic and health status), meaning (life purpose, intelligibility, and manageability), value or self-worth, health, and functional capacity. A sixth domain comprises external factors (living area, housing, accommodation, and family and social contact). Instrument content was largely derived from the literature and existing instruments, with additional items proposed by the developers.

Two instruments within the battery assess life meaning. The Purpose in Life Test comprises 20 items, which sum to give a score from 20 to 140, where high scores indicate a clear purpose in life (Crumbaugh and Maholick, 1964). The Sense of Coherence Test comprises 13-items, which sum to give a score from 13 to 91, where high scores indicate a strong sense of coherence (Antonovsky, 1987).

One instrument, the Self-esteem Scale, assesses self-worth; it comprises ten items which sum to give a score from 10 to 40, where 40 is high self-esteem (Rosenberg, 1965).

Health assessment includes the Psychosomatic Symptom Scale, comprising 12 items which sum to give a score from 12 to 48 score, where 48 is best subjective health (Andersson, 1981). Functional capacity assessment includes the Activities of daily living Ladder (ADL ladder), comprising ten items which sum to give a score from 10 to 30, where higher scores indicate greater independence (Hutler Asberg, 1988). Sensory-motor capacity is assessed by four questions relating to hearing, movement, speech, and vision.

The instrument reportedly comprises 74 items, although this is unclear.

Scores are calculated for each instrument or set of items within each domain. An index score for each domain or the defined ‘model’ is not calculated. A score is not calculated for external conditions.

One study describes the development and evaluation of the QLWMV. This referred to a community-dwelling older population in Finland (Sarvimäki and Stenbock-Hult, 2000), as shown in Table 3.4.

Reliability
High levels of internal consistency reliability have been reported for instruments within several domains, ranging from 0.79 (for the Psychometric Symptom Scale) to 0.86 (for the Purpose in Life Test [PIL]), as shown in Table 5.2.

There is no evidence for test-retest reliability.

Validity

(i) Construct validity: socio-demographic variables and health-service use

Correlation between the external factors domain and other instruments within the QLWMV ranged from 0.16 (family network with Sense of Coherence [SOC]) to 0.25 (social network with Self-esteem Scale [SES]), as shown in Table 5.3.

(ii) Construct validity: other instruments

Correlation between instruments within different QLWMV domains ranged from 0.19 (ADL-ladder with SOC) to 0.62 (SES with PIL).

(iii) Validity: other

Regression analysis was used to explore the relationship between domains. PIL was best explained by ADL, family network, and objective health; SOC by objective and subjective health; and SES by social network, especially contact with friends, and sensory-motor ability.

Acceptability

Home-based interview administration took between 45 minutes and four hours. Interview participation rate was 70%.

p) Self-Evaluation of Life Function Scale (Linn and Linn, 1984)

The Self-Evaluation of Life Function (SELF) scale was designed to evaluate the physical, emotional, and social function of older people (Linn and Linn, 1984). The developers recommend the instrument for research and screening purposes where a short, comprehensive, and inexpensive self-assessment is needed.

Instrument content was derived from existing scales with some items modified for older people. Following completion by older people recruited from different community and hospital settings in the USA, factor analysis supported item reduction to 54 across six domains: depression (11 items), personal control (four items), physical disability (13 items), self-esteem (seven items), social satisfaction (six items), and symptoms of aging (13 items), as shown in Table 5.1. All items use a four-point categorical response scale. Although not specified, items sum to give six domain scores, where higher scores are a less favourable health state.

One study describes the development and evaluation of the SELF. This included respondents from various hospital, institutional, and community settings within the USA (Linn and Linn, 1984), as shown in Table 3.4.

Reliability

Completion by 101 community-based older people showed moderate to high levels of test-retest reliability (three to five day retest) ranging from 0.59 (for self-esteem) to 0.96 (for physical disability), as shown in Table 5.2.

Validity

(i) Construct validity: socio-demographic variables and health-service use

As hypothesised, the six SELF-domains discriminated between groups defined by their living environment (independent community-dwelling or living in an institution) and medical intervention (outpatient treatment or psychiatric counselling).

(ii) Other types of validity assessment

Following completion by respondents from hospital, institutional, and community settings, the one-year predictive validity of the instrument was assessed. Physical disability, depression, and symptoms of ageing were the most frequent predictors of outcome and were specifically predictive of institutionalisation, number of hospitalisations, and number of visits to a physician. Physical disability and symptoms of ageing were predictors of death. Additional factors influenced the predictive ability of the SELF in different settings. High levels of disability, low self-esteem, and poor social satisfaction predicted days sick in bed. Symptoms of ageing, low self-esteem, and disability were predictors of poor self-reported health.

Responsiveness

A mixed group of respondents completed the SELF twice over a three-month period. The group comprised 90 respondents receiving medical care or counselling, 30 from a housing group who were not also receiving treatment, 22 from a nursing home, and a further 22 sex- and age-matched respondents from the housing group. Patients and health-care providers also rated change in health. As hypothesised, SELF change scores discriminated between patients receiving counselling or medical care who reported improvement and those who reported no improvement, and between patients classified

as experiencing little or no change or improvement by the health-care provider. SELF scores also discriminated between nursing-home residents and those from the housing group.

Acceptability
Completion of the SELF took approximately 15 minutes. Few respondents were unable to read items and less than 5% of the sample had to be reminded about missing items.

q) SENOTS program and battery (Stones and Kozma, 1989)

The SENOTS program and battery was developed as a brief, multidimensional instrument for self-assessment of health by cognitively able older people (Stones and Kozma, 1989). Computer-based administration was included to promote its application and usefulness as both a screening and an evaluative instrument. The SENOTS program is the interactive computer program; the SENOTS battery is the multidimensional assessment instrument.

The SENOTS battery comprises 54 items over five domains, as shown in Table 5.1. Instrument content is derived largely from existing instruments with simplified yes/no responses. The five domains are activity limitation (CARE: activity limitation domain), activity propensity (Memorial University of Newfoundland Activities Inventory [MUNAI] - abbreviated version), financial hardship, happiness/depression (Memorial University of Newfoundland Scale of Happiness [MUNSH]) and physical symptoms (CARE: somatic symptoms domain). Three items were removed due to low item-total correlation.

The instrument may be computer self-administered or interview-administered. With the exception of the MUNSH, the yes/no responses are scored 2 and 1, respectively. Some 'yes' responses within the MUNSH have negative scoring (Stones and Kozma, 1989: see appendix for detail). Item summation gives a score of 6 to 84, where 84 is best health.

One study describes the development and evaluation of the SENOTS. This was a community-based older population in Canada (Stones and Kozma, 1989), as shown in Table 3.4.

Reliability

Internal consistency reliability for each domain was not greatly influenced by mode of administration, as shown below and in Table 5.2. There is no evidence for test-retest reliability.

Table 5q Internal consistency reliability of the SENOTS by mode of administration (Stone and Kozma, 1989)

	Mode of administration	
Domain	***Computer***	***Interview***
Activity propensity	0.76	0.79
Activity limitation	0.88	0.91
Financial hardship	0.66	0.67
Happiness/depression	0.88	0.92
Physical symptoms	0.73	0.78

Validity

(i) Construct validity: socio-demographic variables and health-service use

As hypothesised, the SENOTS discriminated between community-dwelling older people and those living in institutions (worse health).

(ii) Validity: other
Inter-correlation between SENOTS domains ranged from –0.07 (activity propensity with financial hardship and activity propensity with physical symptoms) to 0.55 (happiness/depression with activity limitation).

Acceptability
A participation rate of over 85% has been reported.

r) The Wellness Index (Slivinske et al., 1996)

The Wellness Index was developed in the USA as a self-administered assessment of well-being and health status in the older person (Slivinske et al., 1996). The developers recommend its use in clinical practice, screening, and health policy planning.

Instrument content was informed by literature reviews, existing instruments, reference to the OARS framework, and discussion with health professionals and patients. Six domains of well-being are assessed, namely activities of daily life (ADL) and instrumental ADL (IADL) (13 items), economic resources (ten items), morale (20 items), physical health (12 items), religiosity (11 items), and social resources (13 items), as shown in Table 5.1. Item selection from existing instruments involved consultation with administrators, practitioners, and residents from a range of US settings including nursing homes and senior volunteer programs. Pilot evaluations with nursing-home residents (n=61) supported item content and structure. Moderate to high levels of internal consistency reliability were found (Cronbach's alpha ranged from 0.79 to 0.91) and validity was supported.

The 79 items have five-point Likert response scales (1 - strongly disagree to 5 - strongly agree). Items are summed within each domain, where high scores are better health. Whilst principal components analysis gives a five-component solution (excluding economic resources), six domain scores are reported.

One study describes the development and evaluation of the Wellness Index. This was a community-based older population from various settings in the USA (Slivinske et al., 1996), as shown in Table 3.4.

Reliability

The results of internal consistency and test-retest reliability are shown in Table 5.2. High levels of internal consistency reliability were found, ranging from 0.80 (physical health) to 0.94 (ADL/IADL). Small to moderate levels of test-retest reliability (ten-month retest) were reported following completion by 192 older people, ranging from 0.42 (social resources) to 0.69 (physical health).

Validity

(i) Construct validity: socio-demographic variables and health-service use

The Wellness Index discriminated between groups defined by level of independence (assessed by service provision and professional judgement).

(ii) Other types of validity assessment

Correlations between WI domain scores were in accordance with hypotheses and ranged from 0.02 (economic resources with religiosity) to 0.58 (social resources with morale): see Table 5.3. Correlation between WI index and domain scores ranged from 0.52 (economic resources) to 0.79 (morale).

WI domain scores were compared with a clinical assessment of each domain area, as shown in Table 5r below. Correlation ranged from 0.11 (religiosity) to 0.38 (physical health). The index discriminated between groups defined by physician-assessed levels of well-being.

Table 5r Correlation between the Wellness Index and clinical assessment.

Domain	***Clinical assessment***
ADL-IADL	0.30
Economic resources	0.12
Morale	0.22
Physical health	0.38
Religiosity	0.11
Social resources	0.14

Table 5.1 Older people-specific patient-reported health instruments

Instrument (no. items)	*Domains (no. items)*	*Response options*	*Score*	*Completion (time)*
Brief Screening Questionnaire (BSQ) (26)	ADL (6), Cognitive impairment (1), Financial impact (3), Functional mobility (3), Hearing impairment (1), Mental health (1), Polypharmacy (1), Social contact (2), Symptoms (7), Visual impairment (1)	Categorical: yes/no	Summation Index: 0-26; 26 is worst health	Self or interview
Comprehensive Assessment and Referral Evaluation (CARE) (1500)	*4 core domains:* 1. Psychiatric: self-report/test (252), observation/global (79) 2. Physical/medical/nutritional: self-report (272), observation/global (57) 3. Social needs: self-report (265), observation/global (39) 4. Service needs	Categorical: 2 or 3 options	Summation Index: global overview Narrative summary Domain profile	Interview by trained interviewer
CORE-CARE (329)	*6 domains:* Depression, dementia, disability (activity limitation), subjective memory, sleep, somatic symptoms *4 summary scores - 22 indicator scales* 1. Psychiatric: cognition (10), depression (29), subjective memory (9) 2. Physical: somatic symptoms (34), heart (15), stroke effects (9), cancer (6), respiratory (6), arthritis (9), leg problems (9), sleep (8), hearing (14), vision (11), hypertension (4), ambulation (27), activity limitation (39) 3. Social: finance (8), neighbourhood (8), crime (18), isolation (34), retirement dissatisfaction (7) 4. Service needs: service utility (15)	Categorical: 2 or 3 options	*as above*	*as above*
SHORT-CARE (143)	*6 domains:* Depression, dementia, disability, subjective memory, sleep, somatic symptoms *Diagnostic scales:* Depression, dementia, disability	Categorical: 2 or 3 options	*as above*	*as above*
EASY-Care (up to 85)	General health (19) - includes depression (6): Geriatric depression scale (4), additional items (2); single items include hearing (1), loneliness (1), vision (1), global health (1), communication (1) Disability (17): ADL (6), IADL (11) Memory: cognitive impairment test (6) Home/Safety/Support (14): includes financial concerns Health-care services received (22) Looking after your health (7)	Categorical	Summation 6-domain profile: disability (0-100; 100 is maximum health)	Interview
Functional Assessment Inventory (FAI) ***(not clear: '90 items less than OMFAQ')***	ADL impairment (?), Economic resources (?): occupation and income, Mental health (27): mental health, life satisfaction, self-esteem Physical health (?), Social resources (?) *Additional items:* Socio-demographic, Informant section. Interviewer summary (5 domains)	Categorical; some written answers Interviewer: 6-point categorical	Coding scheme (modified from OMFAQ) 5-domain profile Summary ratings	Interview (mean: 30.6 minutes)
Geriatric Postal Screening Survey (GPSS) (10)	Specific conditions Falls/balance (1), Functional impairment (1), Depression (1), Cognitive impairment (1), Urinary incontinence (1) General health status Health perception (2), Polypharmacy (1), Pain (1), Weight loss (1)	Categorical: yes/no	Summation Index: risk score 0-10; 10 is worst health. >4 is high-risk	Self

Geriatric QoL Questionnaire (GQLQ) (25)	1. ADL (24→8) 2. Symptoms (24→8) 3. Emotional function (9)	7-point categorical	Summation. 3-domain profile: high score is best health	Interview (mean: 30 mins, range: 20-60 mins)
Geriatric Screening Questionnaire (GSQ) (6)	Cognitive impairment, Daily activities, Economic status, General health status, Mental health, Social support	Yes/no	Summation Index: high score is worst health	Interview
IOWA Self-Assessment Inventory (ISAI) Revised (56)	Alienation (8), Anxiety/depression (8), Cognitive status (8), Economic resources (8), Mobility (8), Physical health (8), Social support (8)	4-point categorical	Summation. 7-domain profile 8-56, 56 is best health	Self or interview (preliminary ISAI: median 30-45 mins, revised ISAI: median 15 mins)
LEIPAD (31 + 18)	Cognitive function (5), Depression/anxiety (4), Life satisfaction (6), Physical function (5), Self-care (6), Sexual function (2), Social function (3) Moderator scales (18)	4-point categorical	Summation Index: 0-93, 93 is maximum impairment	Self (15-20 minutes)
OARS Multidimensional Functional Assessment Questionnaire (OMFAQ) **Part A (120)**	*Part A:* ADL (IADL 7) (14), Economic resources (15), Mental health (21), Physical health (16) Social resources (9), Demographic items (11), Informant items (10) Interview section: Interview-specific (4), Interviewer assessments (15), Interview ratings (5) Short Portable Mental Status Questionnaire (10) *Part B:* Services Assessment (24)	Categorical, some written answers Interviewer: 5-point categorical	Summary or coding scheme (algorithm) 5-domain profile Index: Cumulative Impairment Score 5-30, 30 is maximum impairment	Interview (Part A: 30 minutes)
Perceived Well-being Scale (PWB) (14)	Psychological well-being (6), Physical well-being (8) Index: General well-being (14)	7-point Likert scale	Summation 2-domain profile Index: 14-98, 98 is best health	-
PGC Multilevel Assessment Instrument (PGCMAI) ***Full (147)*** ***Mid-length (68)*** ***Short (24)***	ADL (16), Cognition (10), Perceived environment (25), Personal adjustment (12), Physical health (49), Social interaction (17), Time use (18)	Check items Interviewer: 5-point categorical	Summation 7-domain profile Interviewer summary assessment	Interview (full: 50 mins)
Quality of Life Cards (QLC) (80)	Affect (20), Life experience (20), Satisfaction/happiness (40)	Pick cards: +1 for positive –1 for negative	Summation Index: –80 to +80, +80 is best health	Interview
Quality of Life Profile - Seniors Version (QOLPSV) ***Full (111)*** ***Short (54)*** ***Brief (24)***	Being: physical (12), psychological (12), spiritual (12) Belonging: physical (12), social (12), community (12) Becoming: practical (13), leisure (13), growth (13)	5-point categorical: importance, enjoyment	Weighted summation 2-domain profile Index: –3.33 to +3.33	Interview (up to 1 hour)
Quality of life-well-being, meaning and value (QLWMV)	Well-being (5): economic, health status, satisfaction with living area Meaning (43): life purpose, intelligibility, manageability (multiple instruments) Value: self-worth (10)	Categorical	Instrument scores; not clear	Interview (range: 45 minutes to 4 hours)

(>74?)	Health (>12): objective, subjective, sensory-motor (4) Functional capacity (>10): ADL External factors: living area, housing, accommodation, family, social contact (n?)			
Self-evaluation of Life (SELF) Scale (54)	Depression (11), Personal control (4), Physical disability (13), Self-esteem (7), Social satisfaction (6), Symptoms of ageing (13)	4-point categorical	Summation 6-domain profile: high score is worse health	Self (approx 15 mins)
SENOTS program and battery (54)	Activity limitation (7), Activity propensity (12), Financial hardship (4), Happiness/depression (24), Physical symptoms (7)	Yes/no	Summation Index: 6-84, 84 is best health	Self or interview
Wellness Index (WI) (79)	ADL/IADL (13), Economic resources (10), Morale (20), Physical health (12), Religiosity (11), Social resources (13)	5-point Likert scale	Summation 6-domain profile, high score is better health	Self

Table 5.2 Reliability of older people-specific instruments

Instrument	*Cronbach's alpha*	*Test-retest correlation [retest period]*
CARE	0.72 (retirement dissatisfaction) to 0.95 (ambulation problems and activity limitation)[1]	-
CORE-CARE	*Indicator scales:* range 0.28 (service utility) to 0.92 (vision disorder) (16 >0.70, 1 >0.90)[1] *Indicator scales:* Psychiatric range 0.84 (cognitive impairment) to 0.87 (depression); Medical/Physical range 0.78 (arthritis) to 0.95 (activity limitations, ambulation) (3 >0.90, vision 0.91); Service needs 0.70; Social needs range 0.73 (neighbourhood) to 0.83 (crime)[2]	Inter-rater (n=2) 0.70 to 0.80[1]
SHORT-CARE	*Diagnostic scales:* 0.64 dementia, 0.75 depression, 0.84 disability[3]	*Diagnostic scales:* Inter-rater 0.78 (disability), 0.82 (depression), 0.88 (dementia)[4] *Diagnostic scales:* Inter-rater (n=13) 0.76 (dementia), 0.91 (disability), 0.94 (depression)[3]
EASY-Care	-	0.04 (cognitive impairment) to 0.82 (stairs) (4 <0.40, 7 >0.70) [2 weeks] Total disability score 0.87 [2 weeks][5]
FAI	-	Inter-rater 0.16 (economic resources) to 0.81 (Short Portable Mental Status Questionnaire) [3-5 weeks][6] Inter-rater 0.53 (mental health) to 0.78 (social resources) [1 week][7]
GPSS	-	Index 0.86 [3 weeks]; items (not listed): range 0.48 to 0.92 [3 weeks] Kappa agreement between risk ratings 0.76 (88.5% agreement)[8]
GSQ	-	Items (not listed): range 0.60– 0.86 [2 weeks][9]
ISAI	*Preliminary ISAI:* Well elderly 0.70 (ADL) to 0.82 (economic resources) Homebound 0.74 (physical health) to 0.92 (cognitive status) (1 >0.90)[10] 0.78 (social resources) to 0.87 (cognitive status)[11] *Revised ISAI:* 0.74 (alienation) to 0.86 (economic resources)[12]	-
LEIPAD	0.43 sexual function; 0.61 life satisfaction, social function; 0.74 self-care, physical function; 0.78 depression/anxiety; 0.79 cognitive function (4 >0.70)[13]	0.81 [2 weeks][13]
OMFAQ	0.68 (IADL);[36] 0.92 (ADL), 0.91 (IADL)[14]	-
PGCMAI	*Short:* 0.04 (social interaction) to 0.87 (ADL) (1 >0.70)[15] *Mid-length:* 0.29 (social interaction) to 0.90 (ADL) (1 >0.90, 3 >0.70)[15] *Full:* 0.71 (time use) to 0.93 (ADL) (1 >0.90; all >0.70)[15]	0.73 (social interaction) to 0.95 (physical health) [3 weeks][15]
PWB	0.91 index (0.78 physical WB, 0.82 psychological WB)[16]	Index 0.78, physical WB 0.65, psychological WB 0.79 [2 years][16] Index 0.60 [4 weeks][17]
QLC	-	0.99 [3 days][18]
QLPSV	*Full:* 0.92-0.98;[19,20] 'all domains and sub-domains >0.90'[21] *Short:* 0.83-0.95[19,20] *Brief:* 0.73-0.92[19,20] *Brief - Time 1:* range 0.47 (Belonging-community) to 0.78 (Being-psychological) (5 >0.70, 4 <0.70; 0.56 Becoming-leisure, 0.64 Becoming-growth and Belonging-physical)	-

	Brief - Time 2: range 0.60 (Belonging-physical) to 0.82 (Becoming-leisure) (7 >0.70; 0.62 Belonging-community)[22]	
QLWMV	0.75 (Self-worth: self-esteem scale) to 0.86 (Meaning: purpose in life)[23]	-
SELF	-	range: 0.59 (self-esteem) to 0.96 (physical disability) [3-5 days]; 5 >0.70, 2 >0.90; 0.79 (personal control), 0.81 (social satisfaction), 0.84 (depression), 0.93 (symptoms of ageing)[24]
SENOTS	*Computer:* 0.66 (financial hardship) to 0.88 (activity limitation, happiness/depression) (4 >0.70) *Interview*: 0.67 (financial hardship) to 0.92 (happiness/depression) (4 >0.70, 2 >0.90)[25]	-
WI	0.80 (physical health), 0.82 (morale), 0.87 (social resources), 0.89 (economic resources), 0.91 (religiosity), 0.94 (ADL/IADL)[26]	0.42 social resources; 0.44 morale; 0.50 ADL/IADL; 0.66 economic resources, religiosity; 0.69 physical health (10 days)[26]

References

1 Golden et al. (1984)
2 Teresi et al. (1984b)
3 Gurland et al. (1984)
4 Teresi et al. (1984a)
5 Philp et al. (2002)
6 Cairl et al. (1996)
7 Robinson et al. (1986)
8 Alessi et al. (2003)
9 Fernandez-Buergo et al. (2002)
10 Morris and Buckwalter (1988)
11 Morris et al. (1989)
12 Morris et al. (1990)
13 De Leo et al. (1998)
14 Breithaupt and McDowell (2001)
15 Lawton et al. (1982)
16 Reker and Wong (1984)
17 Cousins (1997)
18 Rai et al. (1995)
19 Raphael et al. (1995a)
20 Raphael et al. (1995b)
21 Raphael et al. (1997)
22 Irvine et al. (2000)
23 Särvimaki and Stenbock-Hult (2000)
24 Linn and Linn (1984)
25 Stones and Kozma (1989)
26 Slivinski et al. (1996)
27 Smeeth et al. (2001)
28 Bath et al. (2000)
29 Pfeiffer et al. (1981)
30 Pfeiffer et al. (1989)
31 Guyatt et al. (1993b)
32 Condello et al. (2003)
33 Fillenbaum et al. (1985)
34 Fillenbaum and Smyer (1981)
35 Carver et al. (1999)
36 Reuben et al. (1995)
37 McCusker et al. (1999)
38 Stadnyk et al. (1998)
39 Harel and Deimling (1984)
40 Osborne et al. (2003)
41 Wissing and Unosson (2002)

Table 5.3 Validity of older people-specific instruments *(see Table 5.2 for references)*

Instrument	*Socio-demographic variables and health-service use*	*Patient-reported health instruments*
BSQ	*Screening:* high specificity (>90%), low sensitivity (<50%) therefore caution when screening for vision or hearing impairment, depression, cognitive problems[27]	-
CARE	[h] Activity limitation and cognitive impairment: low scores predict family inconvenience and decision to institutionalise, high scores predict families not inconvenienced and deciding not to institutionalise[4]	[h] CARE depression with cognitive impairment 0.12, Global Diagnostic Rating (GDR) 0.75[2] [h] CARE medical conditions with Family Informant Scale (FIS): range 0.45 (arthritis and hypertension) to 0.70 (ADL)[2] [h] CARE service needs (activity limitation, ambulation) with FIS ambulation 0.62, GDR & FIS activity limitation 0.70[2] [h] CARE social needs with GDR: range 0.61 (crime) to 0.64 (finances)[2] [h] CARE indicator scales with GDR: range 0.40 (service needs with total service utilization) to 0.75 (psychiatric disorders with depression)[2] [h] CARE indicator scales with FIS: range 0.30 (service needs with family service provision) to 0.70 (service needs with activity limitation)[2] [h] CARE items with FIS depression 0.33 (psychiatric domain), sleep disorder 0.36 (physical disorder), social isolation problems 0.41 (environmental/social problems)[2]
CORE-CARE	Cognitive and functional impairment, older age, male sex strongest predictors of death at one year. Activity limitation, cognitive impairment, age strongest predictors of service utilization[4]	[h] Indicator scales: activity limitation (AL) with ambulation problems (AM) 0.78 Total service utilisation with AL 0.58 and AM 0.60. Arthritis with AM 0.40. Somatic symptoms (SS) with respiratory symptoms 0.54 and AM 0.51. Depression with sleep disorder 0.55, SS 0.54, AL 0.50[1]
SHORT-CARE	-	Clinician diagnosis with depression and dementia scales: agreement 10 out of 26 (no disorder); clinician diagnosis with psychiatric problems: agreement 12 out of 16[3] SHORT-CARE diagnosed dementia: observed outcomes match expected outcomes[3]
EASY-Care	Levels of deprivation* (Townsend Scores and Under-privileged area scores)[28]	-
FAI	Four settings: ADL strongest predictor, economic resources weakest predictor of impairment [h] Nursing home: greatest impairment all domains; congregate living facilities: highly impaired ADL, mental health, social resources; day-care/senior centres: less impaired all domains[29] [h] Older people living in institutions (worse health) *vs* those attending senior centers and well older people* [30]	FAI with OMFAQ: range 0.27 (economic and social resources) to 0.86 (short psychiatric evaluation)[6] FAI domains: range 0.32 (mental health with physical health) to 0.58 (mental health with ADL)[30]
GPSS	Co-morbidity and health service use* [8] High sensitivity & specificity for risk of falls, depression, urinary incontinence; limited sensitivity and specificity for functional impairment (ADL), cognitive impairment[8]	Groups defined by the GPSS as high- or low-risk: Geriatric Depression Score, short Orientation Memory Cognition test, SF-36 health perception discriminated between groups[8]
GQLQ	-	Change score (12 months): [h] ADL range 0.30 (Rand physical function) to 0.41 (Barthel Index), emotional function range 0.44 (global) to 0.61 (Rand emotional function)[31]
GSQ	*23-item confirmation or exclusion test:* sensitivity 50% or 88%, specificity 89% or 40% *6-item confirmation or exclusion test:* sensitivity 58% or 81%, specificity 89% or 56%[9]	-

ISAI	*Preliminary ISAI:* [h] Social resources, physical health, ADL discriminated between relatively fit or attending a meal program, and homebound or receiving home-delivered meals[10] Does not discriminate groups defined by sex, age, educational level, or living arrangement[10] [h] Domains discriminated between groups defined by: income (economic resources [ER]), age (ER, ADL, cognitive status), education (ER, social resource [SR], mental [MH], physical health [PH], ADL, cognitive status), living arrangements (MH)[11] [h] SR, ER, MH with education level range 0.21-0.27 ADL with age –0.32, ER with income 0.36[11]	*Preliminary ISAI:* [h] Well elderly: ER with MH 0.50, PH with ADL 0.54, SR with MH 0.55, SR with PH 0.57, MH with PH 0.63[10] [h] Homebound elderly: MH with PH 0.57; cognitive status with ADL 0.55, PH 0.66, MH 0.71; PH with ADL 0.70[10] ISAI domains: range 0.19 (cognitive status with ER) to 0.59 (PH with ADL)[11] *Revised ISAI:* [h] Domains range 0.04 (alienation with mobility) to 0.89 (anxiety with mental health)[12]
LEIPAD	Personality disorders*** [32]	[h] Domains with Rotterdam Questionnaire (RQ): 0.10 (sexual function with RQ psychological distress), 0.13 (sexual function with RQ physical distress), 0.14 (social function with RQ ADL), 0.70 (physical function with RQ physical distress), 0.71 (depression/anxiety with RQ psychological distress), 0.79 (self-care with RQ ADL)[13] Index with personality disorders (diagnostic scale) range 0.35 (Passive-aggressive) to 0.68 (Depressive)[32]
OMFAQ	Inverse relationship between IADL scores and survival status at one year[33] Interviewer-rated summary with clinically assessed criteria: mental health 0.67, economic resources 0.68, physical health 0.82, self-care 0.89[34]	OMFAQ index [h] With SF-20 role function 0.56, physical function 0.67[35] *OMFAQ ADL/IADL domains:* [h] ADL with OMFAQ domains: social resource 0.11, economic resources 0.19, mental health 0.54, physical health 0.60[33] ADL with SF-36: range physical functioning 0.36 to role-physical 0.49[36] [h] ADL/IADL summary with Functional Autonomy Measurement System (SMAF) 0.80[37] range: ADL with SMAF-communication 0.31, to IADL with SMAF-IADL 0.77[37] IADL with Physical Performance Test 0.56[36] IADL with Functional Status Questionnaire (FSQ): IADL 0.59, FSQ ADL 0.70[36] [h] IADL with SF-20 role function 0.56, physical function 0.67[35] IADL with SF-36 physical function >0.60[38] *OMFAQ other domains:* Social Resource (SR) with self-rated mental health range 0.08 to 0.31[39] SR with Minnesota Multiphasic Personality Inventory (MMPI): range –0.05 to –0.46[39] SR with interviewer-rated mental health range 0.04 to 0.40[39] Social attachments and social interaction explained 13% of self-rated mental health, 31% of MMPI and interviewer-rated mental health[39] [h] Self-care with AQoL (generic utility) –0.68, with AQoL-independent living –0.82[40] [h] OMFAQ with AQoL domains: range 0.03 (Independent living with AQoL social resources) to –0.40 (Social relationships with AQoL self-care)[40]
PWBS	[h] Community-dwelling versus institutionalised older people[16] [h] Older women defined by age, level of exercise[17]	[h] PWB index with Personal Optimism 0.40[16] With self-reported physical health: psychological WB 0.06, index –0.25, physical WB –0.40; with the Memorial University of Newfoundland Scale of Happiness [MUNSH]:

		physical WB 0.52, psychological WB 0.69, index 0.70[16] [h] PWB index with medication intake –0.39, self-rated global health –0.39, symptoms –0.51[17]
PGCMAI	[h]Residential status: range 0.05 (perceived environment) to 0.54 (ADL)[15] Leg ulcers (presence/absence): mid-length PGCMAI domains of physical health, ADL, cognition, use of time, social behaviour, personal adjustment, quality of the environment* [41]	[h] Respondent with interviewer summary: range 0.36 (perceived environment) to 0.87 (ADL); respondent with independent clinician ratings: range 0.23 (cognition) to 0.65 (physical health)[15] Sub-domain items: range 0.09 (cognitive symptoms) to 0.78 (psychiatric symptoms)[15] Sub-domain items with domain scores: range 0.19 (perceived environment) to 0.87 (ADL)[15]
QLC	-	Delighted-Terrible scale –0.96; visual analogue scale 0.93; index with scores for affect, life experience, satisfaction/happiness domains: range 0.90-0.97[18]
QOLPSV	[h] Small correlation between quality of life scores and socio-demographic variables (income, education, age) - not as hypothesised[19,20,21] *QOLPSV brief:* [h] Discriminates between groups defined by level of nursing care required: Becoming-practical (–0.40), Being-physical (–0.43), Being-spiritual (–0.36), Belonging-physical (–0.46), Belonging-social (–0.50)[22]	With self-reported health status: range 0.37 (Belonging-social) to 0.57 (Being-physical)[19.20,21] [h] With the MUNSH, range Being-physical 0.15 to Belonging-community 0.62; with the Life Satisfaction Scale, range Being-physical 0.19 to Belonging-social 0.37; with the Social Health Battery, range Becoming-growth 0.30 to Belonging-community 0.62; with the National Council on Aging Activity Questionnaire, range Being-psychological 0.11 to Becoming-leisure 0.62[19.20,21] [h] QOLPSV full with QOLPSV short 0.95-0.99 and QOLPSV brief 0.88-0.98 [19,20,21]
QLWMV	External factors (social network) with other QLWMV instruments: range 0.16 (family network with Sense of Coherence scale [SOC]) to 0.25 (social network with Self-esteem Scale [SES])[23]	Instruments within different domains: range 0.19 (ADL ladder with SOC) to 0.62 (SES with Purpose in Life PIL)[23] PIL best explained by ADL, family network, objective health; SOC by objective and subjective health; SES by social network, especially contact with friends, and sensory-motor ability[23]
SELF	Living environment (independent community-dwelling or institutionalised) and medical intervention (outpatient treatment or psychiatric counselling)*** [24] Physical disability, depression, symptoms of aging most frequent predictors of outcome at 1 year, specifically institutionalisation, hospitalisation, visits to physician[24] Physical disability and symptoms of aging predictors of death at 1 year[24] Symptoms of aging, low self-esteem, disability predictors of poor self-reported health[24]	-
SENOTS	[h]Institutionalised elderly (worse health) versus community-dwelling adults[25]	SENOTS domains: range 0.07 (activity propensity with financial hardship, activity propensity with physical symptoms) to 0.55 (happiness/depression with activity limitation)[25]
WI	Independence level (service provision, health professional assessment), levels of well-being (physician assessed)[26]	Domains with clinical assessment: range 0.11 (religiosity) to 0.38 (physical health)[26] [h] Domain scores: range 0.02 (economic resource with religiosity) to 0.58 (social resource with morale)[26] Domain scores with index: range 0.52 (economic resources) to 0.79 (morale)[26]

Key:
[h] = hypothesis supported by correlation
levels of statistical significance: * = p <0.05; ** = p <0.01; *** = p <0.001

Chapter 6: SUMMARY - GENERIC INSTRUMENTS

a) Search strategy

112 articles provided some evidence of measurement and practical properties for the generic instruments included in the review.

b) Patient-reported health instruments

The 15 generic instruments reviewed are listed in Table 6.1. The SF-36, SIP, EuroQol, and NHP have undergone the highest number of evaluations following application in the assessment of older people with 67, ten, ten, and eight published articles, respectively (Table 6.1). A combined total of 11 articles provide evidence for the COOP charts; five of these relate to the WONCA/COOP. The populations in which these instruments have been evaluated are summarised in Table 6.2.

c) Patient and study characteristics

The number of respondents included in the studies ranged from ten (Tidermark et al., 2003b) to 3,263 (Cleary and Jette, 2000): see Table 3.2. A wide age-range was also covered (mean age-range: 64.0 to 86.0 years): see Table 6.2. The evaluations were conducted in a range of settings, including the community, primary care, hospital, day-care, and residential institutions (see Table 6.2). Several studies concern older people with specific co-morbidity, for example, hip fracture (Tidermark et al., 2002a,b), chronic heart disease (Jenkinson et al., 1997; Baldassarre et al., 2002), and stroke (Anderson et al., 1996). Respondents with cognitive impairment were included in only four studies; three of these evaluated the SF-36 (McHorney et al., 1990, cited by McHorney, 1996; Parker et al., 1998; Seymour et al., 2001); one evaluated the SF-12 and HSQ-12 (Petitt et al., 2001).

28 evaluations were performed in the UK, 56 in North America, 19 in Europe, seven in Australia, and two in Japan (Table 6.1, 6.2). 27 studies describe the specific concurrent evaluation of instruments: 19 generic with generic instruments, seven generic with older-person specific instruments, and five generic with disease-specific instruments (see Chapter 8 and Table 8).

d) Description of instruments

The domains covered by the 15 instruments are shown in Table 6.2. When compared using the criteria described by Fitzpatrick et al. (1998), all instruments are multidimensional with between two (QWB) and six (COOP charts, EuroQol, HSQ-12, SF-12, SF-20, SF-36, SIP) domains (Table 6.2). The domain of physical function is included in all instruments. Psychological and social well-being is assessed by all instruments except the QWB; social well-being is assessed by all except the IHQL. 11 instruments assess symptoms, for example, pain and discomfort. There is variation in the content of the remaining instruments domains but eight assess general health. Only three instruments, namely GQL, SIP, and SQL (modified), assess cognitive function.

The shortest instruments have five items (EQ-5D and SQL), the longest has 136 (SIP). With the exception of the QWB, all instruments produce a score profile across domains.

Eight instruments, namely the AQoL, EQ-5D, GQL, IHQL, QLI, QWB, SIP, and SQL, also produce an index score; four instruments, namely the HSQ-12, SF-12, SF-36, and SIP, produce summary scores. The AQoL, EuroQol (EQ-5D), and QWB incorporate utilities or values attached to health states.

Seven instruments were both self-completed by, or interview-administered to, older people, namely the COOP charts, EuroQol, FSQ, NHP, SF-12, SF-36, and SIP. Three instruments, namely the AQoL, GQL, QLI, were self-completed only. The remainder, namel HSQ-12, IHRQL, SF-20, and SQL, were interview-administered only. Proxy completion of the WONCA/COOP charts (Van Balen et al., 2001, 2003), EuroQol (Tamim et al., 2002), NHP (Van Balen et al., 2001, 2003), SF-36 (Pierre et al., 1998; Ball et al., 2001; Yip et al., 2001), and the SIP (Page et al., 1995) was reported. However, the impact of proxy completion was evaluated for the EuroQol and SF-36 only.

e) Reliability

Evidence of reliability is shown in Table 4.2 and summarised in Table 6.1. Internal consistency reliability is reported for nine instruments. The WONCA/COOP charts, EuroQol, IHQL, and the QWB are not amenable to tests of internal consistency. Values for Cronbach's alpha reported for all studies evaluating the AQoL (utility, independent living), GQL, NHP (emotional reactions, physical mobility), QLI, SF-12, and SIP (index, body care and movement, mobility), several studies evaluating the FSQ, and most studies evaluating the SF-36 exceed 0.70, the criterion recommended for studies involving groups of patients (Streiner and Norman, 1995).

The AQoL domains of physical senses, psychological well-being, and social relationships had unacceptable alpha values (Osborne et al., 2003). Two studies reported unacceptable levels of alpha for the FSQ domains of quality of social interaction and ADL (Yarnold et al., 1995; Sherman and Reuben, 1998). A low alpha value was reported for the NHP social isolation domain (Van Balen et al., 2003). Several studies reported low alpha values for SF-36 social function (Table 4.2). Two studies found low alpha values for the SIP domains of sleep and rest, and eating (Rothman et al., 1989; Andresen et al., 1998).

Six instruments have evidence of test-retest reliability, namely the COOP charts, EuroQol, NHP, SF-12, SF-20, and SF-36 (Table 4.2). The EuroQol was the only instrument that did not perform satisfactorily but the lengthy test-retest period of six months means that the results must be interpreted with caution (Brazier et al., 1996). The SF-36 has the greatest evidence for test-retest reliability. In most of these studies, reliability exceeded the criteria necessary for the assessment of groups. Lower levels of reliability were consistently reported for the social function and role-emotional domains.

There was no evidence of reliability for the HSQ-12, IHQL, QWB scale, and SQL index.

Several studies report evidence of data quality at item level for the NHP and SF-36 following completion by older people. Detail is limited for the NHP, but the one published evaluation suggests item-total correlations greater than 0.40 for all domains (Sharples et al., 2000). Most evaluations of the SF-36 report item-total correlations

greater than 0.40. Completion by young-old respondents with depression (Beusterien et al., 1996) resulted in a high Response Consistency Index (RCI); however, interview administration to frail old-old respondents resulted in a lower RCI (Stadnyk et al., 1998).

f) Validity

Patients were not involved in the construction of the AQoL, COOP charts, EuroQoL, FSQ, GQL, HSQ-12, IHQL, QLI, QWB scale, SF-12, SF-20, or the SF-36. Rather, item generation was informed by the literature and existing instruments. The AQoL and COOP charts also specifically included clinicians.

Patients and the lay population were involved in item generation for the NHP, the SIP, and a modified version of the SQL; for the SIP and the modified SQL, health professionals were also involved. However, it is not clear whether older people (aged over 65 years) were included in this process. Modifications to the SQL included the addition of cognition and personal environment domains, modification to the activities of daily life domain to reflect the needs of geriatric assessment, and altered terminology to enhance applicability (Stolee et al., 1996; Stadnyk et al., 1998).

The content validity of generic instruments for older people has not been widely evaluated. However, the omission of memory and cognitive function from the SF-20, and the combining of several activities with different functional demands, reduced the appropriateness of the instrument for assessing older people (Carver et al., 1999).

All instruments have undergone some form of validity testing as shown in Table 4.3 and summarised in Table 6.1.

Internal validity
Four instruments have undergone internal validation using factor or principal component analysis (PCA) to assess dimensionality. Factor analysis in both the general and older populations supported the proposed domain structure of the AQoL (Osbourne et al., 2003). Confirmatory analysis of the SF-12 produced a two-factor solution, but supported a revised model where item 10 (vitality) loaded on physical health (but not mental health), and item 12 (social time) loaded on both mental and physical health in calculation of the summary scores (Resnick and Nahm, 2001).

Four factors were found for the SF-20 (Carver et al., 1999). One general health item ('I have been feeling bad lately') grouped on one factor with all the mental health domain items; the remaining general health items grouped onto a second factor. Physical function and role function items loaded across two additional factors but did not describe domains entirely consistent with the SF-20. Following completion by groups of young-old (Dexter et al., 1996; Wolinsky and Stump, 1996) and frail old-old (Stadnyk et al., 1998), factor analyses of the SF-36 supported the two-factor solution of mental and physical health and the eight-domain structure proposed by instrument developers. Further analyses produced a nine-factor model; the additional factor 'health optimism' included two general health items: 'getting ill' (item 11a) and 'getting worse' (item 11c) (Wolinsky and Stump, 1996).

Factor analysis is not appropriate for the COOP charts, EuroQol, IHQL, and QWB scale, and was not performed in an older population for the FSQ, GQL, HSQ-12, NHP, QLI, SIP, and SQL.

Other instruments and global judgements of health
Further tests of validity included correlations with other instruments and global judgements of health (see Table 4.3). With the exception of the QLI, all instruments have undergone some form of comparison with other patient-reported instruments, the results of which are summarised in Table 6.1.

Several studies hypothesised expected correlations between instrument scores and external variables, highlighted in Table 4.3. However, the hypothesised correlations were often poorly defined and the size of expected correlation was rarely reported.

The AQoL utility and domain scores had correlations of the expected size and direction with scores for a domain-specific and a generic instrument (Osborne et al., 2003). The largest correlation was between physical function domains.

The COOP charts had correlations in the hypothesised direction with scores for several generic instruments (Nelson, 1990; Coast et al., 1998; Van Balen et al., 2003). Evidence suggests that the charts are sensitive to the impact of illness or trauma, with chronic illness most strongly associated with reductions in physical function.

Limited evidence supported hypothesised correlations between the EuroQol and both generic and domain-specific instruments (Coast et al., 1999). Evidence suggests that the index score is sensitive to the impact of trauma, and discriminates between groups defined by a range of variables including pain, mobility, and fracture severity (Tidermark et al., 2002a,b; 2003a).

The FSQ (IADL and ADL domains) had correlations in the hypothesised directions with SF-12 and symptom-specific scores in cardiac patients (Cleary and Jette, 2000). Moderate to strong correlations were reported with physical performance assessments and self-report instruments (Reuben et al., 1995).

There was very limited evidence for the validity of the GQL in older people; the strongest reported correlation was with the Beck Depression Inventory (Andersson et al., 1995).

Regression analysis demonstrated that the HSQ-12 domains of mental health, role-mental, social function, bodily pain, and energy explained 57% of the variance in the SHORT-CARE depression score, whilst the physical function, social function, and energy domains explained 68% of score variance of the SHORT-CARE activities of daily living subscale (Pettit et al., 2001).

The small correlations between the IHQL and domains of the SHORT-CARE instrument did not support hypothesised associations (Livingstone et al., 1998). The authors conclude that the IHQL has limited usefulness in the assessment of older people.

Accumulated evidence supported hypothesised correlations between NHP domains and both generic and domain-specific instruments (Stadnyk et al., 1998; Sharples et al.,

2000; Van Balen et al., 2001, 2003). Evidence suggests that various domains are sensitive to the impact of trauma, and discriminate between groups defined by a wide range of variables, including fitness level and musculoskeletal morbidity (Hunt et al., 1980; Thorsen et al., 1995), chronic illness, depression, anxiety, and pain (Sharples et al., 2000).

The QWB and QWB-SA scales had small to moderate correlations in the hypothesised directions with domains from the SIP and SF-36, and scores for several symptom-specific instruments (Andresen et al., 1995, 1998b). Small to moderate correlations were reported with physical performance assessments (De Bon et al., 1995) and self-report assessments of activity levels (Andresen et al., 1998b).

As hypothesised, the SF-12 MCS explains greater variation in the SHORT-CARE depression scales (Gurland et al., 1984) than the PCS, and the PCS explains greater variation in ADL limitation (Pettit et al., 2001).

The SF-20 had small to large correlations in the hypothesised direction with a range of domain-specific instruments (Carver et al., 1999).

Evidence supported most hypothesised correlations between SF-36 domains and summary scores, and both generic instruments and older people-specific instruments covering a wide range of domains, demonstrating both convergent and divergent validity. However, some correlations for the physical function domains were smaller than hypothesised (Bombardier et al., 1995).

The SIP index, domain scores, and summary score had small to moderate correlations in the hypothesised direction with domain scores from the SF-36 and QWB scale, and scores for several domain-specific instruments (Rothman et al., 1989; Andresen et al., 1995; Andresen et al., 1998b). A strong correlation between the two summary scores was found (Rothman et al., 1989).

The SQL had moderate to large correlations in the hypothesised direction with a range of domain-specific instruments (Carver et al., 1999). The moderate to strong correlation with several SF-36 domains was hypothesised, although a smaller than hypothesised correlation between social function domains was found (Stadnyk et al., 1998).

Proxy completion

Agreement between patients and caregiver proxies for the more observable EQ-5D items or activities, for example, mobility, was greater than agreement between for the more subjective items, for example, depression; agreement improved over time (Tamim et al., 2002).

Strong levels of agreement were found between cognitively intact older people and lay proxies regarding scores for the more observable SF-36 health domains, for example, physical function (PF), role-physical, and general health (Pierre et al., 1998; Yip et al., 2001); moderate levels of agreement were found for the remaining SF-36 domains.

Professional proxies scored lower than cognitively intact older people on all SF-36 domains except bodily pain (BP) and mental health (MH); lay proxies scored lower than patients on all domains (Ball et al., 2001). Difference in agreement between professional and lay proxy completers was statistically significant for PF, BP, and MH.

Following completion of the Functional Independence Measure (FIM) and the SF-36 (PF), evidence suggests that informed professionals are better able to interpret patient health status than patient-nominated lay proxies.

Socio-demographic variables and health-service use
With the exception of the WONCA/COOP, GQL, IHQL, and SQL, all instruments have been compared with socio-demographic variables and health-service use (see Table 4.3). In community-dwelling adults, lower scores on the AQoL were predictive of increased health-care use at 18 months (Osborne et al., 2003). Low scores on COOP emotional condition and overall health charts were predictive of future placement in nursing care and hospitalisation, respectively, for those living in residential homes (Siu et al., 1993b).

The EuroQol (EQ-5D and thermometer) discriminated between groups defined by the number of GP visits, hospital inpatient stays, limiting long-term illness, and level of disability (Brazier et al., 1996). Consensus is lacking with regard to the ability of the EuroQol to discriminate between groups defined by age. Evidence suggests that the FSQ does not discriminate between adults defined by age.

Most domains of the HSQ-12 discriminated between groups defined by age, and all domains by self-reported illness, depression, and limitation in activities of daily life (Bowling and Windsor, 1997; Petitt et al., 2001). Several domains discriminated between groups defined by receipt of health services, impaired vision or hearing, and psychiatric difficulties (Petitt et al., 2001).

The NHP discriminated between groups defined by the number of GP consultations; several domains discriminated between groups defined by marital status, sex, long-standing illness, and disability (Hunt et al., 1980). However, the NHP did not discriminate between groups defined by social class, age, or living status. Scores on the QWB did not discriminate between groups defined by age or sex.

The SF-12 discriminated between groups defined by a range of variables, including use of health and social services, self-reported health, number of chronic illnesses, and level of regular exercise. One study reported both summary and domain scores; domain scores discriminated between groups defined by age (Schofield and Mishra, 1998). A further study reported group discrimination by age for physical health, but not for mental health (Lim and Fisher, 1999).

Where all SF-20 domains, except mental health (MH), discriminated between the general population and older people, MH was the only domain to discriminate groups defined by sex (Carver et al., 1999). In those living in residential homes, low scores on the SF-20 general health and MH domains were predictive of future hospitalisation and placement in nursing care, respectively (Siu et al., 1993b).

Evidence suggests that the SF-36 is sensitive to the impact of different health states, discriminating between a range of socio-demographic features or health-related variables (Table 4.3). Overall evidence suggests a decline in health with age as indicated by scores for physical function, role-physical, and vitality domains. However, many studies also suggest constant or better mental health scores, and often vitality, general health, and social function scores, in older age-groups compared to younger populations. The majority of studies suggest that women report worse levels of health

than men across all domains. Accumulated evidence supports the ability of all or most domains to discriminate between different health states, including long-standing illness or disability and self-reported health, and levels of disease severity. Multiple studies support the ability of specific domains to discriminate between different levels of health-service use including GP and hospital appointments, and need for care. Following completion by the chronically ill, most SF-36 domains, particularly bodily pain, general health, and vitality, were predictive of GP and hospital appointments. Physical function, role-physical, and bodily pain were predictive of hospitalisation (McHorney, 1996). General health and physical function were predictive of mortality. Mental health domains were least predictive in all settings.

Evidence suggests that the SIP, particularly the physical activity domains, is sensitive to the impact of old age (Rothman et al., 1989; Kleipell and Ferrans, 2002). In a single study, the SIP-68 mobility domain had high sensitivity for poor function (91%), low specificity for good function (58%), and discriminated between recurrent fallers and non-fallers (Jannink-Nijlant et al., 1999). The score was predictive of the risk of recurrent falling.

g) Responsiveness

Evidence suggests that most instruments are capable of measuring some change in health, as summarised in Table 6.1. There is no evidence of responsiveness for the GQL, HSQ-12, IHQL, QLI, or QWB scale, and limited evidence for the SIP. The most extensive evidence, across a range of settings, relates to the SF-36.

The ability to discriminate between treatment groups over time was reported for seven instruments, the exceptions being the GQL, SF-12, SIP, and SQL. ES statistics were reported for the COOP charts, EuroQol, NHP, SF-20, and SF-36. Correlation of change scores with change in other variables was reported for the COOP charts, EuroQol, SF-12, SF-20, and SF-36. Although statistical significance was frequently reported, the clinical significance of change scores was rarely addressed.

Where health deterioration in community-dwelling adults was defined by hospitalisation or admission to institutionalised care, limited evidence suggests that the AQoL is more responsive to change over 18 months than the SF-36 and the OARS Multidimensional Functional Assessment Questionnaire (OMFAQ), an older-people specific assessment of functional and general status (see Chapter 5) (Osborne et al., 2003). AQoL baseline score differences discriminated between people who were hospitalised or remained in the community at follow-up.

Small ES were found for the COOP physical function (PF) chart following three months of residential care (Siu et al., 1993b). Small to large correlations between change scores for the COOP charts and the SF-20 were found. However, the COOP PF chart was unable to discriminate better than chance on change in performance-based tests. Small to moderate ES were reported following the management of congestive heart failure (Jenkinson et al., 1997). Moderate to large ES and group discrimination were reported following the surgical repair of hip fracture (Van Balen et al., 2003).

As hypothesised, greater and more rapid improvement in EuroQol scores over four months were reported for patients receiving a total knee replacement than for those suffering from stroke (Coast et al., 1998). In addition, large ES and group

discrimination were reported for the EQ-5D four months after the surgical repair of hip fracture (Tidermark et al., 2003a). Largest change score correlations in the same patient group were reported with the SF-36 domains bodily pain, vitality, and physical function. However, following a trial of cardiac rehabilitation, limited evidence suggested poor responsiveness and no group discrimination for the EQ-5D (Hage et al., 2003). Hypothetical improvements in the use of health resources, age, and health status were associated with small to large ES (Brazier et al., 1996).

Statistically significant score change in the FSQ and group discrimination was reported following the long-term assessment of patients who did, or did not, undergo heart balloon valvuloplasty (Tedesco et al., 1990).

Small to moderate ES were reported for the NHP following the rehabilitation of frail older people with mainly medical conditions (Stadnyk et al., 1998). Small to large effect sizes (ES) were reported following the surgical repair of hip fracture (Van Balen et al., 2001, 2003). Despite a general score improvement across domains, only the energy domain discriminated between groups defined by type of rehabilitation exercise following hip fracture (Mitchell et al., 2001).

A moderate correlation between change scores for the SF-12 physical component summary score (PCS) and the Western Ontario MacMaster Osteoarthritis (WOMAC) questionnaire domains of functional ability, pain, and stiffness was reported following completion by older people receiving drug therapy for moderate to severe osteoarthritis of the knee (Theiler et al., 2002). Improvement in SF-12 PCS was statistically significant, but improvement in the mental component summary score was not.

Following three months of residential care, deterioration or improvement in function was associated with small ES for the SF-20 physical function domain (Siu et al., 1993b). Comparable levels of responsiveness were reported for the SF-20 and COOP physical function domains. The SF-20 physical function domain discriminated better than chance for deterioration in balance and gait.

The SF-36 showed limited responsiveness following the evaluation of community-based continence and mental health services (Hill et al., 1996), the longitudinal evaluation of people with chronic debilitating disease (Wolinsky et al., 1998), and the rehabilitation of frail older people (Stadnyk et al., 1998). In the latter study, domain-specific instruments were more responsive than two generic instruments (SF-36, NHP).

Small to strong ES and group discrimination were reported for the SF-36 following application in two exercise-based trials: a community-based exercise programme (Cochrane et al., 1998) and a six-month cardiac rehabilitation programme (Seki et al., 2003). The highest levels of responsiveness were found for the role physical, general health, and bodily pain domains. For individuals reporting an improvement in depression over six weeks, with the exception of physical function, all domains showed improvement (Beursterien et al., 1996). Score improvement over six months was associated with an improvement according to clinical judgement in the health status of day-hospital patients (Fowler et al., 2000). For the same patients, small to moderate change score correlations were found between the SF-36 and other instruments. Finally, high levels of responsiveness were reported for the PCS, physical function, and general health domains following an 18-month care co-ordination trial (Osborne et al., 2003). Baseline score differences for the physical function, bodily pain, and vitality domains

discriminated between people who were hospitalised or remained in the community at follow-up.

Moderate to strong levels of responsiveness and group discrimination were reported for six SF-36 domains following a meta-analysis of drug trials for osteoarthritis (Lisse et al., 2001) and most domains following a placebo-controlled trial in diabetes (Reza et al., 2001). Small to moderate ES were reported following four weeks of treatment for congestive heart failure (Jenkinson et al., 1997). At four months post-hip fracture repair, strong levels of responsiveness were found for the physical function and bodily pain domains; seven domains discriminated between patients whose improvement was good or poor (Tidermark et al., 2003a). In the same patient group, the strongest change score correlations were reported between the physical function, bodily pain, and vitality domains and the EuroQol.

At six months after surgery for coronary heart disease, statistically significant improvements in SIP index and summary scores were found (Page et al., 1995). Small mean improvements in SIP index and physical health summary scores for patients receiving home-modification advice did not reach clinical or statistical significance over time or discriminate between groups of patients not receiving advice (Liddle et al., 1996).

Large ES were found for the SQL and other domain-specific instruments following a rehabilitation programme for frail older people (Stadnyk et al., 1998). Score reduction following four weeks of rehabilitation after a hip fracture repair did not reach statistical significance or discriminate between groups (Simpson, 2002).

h) Precision

Although ceiling effects may be expected to reduce with age (McHorney, 1996; Ware 1997), it appears that the AQoL (social relationships, physical senses), COOP (daily activities, physical function), FSQ (ADL, IADL), HSQ-12 (several domains), NHP (all domains), SF-20 (all domains), SF-36 (role limitation, social function), and SIP do not discriminate between groups with low levels of morbidity, because of ceiling effects.

The older population generally has more sickness than the general population, which led the SF-36 developers to hypothesise that data quality may be weaker. Floor effects were reported for the role limitation domains. Floor effects have also been reported for several domains within the COOP, SF-20, and the SIP.

i) Acceptability

Completion rates ranged from 75% (IHQL) to 100% (COOP charts and NHP) for interview administration, and from 43% (SIP) to 95% (NHP) for self-completion. Completion rates were not reported for the AQoL, GQL, or SQL. Mean completion times for interview administration ranged from ten minutes (NHP) to 35 minutes (SIP). Self-completion times were frequently not reported; SF-36 self-completion with supervision had a mean completion time of 12.5 minutes (sd 5.5) (Wood Dauphinee et al., 1997).

Instrument completion rates varied with mode of administration, but were generally higher following interview administration than self-completion (for example, Hayes et al., 1995; Parker et al., 1998). Age and administration mode were found to have an independent and statistically significant association with completion rates (Hayes et al., 1995; Parker et al., 1998). For the most extensively studied instruments, evidence suggests that completion difficulties increase with age, declining cognitive ability, and deteriorating health status. Several authors have suggested that self-completion of the SF-36 may be inappropriate for the old-old (Lyons et al., 1994; Parker et al., 1998). This may be the case for most patient-reported health instruments (Hayes et al., 1998).

Issues of acceptability have been extensively studied for the SF-36 and arise mainly in relation to work items, items related to vigorous activity, and repetition of physical activity items (for example, Hayes et al., 1995; Dexter et al., 1996; Parker et al., 1998). Cautious interpretation of the role limitation and social function domains has been advised due to the lack of participation in certain activities expressed by older respondents (Fowler et al., 2000). Difficulty in completing items related to health outlook within the general health domain has been reported in several evaluations (for example, Hayes et al., 1995; Mallinson, 1998; Sharples et al., 2000). The length of question stems associated with specific items in both the SF-12 and SF-36 has caused difficulty for some respondents (Wood Dauphinee et al., 1997; Iglesias et al., 2001). Frequently omitted SIP items relate to sexual activity and interaction with children (Andresen et al., 1998a,b). Where assessed, similarly high levels of patient-reported satisfaction or acceptability have been reported for the HSQ-12 and SF-12 (Petit et al., 2001), and the SF-36 and SIP (Andresen et al., 1998a,b).

j) Instrument evaluations in UK settings

28 articles describe the evaluation of seven instruments in the UK, as summarised in Table 6.1. The most extensively evaluated instrument was the SF-36 (20 articles). The NHP was evaluated in three articles (Hunt et al., 1980; Sharples et al., 2000; Mitchell et al., 2001). Evaluations of the WONCA/COOP charts (Coast et al., 1998; Philp et al., 2001), EuroQol (Brazier et al., 1996; Coast et al., 1998), HSQ-12 (Bowling and Windsor, 1997; Pettit et al., 2001), and SF-12 (Iglesias et al., 2001; Pettit et al., 2001) were each described in two articles; the COOP charts (Jenkinson et al., 1997) and IHQL (Livingstone et al., 1998) were each described in one article.

Table 6.1 Summary of generic instruments: measurement properties

Instrument	*Evaluations (n)*[a]		*Reliability*[b]		*Validity*[b]		*Responsiveness*[b]	
	Total	**UK**	**Thoroughness**	**Results**	**Thoroughness**	**Results**	**Thoroughness**	**Results**
Assessment of Quality of Life instrument (AQoL)	1	0	+	+	++	+/++	+/++	+/++
COOP Charts for Primary Care Practice	6	1	+	+	++	++	+/++	+/++
WONCA/COOP	5	2	+	+	++	++	+/++	+/++
European Quality of Life Questionnaire (EuroQol)	10	2	+	+	++	++	++	++
Functional Status Questionnaire (FSQ)	6	0	+/++	+/++	+/++	+/++	+/++	+/++
Goteborg Quality of Life instrument (GQL)	2	0	+	+	+	+	0	0
Health Status Questionnaire-12 (HSQ-12)	2	2	0	0	+/++	+/++	0	0
Index of Health-related Quality of life (IHQL)	1	1	0	0	+	+	0	0
Nottingham Health Profile (NHP)	8	3	+/++	+/++	+++	++/+++	++	++
Quality of Life Index (QLI)	1	0	+	+	+	+	0	0
Quality of Well-being Scale (QWB)	4	0	0	0	++	++	0	0
Short Form 12-item Health Survey (SF-12)	7	2	+	+	+	+	+	+
Short Form 20-item Health Survey (SF-20)	4	0	+	+	+/++	+/++	+/++	+/++
Short Form 36-item Health Survey (SF-36)	67	20	+++	+++	+++	+++	+++	+++
Sickness Impact Profile (SIP)	10	0	+	+	+++	++/+++	+	+
Spitzer Quality of Life (SQL)	3	0	0	0	+	+	+/++	+/++

[a] Number of evaluations in the older population (aged >60 years)
[b] After McDowell & Newell, 1996: evidence for measurement properties (as Table 2.3)

Thoroughness

0	No reported evidence of testing
+	Basic information only
++	Several types of tests, or several studies reporting evidence
+++	All major forms of evaluation reported; several good quality studies

Results

0	No numerical results reported
+	Weak evidence
++	Adequate evidence
+++	Good evidence

Table 6.2 Summary of generic instruments: health status domains and evaluative settings with older populations

Instrument (no. items)	*Settings and country*	*Mean age in yrs/range*	*Instrument domains (after Fitzpatrick et al., 1998)*							
			Physical function	**Symptoms**	**Global judgement**	**Psychol. well-being**	**Social well-being**	**Cognitive functioning**	**Role activities**	**Personal construct**
AQoL (12-15)	Community Australia	>60.0	x	x		x	x			
COOP (9) WONCA (6)	Community, hospital clinics, inpatients, primary care, residential and nursing homes USA, UK, Netherlands	60.0-84.3	x	x	x	x	x		x	
EuroQol (5+1)	Community, primary care, hospital inpatients, post-surgery (orthopaedics), cardiac rehabilitation UK, Sweden, Italy, Canada	60.0-80.1	x	x	x	x	x		x	
FSQ (34)	Community, surgical intervention, primary care USA	69.0-78.0	x		x	x	x		x	
GQL (15)	Community Sweden	69.9-78.0	x	x		x	x	x		
HSQ-12 (12)	Community UK	65.0-74.0	x	x	x	x	x		x	
IHQL (44)	Community UK	75.7	x	x		x				
NHP (38)	Community, primary care, hospital clinics, surgical intervention UK, Denmark, Canada, Netherlands	68.0-83.0	x	x		x	x			
QLI (64)	Community USA	73.7	x			x	x		x	x
QWB (11)	Community, residential homes USA	72.0-80.0	x				x			
SF-12 (12)	Community, hospital clinics UK, USA, Switzerland	70.0-86.0	x	x	x	x	x		x	
SF-20 (20)	Community, residential homes USA, Canada	76.0-84.0	x	x	x	x	x		x	
SF-36 (36)	Community, nursing/retirement homes, surgical intervention, hospital inpatients, day-hospitals, primary care, drug trials, rehabilitation US, UK, Canada, Australia, Sweden, Japan	64.0-82.0	x	x	x	x	x		x	
SIP (136)	Primary care, nursing homes, hospital clinics USA Australia, Netherlands	64.0-82.0	x	x		x	x	x	x	
SQL (5)	Community, rehabilitation Canada	60.0->80.0	x		x	x	x			

Chapter 7: SUMMARY - OLDER PEOPLE-SPECIFIC INSTRUMENTS

a) Search strategy

46 articles provided some evidence of measurement and practical properties for the older people-specific instruments included in the review.

b) Patient-reported health instruments

The 18 older people-specific instruments that were reviewed are listed in Table 7.1. The OARS Multidimensional Functional Assessment Questionnaire (OMFAQ), Comprehensive Assessment and Referral Evaluation (CARE), Functional Assessment Inventory (FAI), and Quality of Life Profile - Senior Version (QOLPSV) have undergone the largest number of evaluations with 12, five, four, and four published articles, respectively (Table 7.2). However, all of the articles describing the evaluation of the CARE instrument (Gurland et al., 1977; Teresi et al., 1984a,b; Golden et al., 1984), and three of the articles describing the QOLPSV (Raphael et al., 1995a,b, 1997) refer to the same study population. The EASY-Care, ISAI, and PGCMAI have three published evaluations. The majority of instruments have been evaluated in single studies (see Table 7.1).

c) Patient and study characteristics

The populations in which these instruments have been evaluated are summarised in Table 7.2. The number of respondents included in the studies ranged from 18 (Cousins, 1997) to over 5,000 (Smeeth et al., 2001). The age ranged from a mean of 60.0 years to 83.5 years. The evaluations were conducted in community, primary care, nursing home, and hospital settings (Table 7.2).

Although respondents with cognitive impairment were specifically excluded from several studies (for example, Guyatt et al., 1993b; Philp et al., 2002), the majority of studies do not report respondent cognitive status. Several studies report the use of proxy completion for cognitively impaired people (FAI: Pfeiffer et al., 1981, 1989; OMFAQ: Breithaupt and McDowell, 1996).

Only the CARE, EASY-Care, and BSQ have published UK evaluations, with five, four, and one article, respectively (see Table 7.1). The majority of evaluations were in the United States (USA) (18) and Canada (13).

Eight studies describe the specific concurrent evaluation of older people-specific instruments (Cairl et al., 1983) and older people-specific with generic instruments (Guyatt et al., 1993b; Reuben et al., 1995; Stadnyk et al., 1998; Irvine et al., 2000; Jaglal et al., 2000; Philp et al., 2001; Osborne et al., 2003): see Chapter 8 and Table 8.

d) Description of instruments

The domains covered by the instruments are shown in Table 7.2. With the exception of the Quality of Life Cards (personal construct only), when compared using the criteria described by Fitzpatrick et al. (1998), all instruments are multidimensional with

between two (Perceived Well-being Scale) and nine domains (EASY-Care): see Table 7.2. Most include physical function, psychological, and social well-being. There is variation in the content of the remaining instruments, but nine also assess symptoms. Eight instruments, namely the BSQ, EASY-Care, GSQ, GPSS, ISAI, LEIPAD, OMFAQ, and PGCMAI, assess cognitive functioning . Three, namely EASY-Care, GPSS, and GSQ, include global judgements of health.

The shortest instrument has six items (GSQ), the longest has 1500 (CARE). Several shortened versions of instruments have been developed, including the SHORT-CARE (143 items) and QOLPSV (short version: 54 items, brief version: 27 items). Ten instruments, namely CARE, EASY-Care, FAI, GQLQ, ISAI, OMFAQ, PWB, MAI, SELF, and WI, produce a score profile across all domains. Nine instruments, namely the BSQ, GPSS, GSQ, LEIPAD, OMFAQ, PWB, QLC, QOLPSV, and SENOTS, have index scores. The OMFAQ and PWB produce both profile and index scores. Clarity of the scoring procedure is lacking for the BSQ and QLWMV.

Ten instruments, namely CARE, EASY-Care, FAI, GQLQ, GSQ, LEIPAD, OMFAQ, PGCMAI, QLC, and QLWMV, require interview administration, with reported completion times ranging from 15 minutes (LEIPAD) to four hours (QLWMV; range: 45 minutes to four hours). The BSQ, QOLPSV, SELF, SENOTS, and WI may be interview- or self-administered. The GPSS and ISAI are designed for self-administration, with a reported mean completion time of 15 minutes for both. Method of administration was not reported for the PWB.

e) Reliability

Evidence for reliability is shown in Table 5.2 and summarised in Table 7.1. Internal consistency is reported for ten instruments but not for the BSQ, EASY-Care, FAI, GPSS, GSQ, QLC, and SELF. Tests of internal consistency are inappropriate for the GQLQ. Levels of Cronbach's alpha exceeding 0.70, the criterion recommended for studies involving groups of patients (Streiner and Norman, 1995), are reported for all studies evaluating the ISAI (all domains), the LEIPAD (four domains), the PWB (both domains and index), the PGCMAI (full length: all domains, mid-length: three domains, short version: one domain), the QLWMV (all domains), and the WI (all domains). Levels of Cronbach's alpha exceeding 0.70 are reported for several studies evaluating the CARE, SHORT-CARE diagnostic scales, and the QOLPSV (brief version: five domains).

Alpha levels exceeding 0.90, the criterion recommended in the assessment of individual patients (Fitzpatrick et al., 1998), were reported for several domains within the CARE and CORE-CARE indicator scales, ISAI (cognitive status), OMFAQ (ADL, IADL), PGCMAI (mid- and full-length ADL), PWB (index), QOLPSV (all versions), SENOTS (activity limitation, depression/happiness), and the WI (ADL/IADL).

Lower alpha values were reported for several domains within the CARE (service utility 0.28), LEIPAD (sexual function 0.43), short and mid-length versions of the PGCMAI (social interaction less than 0.30), the brief version of the QOLPSV (Belonging-community and Belonging-physical less than 0.60), and the SENOTS battery (financial hardship 0.66).

Eight instruments, namely EASY-Care, GPSS, LEIPAD, PWB, MAI, QLC, SELF, and WI, have limited evidence for test-retest reliability with retest periods ranging from three days to two years: see Table 5.2. All domains within the PGCMAI, individual domains within EASY-Care (disability score 0.87) and the SELF (physical disability 0.96), the GPSS (index 0.86; individual items), and the index score for the LEIPAD and QLC exceed 0.70. Two domains within the SELF and the QLC have levels of reliability greater than 0.90. Several instrument domains have low levels of reliability that do not support their application in the assessment of groups, namely EASY-Care communication, telephone, feeding, and cognitive impairment (less than 0.40), PWB index (0.60), SELF self-esteem (0.59), and WI social resources (0.42). Few authors indicate whether reliability is assessed in people reporting no change in health over the retest period.

Only the LEIPAD, PWB, MAI, and WI have evidence of both internal consistency and test-retest reliability. Evidence of reliability is lacking for the BSQ and GQLQ. The FAI has evidence of inter-observer reliability in the range 0.16 (economic resources) (Cairl et al., 1983) to 0.78 (social resources) (Robinson et al., 1986).

f) Validity

Patients were involved in item generation for the GQLQ, QLPSV, and WI. Early versions of the OMFAQ were piloted with patient groups. The EASY-Care, GPSS, GQLQ, ISAI, LEIPAD, OMFAQ, and WI incorporated the expert opinion of health professionals. The PWB used psychology students as a resource and the GSQ used a survey of risk factors for poor health in older people. The literature and existing instruments provided the main source of items for the remaining instruments. The FAI, ISAI, MAI, and WI drew heavily on the content of the OMFAQ.

The content validity of instruments is not widely reported. The ISAI used health professionals to confirm item content and readability.

All instruments have undergone some form of validity testing as shown in Table 5.3 and summarised in Table 7.1.

Internal validity
Six instruments have undergone internal validation using factor or principal component analysis to assess dimensionality. Factor analysis supported the proposed domain structure of the revised ISAI, LEIPAD, OMFAQ, PWB, SELF, and WI.

The seven-domain structure of the revised ISAI was informed by factor analysis of the preliminary ISAI (Morris et al., 1990). Two factors were found for the LEIPAD: three items grouped on the psychosocial factor, two items grouped on the physical factor (Dr Leo et al., 1998). The remaining two items did not clearly group onto either factor. Factor analysis supported the proposed domain structure of the OMFAQ (George and Fillenbaum, 1985; Harel and Deimling, 1984) and the PWB (Reker and Wong, 1984). During instrument development, factor analysis of the SELF supported item reduction to 54 items across six domains (Linn and Linn, 1984). Although a five-domain solution was described for the WI, a six-domain score is used (Slivinski et al., 1996).

Factor analysis is not appropriate for the GQLQ and was not performed for the remaining instruments.

Modern psychometric methods have been used to assess the internal validity of the OMFAQ ADL domain (Breithaupt and McDowell, 2001). The results describe the two ADL and IADL sub-scales, and suggest that each measures most precisely at different functional levels: ADL is most precise at lower functional levels, while the IADL is most precise at higher levels. The results support the independent use of the two sub-scales.

Other instruments and global judgements of health
Further tests of validity included correlations with other instruments and global judgements of health (see Table 5.3). With the exception of the BSQ, EASY-Care, GSQ, and SELF, all instruments have undergone some form of comparison with other patient-reported instruments, the results of which are detailed in Table 5.3 and summarised in Table 7.1. The few studies that proposed expected associations between individual domains or index scores and external variables are shown in Table 5.3.

The CARE domains had correlations in the hypothesised direction with scores for several domain- and symptom-specific instruments (Teresi et al., 1984b; Golden et al., 1984). The correlation between the CARE instrument and well-established generic or other older people-specific instruments has not been assessed. However, agreement between SHORT-CARE defined psychiatric problems and clinician-assessed problems have been reported (Gurland et al., 1984).

Small to strong correlations between the FAI and its parent instrument, the OMFAQ, (Cairl et al., 1983), and small to moderate correlations between FAI domains (Pfeiffer et al., 1989) were found.

Moderate to strong change score correlations were found between the GQLQ and several domain-specific instruments following a one-year clinical trial of day-hospital versus conventional care for frail older people (Guyatt et al., 1993b).

Moderate to strong levels of domain correlation were found for the preliminary ISAI (Morris and Buckwalter, 1988), but smaller correlations were reported for the revised format (Morris et al., 1990).

The LEIPAD had small to large correlations in the hypothesised direction with several domain-specific instruments (De Leo et al., 1998; Condello et al., 2003).

Accumulated evidence supported the majority of hypothesised correlations between the OMFAQ index and domain scores, and several domain-specific and generic instruments covering a range of domains (Harel and Deimling, 1984; Fillenbaum, 1985; Reuben et al., 1995; McCusker et al., 1999; Stadnyk et al., 1998; Carver et al., 1999; Osborne et al., 2003). Moderate to strong correlations between interviewer-assessed OMFAQ ratings and clinician-assessed criteria were found (Fillenbaum and Smyer, 1981). Most evidence supported the validity of the ADL and IADL domains.

The PWB index and domain scores had small to large correlations in the hypothesised directions with several domain-specific instruments (Reker and Wong, 1984; Cousins, 1997). Small to large correlations between respondent-completed PGCMAI and interviewer summaries or clinician ratings, and between sub-domain items and domain total scores were found (Lawton et al., 1982).

Strong correlations between QLC domains, and between the QLC and the Delighted-Terrible scale and a visual analogue scale were found (Rai et al., 1995). The QOLPSV had small to large correlations in the hypothesised directions with several domain-specific instruments and with self-reported health status (Raphael et al. 1995a,b, 1997). The strong correlations between the different versions of the instrument were hypothesised.

Only the inter-relationship between different instruments within the QLWMV has been reported, and small to large correlations have been found (Sarvimäki and Stenbock-Hult, 2000). Small to moderate correlations between SENOTS domains were found (Stones and Kozma, 1989). Small to moderate correlations between WI domain scores, and moderate to large correlations between domain and index scores were found (Slivinski et al., 1996).

Sensitivity and specificity
High specificity (over 90% for all domains) following interview or self-completion supports the diagnostic accuracy of the BSQ, but low sensitivity (less than 50% for all domains) suggests it should be used with caution when screening for poor vision, hearing impairment, depression, and cognitive problems (Smeeth et al., 2001). The GPSS has high sensitivity and specificity for risk of falls, depression, and urinary incontinence, but limited accuracy for the evaluation of functional and cognitive impairment (Alessi et al., 2003). Where the GPSS discriminated between groups defined by levels of co-morbidity and health-service use, groups defined as high- or low-risk on the GPSS were discriminated by scores for several domain-specific and generic instruments. The GSQ had a sensitivity of 65% and a specificity of 98% when used as a confirmation test for positive or negative health in a comprehensive geriatric assessment (Fernandez Buergo et al., 2002).

Socio-demographic variables and health-service use
With the exception of the BSQ, GQLQ, GSQ, LEIPAD, and QLC, all instruments have been compared with socio-demographic variables and health-service use (see Table 5.3). In community-dwelling adults, the CARE activity limitation and cognitive impairment domains were predictive of a family's perceived level of inconvenience and decision to institutionalise a relative (Teresi et al., 1984a). Older age, and cognitive and functional impairment were strong predictors of service utilisation and death after one year.

The EASY-Care discriminated between groups defined by levels of deprivation (defined by the Townsend Score) (Bath et al., 2000). The FAI discriminated between groups defined by service setting and health-care utilisation, where activities of daily living were the strongest predictors of impairment (Pfeiffer et al., 1981, 1989).

The GPSS discriminated between groups defined by levels of co-morbidity and health-service use; groups defined as high- or low-risk on the GPSS were discriminated by scores for several domain-specific and generic instruments (Alessi et al., 2003).

As hypothesised, the ISAI (preliminary version) discriminated between groups defined by level of fitness and dependency for home-care (Morris Buckwalter, 1988), income, age, education, and living arrangements (Morris et al., 1989). However, small correlations with income, age, and education were reported.

The LEIPAD discriminated between groups with diagnosed personality disorders (Condello et al., 2003). The OMFAQ IADL scores were predictive of survival status at one year (Fillenbaum et al., 1985). The PWB discriminated between groups of community-dwelling older people and those living in institutions (Reker and Wong, 1984), and between women defined by age and level of exercise (Cousins, 1997).

The (mid-length) PGCMAI discriminated between groups defined by the presence of leg ulcers (Wissing and Unosson, 2002). Small to moderate correlations between PGCMAI domains and residential status were found (Lawton et al., 1982).

Where several domains of the (brief) QOLPSV discriminated between groups of patients defined by the level of nursing care required (Irvine et al., 2000), correlation between the (full) QOLPSV scores and socio-demographic variables were smaller than hypothesised (Raphael et al. 1995a,b, 1997).

The SELF discriminated between groups defined by their living environment and medical interventions (Linn and Linn, 1984). SELF physical disability, depression, and symptoms of aging were the most frequent predictors of institutionalisation, number of hospitalisations, and physician visits after one year. SELF physical disability and symptoms of aging were predictors of death.

As hypothesised, the SENOTS discriminated between community-dwelling older people and those living in institutions (Stones and Kozma, 1989). The WI discriminated between groups defined by level of independence (Slivinski et al., 1996).

g) Responsiveness

Limited evidence of responsiveness was found for five instruments, namely the GQLQ, OMFAQ, PGCMAI, QLPSV, and SELF.

Small to moderate levels of responsiveness were found for the GQLQ following a one-year trial of day-care in frail older people (Guyatt et al., 1989). These values were comparable to other domain-specific instruments.

Evidence suggests that the OMFAQ detects change over time as a result of life events and in response to receipt of services (George and Fillenbaum, 1985). Although less responsive than the SF-36 and a condition-specific instrument, moderate to large effect sizes (ES) and group discrimination was reported for the OMFAQ physical function domain following surgical repair of hip fracture (Jaglal et al., 2000). Low levels of responsiveness were found following an 18-month care co-ordination trial, where deterioration was defined by hospitalisation (Osborne et al., 2003). Baseline score differences (self-care and social resources domains only) discriminated between people who were hospitalised or remained community-dwelling at follow-up.

As hypothesised, a statistically significant improvement in several QOLPSV domains was found following nursing agency care, but the instrument was less responsive than the SF-36 (Irvine et al., 2000). Moreover, score change did not discriminate between acute, chronic, or palliative care patients, or groups defined by number of nurse-visits.

SELF change scores discriminated between patients receiving counselling or medical care who reported improvement and those who reported no improvement, or were assessed as not having improved by the health-care provider (Linn and Linn, 1984).

h) Precision

Response distribution was reported for the OMFAQ ADL and IADL items only and ceiling effects were identified (Reuben et al., 1995; Breithaupt and McDowell, 2001). Floor effects have not been reported.

i) Acceptability

Although frequently not reported, completion rates were generally higher with interview administration (often more than 85%) than with postal self-completion. Several studies reported high completion rates with self-completion. The overall response rates for the BSQ were higher with postal self-completion (83.5%) than with interview administration (mean range 73.9% to 75.9%), but when groups were defined by age, older age-groups had lower self-completion rates (Smeeth et al., 2001). Furthermore, 21% of postal responders required help with completion and the proportion of missing or invalid responses was higher in this group. The QOLPSV had the lowest reported self-completion rate (67%) (Raphael et al. 1995a, 1995b, 1997). High participation rates have been reported for both proxy (81%) (Breithaupt and McDowell, 1996) and respondent completion of the OMFAQ (Reuben et al., 1995).

Interview-administered instruments had associated increased completion times (range 30 minutes for GQLQ and OMFAQ, to four hours for QLWMV) when compared to self-completion instruments (ISAI, LEIPAD, and SELF: range 15-20 minutes).

Despite the high response rate (90%), non-responders to the GPSS were contacted by telephone (Alessi et al., 2003). Many non-responders were classified as being at high risk for health impairment, and the authors suggest that persuasive methods to increase response rates, for example, telephone contact and home-visits, are required when using questionnaires for screening purposes.

Evidence of acceptability is lacking for the PWB, QLC, and WI.

j) Instrument evaluations in UK settings

Ten articles describe the evaluation of three instruments in the UK, as summarised in Table 7.1. The most extensively evaluated instruments are the CARE and EASY-Care. However, five CARE publications relate to the same patient population.

Table 7.1 Summary of older people-specific instruments: measurement properties

Instrument	*Evaluations (n)*[a]		*Reliability*[b]		*Validity*[b]		*Responsiveness*[b]	
	Total	**UK**	**Thoroughness**	**Results**	**Thoroughness**	**Results**	**Thoroughness**	**Results**
Brief Screening Questionnaire (BSQ)	1	1	0	0	++	+/++	0	0
Comprehensive Assessment and Referral Evaluation (CARE)	1 (6)[c]	1 (6)[c]	+	+	++	++	0	0
CORE-CARE	3	3	++	++	++	++	0	0
SHORT-CARE	2	2	++	++	++	++	0	0
EASY-Care	3	3	+	+	+/++	+/++	0	0
Functional Assessment Inventory (FAI)	4	0	+	+	+	+	0	0
Geriatric Postal Screening Survey (GPSS)	1	0	+	+	+/++	+/++	0	0
Geriatric Quality of Life Questionnaire (GQLQ)	1	0	0	0	+	+	++	++
Geriatric Screening Questionnaire (GSQ)	1	0	+	+	+++	++/+++	0	0
IOWA Self-Assessment Inventory (ISAI)	3	0	+	++	+	+	0	0
LEIPAD	2	0	++	++	++	++	0	0
OARS Multidimensional Functional Assessment Questionnaire (OMFAQ)	12	0	+	++	+	+	+	+
Perceived Well-being Scale (PWB)	2	0	++	++	+/++	+/++	0	0
Philadelphia Geriatric Centre Multilevel Assessment Instrument (PGCMAI)	3	0	++	++	+++	+++	+	+
Quality of Life Cards (QLC)	1	0	+	+	+++	++/+++	0	0
Quality of Life Profile: Seniors Version (QOLPSV)	4[d]	0	+	+	+	+	++	+
Quality of life - well-being, meaning and value (QLWMV)	1	0	+	+			0	0
Self-evaluation of Life (SELF) Scale	1	0	+	+			+	+
SENOTS program and battery	1	0	+	+			0	0
Wellness Index (WI)	1	0	++	+			0	0

[a] Number of evaluations in the older population (aged >60 years)
[b] After McDowell & Newell, 1996: see Tables 2.3 and 6.1
[c] Five evaluations for the CARE instruments refer to the same patient population
[d] Three evaluations refer to the same patient population

Table 7.2 Summary of older people-specific instruments: health status domains and evaluative settings

Instrument (no. items)	*Settings and country*	*Mean age/ age-range in yrs*	*Instrument domains (after Fitzpatrick et al., 1998)*								
			Physical function	**Symptoms**	**Global judgement**	**Psychol. well-being**	**Social well-being**	**Cognitive functioning**	**Role activities**	**Personal construct**	**Treatment satisfaction**
BSQ (26)	Community UK	>75	x	x		x	x	x	x		
CARE (1500)	Community USA, UK	>65	x	x		x	x				
CORE-CARE (329)	Community USA, UK	>65	x	x		x	x				
SHORT-CARE (143)	Community USA, UK	>65	x	x		x	x				
EASY-Care (up to 85)	Community; primary care; rehabilitation units UK	75.0-81.0	x	x	x	x	x	x	x	x	x
FAI (not clear)	Nursing homes; community; institutions; primary care USA	60.0-83.0	x			x	x		x		
GPSS (10)	Community USA	>65	x	x		x		x			
GQLQ (25)	Day-hospital; outpatients Canada	79.6-78.2	x		x	x	x	x			
GSQ (6)	Community Spain	>65	x	x	x	x		x			
ISAI Revised (56)	Community; meals programmes; home-care USA	75-79	x			x	x	x			
LEIPAD (31)	Community Italy, Netherlands, Finland	>60	x			x	x	x		x	
OMFAQ Part A (120)	Community; emergency care; primary care USA, Canada	60.0-77.0	x			x	x	x			
PWB (14)	Community Canada	68.0	x			x					

PGCMAI (147)	Community; primary care USA, Sweden	74.0-82.0	x			x	x	x			
QLC (80)	Day-hospital Netherlands	79.0-83.5				x				x	
QOLPSV (111)	Community; primary care Canada	61.0-73.0	x			x	x			x	
QLWMV (>74)	Community Finland	75.0-97.0	x	x		x	x			x	
SELF (54)	Community; hospital in- and outpatients; nursing homes USA	70.4	x	x		x				x	
SENOTS (57)	Community: institutions Canada	77.8	x	x		x	x				
WI (79)	Community, nursing homes USA	73.4	x			x	x			x	

Chapter 8: DISCUSSION AND RECOMMENDATIONS

The application of patient-reported health instruments in the evaluation of health-care has become increasingly important (Garratt et al., 2002a), specifically in the assessment of older people (NSF-OP, 2001). This review has identified an increase in the number of instrument evaluations and applications with older people, particularly since 2000. Older people are a growing and diverse population within society, with the highest demand for health- and social care (NSF-OP, 2001). It is therefore important to determine how old age, and associated management and treatment programmes, affect health from the perspective of the older person, and how best to assess health for screening, monitoring, and evaluative purposes.

a) Quantity of HRQL assessment in older people

There has been an exponential growth in the availability of patient-reported health instruments over the last decade with the result that there are many instruments from which to choose for assessment purposes (Garratt et al., 2002a). This growth has been greatest within the specialities of cancer, rheumatology, and musculoskeletal medicine. A huge growth has also been observed in the field of gerontology.

There are two broad approaches to measuring health outcomes from the perspective of the older person, namely generic instruments which aim to cover aspects of health and quality of life relevant to the general population, and older people-specific instruments which aim to cover aspects of health relevant specifically to the older population. This review has focused on evaluations of generic and older people-specific measures of HRQL used in the assessment of older people (aged over 60 years). 15 generic and 18 older people-specific, patient-reported multidimensional measures of HRQL were reviewed.

The majority of instruments have been developed and evaluated in older populations in the United States (USA) and Canada. The SF-36 was by far the most widely evaluated generic instrument. The OMFAQ was the most widely evaluated specific instrument. The majority of older people-specific instruments have just one published evaluation of their measurement properties. Seven generic and three older people-specific instruments have been evaluated in UK populations; the SF-36 and older people-specific CARE and EASY-Care are the most widely evaluated in the UK. The CARE was co-developed in the USA and UK (Gurland et al., 1977), and the EASY-Care in the UK and other European countries (Philp et al., 1997).

b) HRQL in older people - instrument selection

The most extensive evidence was found for the SF-36. The generic EuroQol, SIP, COOP, NHP, SF-12, and the older people-specific OMFAQ have also been widely evaluated; except for the SIP and OMFAQ, these instruments have been evaluated in UK populations. There was much less evidence for the remaining instruments.

Instrument description

When selecting a patient-reported instrument, the appropriateness of item content, relationship to the proposed application and population group, and level of respondent

and clinician/researcher burden in terms of time, cost, and feasibility of application should be considered (Patrick and Erickson, 1993; Fitzpatrick et al., 1998).

The shortest generic instruments were the EuroQol and SQL (five items); the longest was the SIP (136 items). Except for the QWB and SF-12, all produce score profiles. The FSQ, SF-12, SF-36, and SIP have summary scores. The GQL, QLI, SIP, and SQL also produce index scores. The AQoL, IHQL, QWB, and EuroQoL produce scores that include values for health states in the form of utilities.

The shortest and longest specific instruments were the GSQ (six items) and the CARE (1500 items), respectively. 11 instruments have profile and ten have index scores; the CARE, OMFAQ, PWB, and QOLPSV produce both profile and index scores.

All the generic instruments are multidimensional, with between three (GQL, IHQL) and twelve (SIP) domains. When domain coverage is compared using the domains described by Fitzpatrick et al. (1998), instruments assess between four (NHP, SQL) and six domains (COOP, EuroQol, HSQ-12, SF-12, SF-20, SF-36, SIP). Except for the IHQL and QWB, physical function, psychological well-being, and social well-being are common to all instruments. The GQL and SIP are the only generic instruments to assess cognitive function. Although few of the articles reviewed specifically evaluated the validity of instrument content for older people, several generic instruments are criticised for omitting the assessment of memory and cognitive ability, and for inappropriately combining items which address a range of physical activities. The COOP, EuroQol (EQ-thermometer), FSQ, SF-12, and SQL, include single item domains for the global judgement of overall health, which may limit the ability to record the influence of different factors on health, and may influence interpretation (Fitzpatrick et al., 1998). The SF-20 and SF-36 provide more detailed assessments of general health.

All older people-specific instruments are multidimensional, with between two (PWB) and ten (BSQ) domains. When domain coverage is compared using the domains described by Fitzpatrick et al. (1998), all instruments assess between two (PWB, QLC) and nine (EASY-Care) domains. Physical function, psychological well-being, and social well-being were assessed by the majority of instruments. More of the specific instruments (BSQ, EASY-Care, GSQ, GPSS, ISAI, LEIPAD, OMFAQ, PGCMAI) assessed cognitive ability than did the generic instruments. The OMFAQ has been criticised for combining cognitive ability and psychological well-being in a single domain (mental health) (Morris et al., 1989). There was considerable variation in the spread of items across the remainder of instrument domains.

Undue length may limit the scope for application of several instruments, for example, the original CARE contains up to 1500 items. The 85-item EASY-Care combines comprehensive domain coverage with fewer items than several of the more established generic and older people-specific instruments including the SIP, OMFAQ, and CARE.

Reliability

The most extensive evidence of reliability was found for the SF-36. Four generic instruments, namely the NHP, SF-12, SF-20, and SF-36, have evidence of internal consistency and test-retest reliability. The range of reliability estimates supports application at the group level and, in some instances, at the individual level. There is less evidence supporting the application of the COOP and EuroQol at the group level.

Several studies report higher levels of internal consistency reliability for domains within the FSQ (IADL domain), QLI, SF-12, SF-36 (physical function, bodily pain, role emotional, mental health), and the SIP (index), supporting application in individual assessment. Lower levels of internal consistency for domains within the AQoL, FSQ, and social function domains within the NHP, SF-36, and SIP have been reported. Although not falling below 0.67 when completed by older people (Wollinsky et al., 1998), lower levels of internal consistency for the SF-36 general health domain may be explained by evidence that some older people have difficulty answering items within this domain (Hayes et al., 1995; Mallinson, 1998; Sharples et al., 2000). For example, the item 'I expect my health to get worse' was viewed as unnecessarily negative (Hayes et al., 1995).

Low levels of test-retest reliability have been reported for the SF-36 role limitation domains (Andresen et al., 1996, 1999; Stadnyk et al., 1998; Sharples et al., 2000). The instrument developers have stated that these domains are appropriate to retired individuals (Ware, 1997). Others have suggested that the lower levels of reliability may reflect the difficulty experienced by older people in terms of role perception (Sharples et al., 2000).

The HSQ-12, IHQL, QWB, and SQL do not have evidence of internal consistency or test-retest reliability in older people and annot therefore be recommended for application. The AQoL, FSQ, GQL, QLI, and SIP lack evidence of test-retest reliability and the AQoL, GQL, and QLI have limited evidence of internal consistency, which limits the extent to which these instrument can be recommended. However, the AQoL is a new instrument and further evidence of instrument performance is required.

Most of the older people-specific instruments have limited evidence of reliability. Four instruments, namely the LEIPAD, PGCMAI, PWB, and WI, have evidence of internal consistency and test-retest reliability. The range of reliability estimates supports their application at the group level and, for the PGCMAI and WI, at the individual level. Limited evidence of internal consistency supports application of the ISAI, OMFAQ (ADL and IADL domains only), QOLPSV, and SENOTS at the group level, and, in some instances, at the individual level.

The BSQ and GQLQ do not have evidence of internal consistency or test-retest reliability in older people and cannot therefore be recommended for application. The CARE (all forms), ISAI, OMFAQ, QOLPSV, QLWMV, and SENOTS lack evidence of test-retest reliability and the EASY-Care, FAI, GPSS, GSQ, QLC, and SELF lack evidence of internal consistency, which limits the extent to which these instrument can be recommended. However, the three screening instruments, namely the BSQ, GPSS, and GSQ, and the EASY-Care are relatively new instruments and further evidence of instrument performance is required.

There was a wide range of test-retest intervals and few authors described the assessment of reliability in patients indicating no change in health. Evidence for test-retest reliability should be sought within an appropriate time-frame and with a supportive transition question to assess whether the individual's general health has remained stable between administrations (Streiner and Norman, 1995).

Validity

To support the comprehensive measurement of the domain of interest, item derivation and confirmation should be generated primarily from the views of the relevant population (Fitzpatrick et al., 1998). Patients and members of the lay public were only directly involved in item generation for two generic instruments (NHP and SIP) and three older people-specific instruments (the GQLQ, QOLPSV, and WI). Although patient participation enhances the validity of instrument content (Fitzpatrick et al., 1998), it is not clear whether people aged over 65 years were involved in item generation for the generic instruments. Item relevance, and hence the acceptability of instruments to the older population, should be considered when instruments are selected.

Empirical evidence supports the proposed health domains assessed by three generic instruments, namely the AQoL (Osborne et al., 2003), SF-12 (Resnick and Nahm, 2001), and SF-36 (Dexter et al., 1996; Wolinsky and Stump, 1996; Stadnyk et al., 1998) and six older people-specific instruments, namely the ISAI (Morris et al., 1990), LEIPAD (De Leo et al., 1998), OMFAQ (George and Fillenbaum, 1985; Harel and Deiling, 1984), PWB (Reker and Wong, 1984), SELF (Linn and Linn, 1984), and WI (Slivinski et al., 1996).

The interpretation of construct validity for many instruments was hindered by a lack of hypotheses relating to the size and direction of expected correlations, which has limited the interpretation of results in previous instrument reviews (Garratt et al., 2002b). Most instruments were assessed for validity through comparison with other instruments; global judgements of health; or clinical, socio-demographic, and health-service use variables. With the exception of the generic QLI and SF-12, and the older people-specific BSQ, EASY-Care, GSQ, and SELF, all instruments have evidence for validity through comparison with instruments that measure similar or related constructs. This is most extensive for the SF-36. The OMFAQ and CARE are the older people-specific instruments with the most extensive evidence.

With the exception of four generic instruments (COOP, GQL, IHQL, and SQL) and five older people-specific instruments (BSQ, GQLQ, GSQ, LEIPAD, and QLC), all instruments have evidence to support their ability to discriminate between groups defined by a range of socio-demographic, health, and health-service use variables. This was most extensive for the generic EuroQol, HSQ-12, NHP, SF-12, SF-36, and SIP, and the older people-specific FAI, GPSS, ISAI (preliminary), PWB, PGCMAI, QOLPSV, SENOTS, and WI instruments. Specific domains within these instruments discriminate between levels of health-service use, including need for care. The generic AQoL, COOP, SF-20, SF-36, and SIP, and older people-specific CARE, GPSS, OMFAQ, and SELF have evidence of predictive validity. Following completion by the chronically ill, most SF-36 domains were shown to be predictive of physician and hospital visits; physical function, bodily pain, and role physical domains were predictive of hospitalisation; general health and physical function domains were predictive of mortality (McHorney, 1996).

The generic EuroQol, HSQ-12, NHP, SF-12, SF-20, SF-36, and older people-specific ISAI and PWB have evidence to support the capacity to discriminate groups by age. Evidence for the SF-12 and SF-36 suggests a decline in physical health with age, but a constancy or improvement in mental health (McHorney et al., 1994b; Schofield and Mishra, 1998; Walter al, 2001; Baldassarre et al., 2002; Girotto et al., 2003). Similar

findings have been reported in population-based assessments in both the USA (McHorney et al., 1994b; Ware et al., 1994) and Australia (Schofield and Mishra, 1998). Cognitive ability is reported to deteriorate with age (NSF-OP, 2001), but few theories explain an enhancement in mental health. However, older people often experience difficulty in acknowledging or reporting mental health problems (Buckwalker and Piven, 1999, cited by Resnick and Nahm, 2001).

Older people may also be unwilling to report symptoms or reduced function, for example, pain or decreased mobility, considering these to be a part of normal ageing and not a reflection of health or illness (American Geriatrics Society, 1998; Resnick and Nahm, 2001, p158.). Notwithstanding reduced physical function and energy levels, older people have reported levels of global health comparable to younger populations, which may be explained by a difference in expectation or perception of what global health should be (Mangione et al., 1993). Alternatively, older people may be more willing to report more generic symptoms, for example, tiredness and pain, than to admit issues of role limitation and change in normative role function (McHorney, 1996).

These findings demonstrate the importance of item relevance and content validity in relation to instrument development, measurement properties, and practical issues such as respondent acceptability, score interpretation, and application. Seeking the views of older people with regard to instrument content and relevance is strongly recommended (McHorney, 1996).

Responsiveness

The ability to record change in health status above that described by measurement error is an essential requirement for evaluative instruments (Kirshner and Guyatt, 1985; Fitzpatrick et al., 1998). However, the interpretation of instrument responsiveness may be influenced by the method adopted to calculate responsiveness (Fitzpatrick et al., 1998; Husted et al., 2000) and the specific intervention (Wiebe et al., 2003; Beaton et al., 2001).

Ten generic instruments have evidence of responsiveness, the exceptions being the GQL, HSQ-12, IHQL, QLI, and QWB. Limited evidence was available for five older-people-specific instruments, namely the GQLQ, OMFAQ, PGCMAI, QOLPSV, and SELF. The most extensive evidence of responsiveness across a range of settings has been reported for the SF-36, with evidence to support small to large levels of responsiveness for improvement and deterioration in health across most domains. For example, following a comparative evaluation with the EuroQol in community-dwelling older women, the SF-36 had greater sensitivity to change across lower levels of morbidity (Brazier et al., 1996). Although both demonstrated high levels of responsiveness following the surgical repair of hip fracture, the EuroQol was more responsive than the SF-36 (Tidermark et al., 2003a). However, following a rehabilitation programme for frail older people, domain-specific instruments were more responsive than the SF-36 and NHP (Stadnyk et al., 1998).

High levels of responsiveness were reported for the EuroQol and NHP following interventions where change in health was expected to be substantive, including surgical repair of hip fracture (Van Balen et al., 2001, 2003; Tidermark et al., 2003a). However, limited evidence suggests that the NHP may be insensitive to the small but important changes in HRQL following physical therapy (Stadnyk et al., 1998; Mitchell et al., 2001).

An evaluation of drug therapy for osteoarthritis reported good responsiveness for the SF-12 physical component summary score (Thieler et al., 2002). Limited evidence suggests poor responsiveness for the SIP (Page et al., 1995; Liddle et al., 1996). There is limited or no evidence of responsiveness for the remaining generic instruments.

Small to moderate levels of responsiveness, comparable to other domain-specific instruments, were found for the GQLQ following the evaluation of day-care in a group of frail older people (Guyatt et al., 1993b). However, despite good content validity, enhanced responsiveness or validity in comparison to existing, simpler instruments was not demonstrated. The OMFAQ (physical health) was less responsive than the SF-36 and a condition-specific instrument following the surgical repair of hip fracture (Jaglal et al., 2000), and less responsive than the generic AQoL and SF-36 following the assessment of community care coordination (Osborne et al., 2000). The QOLPSV showed statistically significant improvements in score following the nursing care of chronically ill older people, but was less responsive than the SF-36 and did not discriminate between acute, chronic, or palliative care patients (Irvine et al., 2000). Limited evidence supported the responsiveness of the SELF (Linn and Linn, 1984).

Although a necessary measurement property of instruments intended for application in evaluative studies for the measurement of longitudinal changes in health, responsiveness has been the most neglected area of evaluating instruments for use with older people. In addition, the level of change in HRQL that is important to patients, the Minimal Important Difference (MID), has not been addressed. Instruments should be administered longitudinally, before and after changes in treatment known to improve health-related quality of life, and health transition ratings should be included as external criteria of change in patient health (Husted et al., 2000). Where possible, the relative responsiveness of instruments should be assessed concurrently (Guyatt et al., 1993a; Garratt et al., 2002b; Wiebe et al., 2003).

Precision

Score distribution and end effects were reported for several generic instruments but only for the older people-specific OMFAQ (ADL domain). Although expected to reduce with age, ceiling effects were reported for several instruments, specifically for physical mobility domains within the COOP and FSQ; several domains within the HSQ-12; all domains within the NHP and SF-20; role limitation and social function domains of the AQoL, SF-36, and the SIP; and the ADL and IADL domain of the OMFAQ. This suggests that domains within these instruments may not discriminate between older people with low morbidity levels. However, measuring improvement in patients with excellent health may be less of a concern than measuring deterioration.

Due to the potential for older people to have more sickness than the general population, floor effects following completion of the SF-36 by older people were reported for the two role limitation domains (McHorney, 1996). Floor effects were also reported for several COOP charts, SF-20 (role function), and SIP (summary scores and several domains). Measuring deterioration in the health of patients whose health is already poor is an important requirement of health assessment in general (Bindman et al., 1990), and particularly with older people. However, the developers of the modified SF-36 version 2 (v2) suggest that the improved range of response categories enhances instrument precision and reduces the floor effects observed in the general population (Quality

Metric Incorporated web-site: www.sf-36.org/community/sf36v2andsf12v2.shtml). The SF-12 v2 and SF-36 v2 have not been evaluated in an older population.

Acceptability

Evidence across most instruments suggests that completion difficulties increase with age, deteriorating health status, and declining cognitive ability. Although similarly high levels of patient satisfaction have been reported for the SF-12, SF-36, and SIP, few other instruments have been as extensively evaluated for acceptability in older people as the SF-36.

Score calculation is dependent upon instrument completion rates, and although many instruments accommodate the omission of several items, where item omission is high, validity is threatened. Where more than 10% of data is missing, this is considered a substantial loss and is particularly important where non-random item omission is identified (McHorney, 1996). Items may be omitted due to perceived ambiguity or non-relevance. For example, older people frequently omit items from the SF-36 and SIP related to work, vigorous activity, health outlook, sexual activity, and social function (Andresen et al., 1998a,b).

For the SF-12 and SF-36, the mixed response formats and question length, or 'strings', reportedly cause confusion in respondents, and modifications of item format have been proposed which apply to the York SF-12 (Iglesias et al., 2001) and new versions of both instruments (Ware et al. 2000; 2002). Informed by the cross-cultural translation of instruments and completion difficulties experienced across populations, version 2 modifications include simplified instructions and wording to reduce ambiguity, improved layout, and five-point response options which run horizontally left to right for most items. The developers report improved measurement properties and acceptability in the general population, but there is no published evidence for the performance of these instruments with older people.

Instrument length and mode of administration impose a burden on both responders and staff in terms of completion, data entry, analysis, and cost. The mode of administration is an important consideration in maximising data. The best completion rates were reported for interview administration of all instruments, with many instruments achieving a 100% completion rate. However, this is associated with longer completion times and increased cost.

Although the results of this review would suggest that this mode of administration is used less frequently, telephone interviews are less costly than personal interviews and also achieve good completion rates (McHorney et al., 1994b). However, hearing impairment in the older age group may limit the usefulness of telephone administration.

Self-administration, either by post or within a clinic or hospital setting, is the cheapest mode of administration (McHorney et al., 1994b), but many studies report low completion rates. Self-completion rates of less than 50% have been reported for the SF-36 (Hayes et al., 1995) and SIP (Jannink-Nijlant et al., 1999). Although good self-completion rates have been reported for the EuroQol, one study estimated that where 11% of people aged 65 years would require interview administration, this increased dramatically in older age-groups (Coast et al., 1999). Sight impairment, limited reading ability together with fine motor disability due to (for example) arthritis, age, impaired cognitive ability, and general ill-health are the principal reasons for instrument non-

completion in older people (McHorney, 1996; Coast et al., 1999). Respondents may therefore be drawn from a limited range of the healthier young-old, capable of self-completing a questionnaire and returning it in the post.

Furthermore, when postal non-responders to the older people-specific GPSS were contacted by telephone, a large number were classified as being at high risk of functional decline (Alessi et al., 2003). This led the authors to suggest that persuasive methods are required to increase response rates for postal self-completion. Combined use of both mail and telephone administration of the SF-36 in a general population in the USA that included a large number of respondents over the age of 65 years gave a higher response rate (77.1%) than either mail self-completion (65.1%) or telephone administration (65.3%) alone (McHorney et al., 1994b).

Small print and unfamiliarity with printed questionnaires further add to self-completion difficulties for older people. McHorney (1996) suggests that a larger typeface and greater use of white space in questionnaire design aids completion. Only the developers of the GPSS specifically indicate the use of large print; high response rates were reported for both the development (88%) and main surveys (90%), with only 11% of respondents requiring assistance (Alessi et al., 2003). In addition, limited reading skills may be over-represented in the older population (McHorney, 1996), and the required reading level for questionnaires should be considered. Evidence suggests that the NHP and SF-36 have comparable levels of readability (Sharples et al., 2000).

Mixed mode survey design has been reported by several authors (for example, Coast et al., 1998; Smeeth et al., 2001; Fowler et al., 2000) but differences in administration may threaten validity. Evidence suggests that respondents are more likely to report more positive health states with interview administration (McHorney et al., 1994b; Fitzpatrick et al., 1998). Older people who were interview-administered the BSQ reported better levels of health than those who self-completed the instrument (Smeeth et al., 2001).

An alternative method of instrument administration involves proxy completion by informed health professionals or nominated lay-persons. Following proxy completion of the SF-36, greater levels of agreement for a patient's perceived health status were found between patients and informed professionals than between patients and nominated lay proxies (Pierre et al., 1998; Ball et al., 2001). The results of proxy completion of the EuroQol (Tamim et al., 2002) and the SF-36 (Pierre et al., 1998; Yip et al., 2001) suggest that higher levels of agreement are found regarding the assessment of more observable aspects of health compared to more subjective constructs. Proxies may overestimate health limitations, particularly for less observable health constructs such as emotions and mental health status. The OMFAQ and FAI incorporate interviewer ratings alongside respondent answers in calculating final scores. The dependence of these instruments on interviewer training and the 'clinical insight' necessary to support the translation of item responses into summary scores has been criticised (Morris et al., 1989).

Cognitive impairment

The point at which an individual with cognitive impairment becomes unable to give a valid report on their health is not known (Fletcher et al., 1992; Albert, 1997). However, in 1988 it was recommended that patient-reported health instruments should not be used in the assessment of cognitively impaired older people (Society for General Internal Medicine Task Force of Health Assessment Guidelines for Geriatric Assessment, 1988, cited by McHorney, 1996). Consequently, most of the studies examined in this review exclude cognitively impaired respondents; only four of the articles reviewed specifically included patients with cognitive impairment (McHorney et al., 1990, cited by McHorney, 1996: SF-36; Parker et al., 1998: SF-36; Pettit et al., 2001: HSQ-12, SF-12; Seymour et al., 2001: SF-36).

Unfortunately, this limits the assessment of patient-reported health across the broad spectrum of old age and hence the evaluation of instrument performance. Evaluation of the SF-36 in a general older population in the USA, of whom 5.8% were cognitively impaired, suggested that, notwithstanding longer self-completion times and increased missing items, reliability and validity across most domains were comparable (McHorney et al., 1990, cited by McHorney, 1996). However, following interview completion in a UK population of physically disabled older people with and without cognitive impairment, lower levels of reliability and validity were found for the cognitively impaired group (Seymour et al., 2001). The application of patient-reported health instruments across the spectrum of cognitive impairment in older people is required to further inform evaluation of instrument performance.

c) Concurrent evaluations

Both concurrent evaluations and reviews of measurement properties inform instrument selection and standardisation (Garratt et al., 2002a). However, there are few concurrent instrument evaluations, particularly in relation to responsiveness, both generally (Fitzpatrick et al., 1998; Garratt et al., 2002a), and specifically within the assessment of older people. Concurrent evaluations between generic, generic and older people-specific, and generic or older people-specific with domain-specific instruments are shown in Table 8. Most evaluations include the SF-36.

Several evaluations report similar levels of reliability and evidence for validity between the SF-36 and EuroQol (Brazier et al., 1996; Tidermark et al., 2003a), and between the SF-36 and NHP (Crockett et al., 1996; Stadnyk et al., 1998; Sharples et al., 2000). The SF-36 appears to be more responsive to change in health across lower levels of morbidity (Brazier et al., 1996; Sharples et al., 2000; Osborne et al., 2003), but the EuroQol (Brazier et al., 1996; Tidermark et al., 2003a) and NHP (Sharples et al., 2000; Van Balen et al., 2001, 2003) may be more responsive where substantive changes in health status are expected. Comparable levels of responsiveness have been reported between the COOP and SF-36 (Jenkinson et al., 1997), and WONCA/COOP and NHP (Van Balen et al., 2001,2003).

Evaluations comparing generic instruments suggest that the SIP is not suitable for the assessment of community-based older people, largely due to ceiling effects and time needed for administration (Weinberger et al., 1991; Andresen et al., 1995; Andresen et al., 1998a).

Seven concurrent evaluations of generic and older people-specific instruments were reviewed (Guyatt et al., 1993b; Reuben et al., 1995; Stadnyk et al., 1998; Irvine et al., 2000; Jaglal et al., 2000; Philp et al., 2001; Osborne et al., 2003): see Table 8. Reliability and content validity were infrequently assessed (Jaglal et al., 2000). Two studies reported higher (Stadnyk et al., 1998) or comparable (Guyatt et al., 1993) levels of responsiveness for older people-specific instruments in comparison to generic instruments. Higher levels of responsiveness were reported for the SF-36 when compared with the QOLPSV (Irvine et al., 2000) and OMFAQ (Jaglal et al., 2000; Osborne et al., 2003).

In accordance with recommendations for the general population (Guyatt et al., 1993a; Fitzpatrick et al., 1998), several evaluations comparing generic and disease-specific instruments generally supported their combined use with older people (Bombardier et al., 1995; Jenkinson et al., 1995; Jaglal et al., 2000; Groessl et al., 2003). Disease- and population-specific instruments may have greater clinical appeal due to the specificity of content, and an associated increased responsiveness to specific change in condition. The broad content of generic instruments facilitates the identification of co-morbid features and treatment side-effects that may not be captured by specific instruments, but may also reduce responsiveness to small but important changes.

Do older people-specific instruments perform better than generic instruments?
There is insufficient evidence from concurrent evaluations of generic and older people-specific instruments. Supported by recommendations from this review, comparative empirical evaluations of widely used generic and new or widely used older people-specific instruments, global assessments, and domain-specific instruments are required.

d) Screening the older population

Three generic instruments have been evaluated for screening purposes. The validity and utility as screening tools of the CARE activity limitation (AL) and cognitive impairment (CI) domains was assessed using measures of family inconvenience and decision to institutionalise an older relative as criterion variables (Teresi et al., 1984a). As hypothesised, both domains discriminated between family groups defined by their level of perceived inconvenience and decision to institutionalise. Although many community-dwelling older people may be correctly diagnosed with pervasive depression or dementia using the SHORT-CARE diagnostic scales (sensitivity 84% and 91%, respectively), the high true negative rates (specificity 35% and 30%, respectively) suggest they should be used with caution (Gurland et al., 1984).

The COOP emotional condition (EC) chart highlighted possible depression in 32.7% of a community-based population. The concurrent review of medical records revealed a medical diagnosis of depression in only 7% (Doetch et al., 1994). Completion of a range of depression measures suggested an illness prevalence in the range 16.5% to 34.7%, supporting the possible role of the COOP EC in screening for depression in older people. The SIP (68-item) mobility domain had high sensitivity (91%) for poor function but low specificity (58%) for good function (Jannink-Nijlant et al., 1999). Furthermore, it discriminated between groups defined as recurrent fallers and non-fallers, and identified people at risk of recurrent falling.

Three older people-specific screening instruments have been reviewed, namely the BSQ (Smeeth et al., 2001), GPSS (Alessi et al., 2003), and GSQ (Fernandez Buergo et al.,

2002). The diagnostic accuracy of the BSQ was supported by high specificity (greater than 90% for all domains), but low sensitivity (less than 50% for all domains) suggests it should be used with caution when screening for depression and impairments in cognition, hearing, or vision (Smeeth et al., 2001). Although assessment accuracy was limited for functional and cognitive impairment, the GPSS had high sensitivity and specificity for the risk of falls, depression, and urinary incontinence (Alessi et al., 2003). The GPSS discriminated between groups defined by levels of co-morbidity and health-service use, and groups defined as high- or low-risk on the GPSS were discriminated by scores on several domain-specific and generic instruments. The GSQ had a low sensitivity (58%) but high specificity (89%) when used as a confirmation test for positive or negative health in a comprehensive geriatric assessment (Fernandez Buergo et al., 2002).

Although a role in screening has been described by the developers of six additional older people-specific instruments, namely the FAI, ISAI, OMFAQ, QOLPSV (brief), SELF, and SENOTS, there is limited published evidence to support this role.

e) Review limitations

This broad-based review included patient-reported multidimensional instruments, which purport to measure health-related quality of life and had evidence of reliability or validity following application in the assessment of older people. Although based on an extensive database search, the review is limited by the exclusion of non-English language publications.

Developers of the older people-specific instruments were contacted by mail but this did not result in a high response rate. Many letters were returned, it being difficult to locate an appropriate contact address other than that identified from publications. Inconsistencies in the reporting of several instruments were identified and contact with the development team could have provided further clarification.

f) Recommendations

The review provides an extensive synthesis of evidence describing how the instruments identified perform in measuring health-related quality of life in older people. Consideration for application in clinical trials, routine practice, or the community setting requires an instrument with content relevant to the proposed application, which fulfils essential measurement properties, is brief and simple to administer, and is acceptable to the respondent, thus ensuring maximum completion rates (Fitzpatrick et al., 1998; Eiser and Moore, 2001).

For the SF-36, EuroQol, and NHP there is relatively good evidence of reliability, supporting their application in the assessment of groups and, for the SF-36 and NHP in some instances, in the assessment of individuals; good evidence of validity and responsiveness was also found. The SF-36 has relatively good evidence of responsiveness across a range of settings and populations, which suggests that it is sensitive to change, particularly in community-dwelling older people and in those with lower levels of morbidity. Both the EuroQol and NHP have high levels of responsiveness following interventions resulting in substantive changes in health (Van Balen et al., 2001, 2003; Tidermark et al., 2003a). In the rehabilitation of frail older people, the NHP and SF-36 were less responsive than older people-specific (OMFAQ)

and domain-specific instruments (Stadnyk et al., 1998). However, in the assessment of chronically ill (Irvine et al., 2000) and community-dwelling older people (Osborne et al., 2003), the SF-36 was more responsive than older people-specific measures of HRQL. The SF-36 produces profile and summary scores; the EuroQol produces profile, index, and utility scores; the NHP produces a profile score. Utility weights are available for the SF-36 and SF-12.

There is relatively good evidence of reliability for the COOP charts and SF-12, supporting their application in the assessment of groups, but limited evidence of reliability for the SIP. There is moderate evidence of validity for the COOP charts and SIP, and limited evidence for the SF-12. The COOP charts had limited responsiveness following a rehabilitation programme for older people, but higher levels following the surgical repair of hip fracture. There is limited evidence of responsiveness for the SF-12, and weak evidence for the SIP. Evidence for the remaining instruments is weak. The new AQoL and older people-specific modifications to the SQL require further evaluation. The SF-12 v2 and SF-36 v2 have yet to be evaluated in an older population. The IHQL and QWB lack evidence for reliability and responsiveness and are not recommended for the assessment of older people.

Four older people-specific instruments, namely the OMFAQ, PGCMAI, QOLPSV, and SELF, have relatively good evidence of reliability supporting their application in the assessment of groups, good evidence of validity, and limited evidence of responsiveness. With the exception of the GQLQ, the remaining instruments lack evidence of responsiveness. Despite the large number of evaluations, evidence of reliability for the OMFAQ was reported in only two studies, and is limited to internal consistency reliability. With the exceptions of the EASY-Care, FAI, GPSS, LEIPAD, PGCMAI, QLC, SELF, and WI, all older people-specific instruments lack evidence of test-retest reliability. There is relatively good evidence of reliability for the CARE, ISAI, LEIPAD, and PWB, supporting their application in group evaluation, and some evidence of validity. However, the correlation between these instruments and well-established generic, disease-specific, and other older people-specific instruments has not been reported. Evidence for the remaining instruments is weak.

The newly developed EASY-Care and GPSS require further evaluation. The EASY-Care covers the most extensive range of health domains of all the instruments reviewed, with an economical number of items. The EASY-Care is an important development in the comprehensive assessment of older people and the single assessment process. Limited evidence suggests acceptable reliability and respondent acceptability, but evidence of validity and responsiveness is lacking. In addition, the GPSS provides a new self-completed instrument for the postal screening of community-dwelling older people, to identify those who would most benefit from a comprehensive assessment.

Two broad methods for the measurement of HRQL have been reviewed, namely generic instruments and those specific to the assessment of older people. Generic instruments are suitable for comparisons between general and specific populations, where the availability of normative data supports the interpretation of data. Generic instruments are also particularly relevant to economic evaluation. Their use in general population surveys and the results of this review support the application of several generic instruments in the assessment of community-dwelling older people. For example, the evidence reviewed suggests that the SF-36 is more responsive than older people-specific instruments (OMFAQ, QOLPSV) with community-dwelling adults. Where a more

detailed and broad-ranging assessment of HRQL is required, particularly in older people with lower levels of morbidity, the SF-36 is recommended; initial evaluation of the SF-12 v2 and SF-36 v2 in older people is also recommended. Where a more succinct assessment of HRQL is required, particularly for patients in whom a substantive change in health is expected, the EuroQol is recommended, but further evidence of reliability and respondent acceptability is required. However, the content of some items of generic instrument may have less relevance for, and reduce acceptability and responsiveness in, the very old and those with physical disabilities.

Older people-specific instruments aim to have greater relevance to the immediate health concerns of the older population. This may enhance respondent acceptability and instrument responsiveness to specific changes in health. Instrument specificity may increase applicability to particular older populations or settings, for example, frail elderly people in hospital settings, but reduce applicability to the general older population. However, few specific instruments included older people in item derivation and evidence of responsiveness is limited. The OMFAQ had the greatest number of evaluations with good evidence supporting instrument validity, but this was mostly limited to the performance of the ADL domain; evidence of reliability was limited and responsiveness was poor. Further evaluation of the newly developed EASY-Care and GPSS is recommended.

Further evaluation and, where appropriate, refinement of existing instruments is required before new instruments are developed. Seeking the views of older people with regard to instrument format, relevance, and mode of completion is strongly recommended (McHorney, 1996). Where it is deemed necessary to develop new instruments, particularly older people-specific instruments, the close involvement of older people in the development of instruments is recommended.

Supported by recommendations from this review, comparative empirical evaluations of widely-used generic and new or widely used older people-specific instruments, global assessments, and domain-specific instruments are required for both the general older population and for specific older patient populations. This research will inform decisions regarding the selection of instruments for future application in research and clinical practice.

In conclusion, this comprehensive review has evaluated the evidence for the measurement properties of instrumments, and practical issues, and is designed to inform instrument selection for applications where self-rated health assessment in older people is required. Instruments that cover a range of health domains within the construct of HRQL have been reviewed; the appropriateness of item content in relation to the questions that any application seeks to address must be evaluated (Patrick and Erickson, 1993; Fitzpatrick et al., 1998). Clinicians and researchers wishing to select instruments for particular applications must consider these methodological and practical issues as well as issues of appropriateness.

Table 8 Summary of concurrent evaluations of reviewed instruments

Concurrent evaluations including the SF-36									
Study	***Instruments***			***Measurement properties***					***Results***
	Generic	**OP**	**Other**	**Reliability**	**Validity**	**Responsive-ness**	**Precision**	**Acceptability**	
Andreson et al. (1995)	QWB, SF-36 (PF,RP,GH), SIP	-	-	Test-retest: SF-36 (PF only)	Moderate correlation between related domains	-	Ceiling effects: SIP, SF-36 (PF,RP)	QWB difficult to administer	No one tool suitable for all applications; SIP unsuitable for healthy community elders
Andreson et al. (1998a)	SF-36, SIP	-	-	Alpha values similar	Moderate correlation between related domains	-	Ceiling effects: SIP	Satisfaction, data similar; SF-36 quicker to complete	SF-36 recommended for community assessments
Bombardier et al. (1995)	SF-36	-	WOMAC	-	Measure different but complementary aspects of the health construct; small to moderate correlation	-	Consistently higher scores: WOMAC	-	Use of both generic and disease-specific instruments supported
Brazier et al. (1996)	EuroQol, SF-36	-	-	Test-retest similar	Large correlation between related domains	SF-36 more sensitive to lower morbidity levels (hypothetical)	Floor effects: SF-36 (PF,RP,RE)	SF-36 lower completion rates	EuroQol for brevity and where change in health is substantive SF-36 for detail and greater sensitivity
Crockett et al. (1996)	NHP, SF-36	-	-	-	Small to large correlation between related domains	-	-	-	No one instrument recommended
Irvine et al. (2000)	SF-36	QOLPSV	-	SF-36 higher alpha values	-	SF-36 more responsive (except GH)	Similarly low missing data	-	SF-36 more reliable and responsive
Jaglal et al. (2000)	SF-36	OMFAQ (PH)	LEM	OMFAQ content less applicable to change post-fracture	Large correlation between related domains (SF-36, LEM)	LEM, SF-36 (PF,RP,BP) comparably high; OMFAQ moderate	-	Similar	Use of both generic and disease-specific (LEM) instruments supported post-fracture in community-dwelling elders

Jenkinson et al. (1995)	SF-36	-	PDQ	Alpha values comparable	Small to large correlation with other instruments	-	-	-	Use of both generic and disease-specific instruments supported
Jenkinson et al. (1997)	COOP, SF-36	-	-	-	-	Comparably small effect size (ES)	-	-	Both instruments have low levels of responsiveness
Osborne et al. (2003)	AQoL, SF-36	OMFAQ	-	-	Large correlation between related domains (AQoL with SF, OMFAQ)	AQoL (IL) and SF (BP) comparable responsiveness	Ceiling effects: AQoL (SR,PS), SF-36 (RP,SF,RE) Floor effects: SF-36 (RP,RE)		AQoL and SF-36 comparable responsiveness AQoL fewer end effects OMFAQ no end effects, but low responsiveness
Reuben et al. (1995)	FSQ (I/ADL) SF-36 (PF)	OMFAQ (ADL)	Katz ADL, PPT	-	Inconsistent small to moderate correlations between related domains		Interview completion rate higher than self-completion	Mixed completion: only SF-36 and PPT self-completed	Instruments measure different levels of PF. Impact of different response options, task combination, self-completion. Composite index may be most appropriate
Schofield & Mishra (1998)	SF-12, SF-36	-	-	-	Both discriminate between age-groups; domain scores differ	-	-	SF-12 summary scores more useful than single domains	SF-36 summary and domain scores more reliable and precise for assessing change over time and between groups
Sharples et al. (2000)	NHP, SF-36	-	Katz ADL, HADS, GPT	Test-retest, alpha values similar	Large correlation between related domains: comparable between instruments	-	Ceiling effects: NHP all domains, SF-36 (RE,RP, SF,BP)	Readability: NHP 93.9% SF-36 86.8% Item completion: both 99.6%	Similar reliability and validity SF-36 more sensitive to minor morbidity; includes general health NHP includes sleep
Sherman & Reuben (1998)	FSQ (I/ADL), SF-36 (PF)	-	Perform-ance tests	Alpha values: SF-36 >0.90, others >0.60	Moderate to large correlation between related domains	-	Ceiling effects FSQ (I/ADL)	-	SF-36 recommended: simple and reliable. Performance and self-completed instruments measure different aspects of function
Stadnyk et al. (1998)	NHP, SF-36, SQL	OMFAQ (IADL)	BI	-	With SF-36: moderate to large correlations between related domains	SQL, BI, OMFAQ most responsive; SF-36, NHP comparably low responsiveness	-	-	SF-36 reliable and valid but less responsive than SQL and OMFAQ in frail older people undergoing rehabilitation

Tidermark et al. (2003a)	EuroQol, SF-36 (GH, BP, PF)	-	-	-	Small to large correlation between related domains; both have discrimin-ative validity	Large ES for both instruments, larger for EuroQol	-	-	Small to large correlations between related domains Both have good levels of responsiveness; better for EuroQol
Weinberger et al. (1991)	SF-36, SIP	-	-	-	Concurrent: large correlation between related domains	-	SIP: reports higher levels of function	Time: SF-36 quicker to complete	SF-36 shorter administration
Concurrent evaluations not including the SF-36									
Study	***Instruments***			***Measurement Properties***					***Results***
	Generic	**OP**	**Other**	**Reliability**	**Validity**	**Responsive-ness**	**Precision**	**Acceptability**	
Cairl et al. (1983)	-	FAI, OMFAQ	-	-	Criterion: small to large correlations	-	-	FAI quicker to complete	Further evidence required for FAI
Groessl et al. (2003)	QWB	-	AIMS	-	Small to moderate correlation	[ss]Change score correlation between QWB and AIMS total, physical activity, health, and pain scores	-	-	Ability of QWB to measure QoL in older people with osteoarthritis supported
Guyatt et al. (1993b)	Rand physical and emotional function	GQLQ	BI	-	GQLQ good content	Comparable responsiveness	-	-	GQLQ no real advantage in comparison to simpler instruments
Iglesias et al. (2001)	SF-12, York SF-12	-	-	Alpha values comparable	-	-	-	Response rates comparable	York SF-12 slight improvement over SF-12 in reliability and item response
Philp et al. (2001)	WONCA/ COOP	EASY-Care	-	-	-	-	-	EASY-Care quicker	GP, staff, patients more satisfied with EASY-Care
Siu et al. (1993a)	COOP, SF-20	-	-	-	Emotional health predictive of skilled care; overall health predictive of hospitalisation	Change in score not associated with later placement in skilled care	-	-	Both demonstrate predictive validity

Siu et al. (1993b)	COOP, SF-20	-	(Katz ADL)	-	Change score:- small to large correlation between pain, social, mental health (not physical function [PF])	SF-20 PF most discriminative for worsening performance-based activities	-	-	PF domains: comparably low levels of responsiveness, and less responsive to improvement in health than other domains
Tedesco et al. (1990)	FSQ (ADL, IADL)	-	NYHA	-	FSQ better able to assess functional impairment	Change in FSQ scores predict symptomatic deterioration	-	-	FSQ recommended for assessment of functional status
Van Balen et al., (2001)	WONCA/ COOP (W/C), NHP	-	(RAP)	-	-	5 NHP domains (not E), 4 W/C domains (not SA,E) responsive; mobility and pain most responsive	-	-	NHP wider coverage and more responsive
Van Balen et al. (2003)	WONCA/ COOP (W/C), NHP	-	BI, RAP	Alpha values: NHP 5 >0.70 (SI 0.52), RAP 3 >0.70 (1=0.13), BI 0.92	All cover function NHP broader content than W/C	ES for NHP and W/C comparably small to large, RAP small to large, BI large	Ceiling effects: NHP (SI, Sleep), W/C (SA) Floor effects: W/C (PF)	All <10 minutes completion	NHP recommended for change in emotion, pain, energy RAP recommended for functional status

Key:
[ss] statistically significant

AIMS	Arthritis Impact Measurement Scale	NYHA	New York Heart Association Functional Scale
BI	Barthel Index	OP	Older people-specific instruments
GPT	Guralnik Performance Test	PDQ	Parkinsons Disability Questionnaire
HADS	Hospital Anxiety and Depression Score	PPT	Physical Performance Test
Katz ADL	Katz Activities of Daily Living Scale	RAP	Rehabilitation Activities Profile
LEM	Lower Extremity Measure	WOMAC	Western Ontario and McMaster Universities Arthritis Index

REFERENCES

Albert SM. (1997) Assessing health-related quality of life in chronic care populations. *Journal of Mental Health and Aging;* 3: 101-118.

Albrecht GL. (1994) Chapter 2: Subjective health assessment. In: Jenkinson C (Ed). Measuring health and medical outcomes. London: UCL Press.

Alessi CA, Josephson KR, Harker JO, Pietruszka FM, Hoyl MT, Rubenstein LZ. (2003) The yield, reliability, and validity of a postal survey for screening community-dwelling older people. *Journal of the American Geriatrics Society;* Feb; 51(2): 194-202.

Altman DG. (1996) Practical Statistics for Medical Research. Chapman and Hall, London.

American Geriatrics Society (AGS) Panel on Chronic Pain in Older Persons. (1998) *Journal of the American Geriatrics Society;* 46: 635-651.

Anderson C, Laubscher S, Burns R. (1996) Validation of the Short Form 36 (SF-36) health survey questionnaire among stroke patients. *Stroke;* 27: 1812-1816.

Andersson L. (1981) The Psychosomatic Symptoms Scale. Stockholm Gerontology Research Centre, Stockholm.

Andersson G, Melin L, Lindberg P, Scott B. (1995) Dispositional optimism, dysphoria, health, and coping with hearing impairment in elderly adults. *Audiology;* 34: 76-84.

Andresen EM, Patrick DL, Carter WB, Malmgren JA. (1995) Comparing the performance of health status measures for healthy older adults. *Journal of the American Geriatrics Society;* 43: 1030-1034.

Andresen EM, Bowley N, Rothenberg BM, Panzer R, Katz P. (1996) Test-retest performance of a mailed version of the Medical Outcomes Study 36-item Short-Form health survey among older adults. *Medical Care;* 34: 1165-1170.

Andresen EM, Rothenberg B, Panzer R, Katz P, McDermott MP. (1998a) Selecting a generic measure of health-related quality of life for use among older adults: a comparison of candidate instruments. *Evaluation and the Health Professions;* 21: 244-264.

Andresen EM, Rothenberg BM, Kaplan RM. (1998b) Performance of a self-administered mailed version of the Quality of Well-Being (QWB-SA) questionnaire among older adults. *Medical Care;* 36: 1349-1360.

Andresen EM, Gravitt GW, Aydelotte ME, Podgorski CA. (1999) Limitations of the SF-36 in a sample of nursing-home residents. *Age and Ageing;* 28: 562-566.

Antonovsky A. (1987) Unraveling the mystery of health. Jossey-Bass, San Francisco.

Arnold SB. (1991) Chapter 3: Measurement of quality of life in the frail elderly. In: Birren JE, Lubben JE, Rowe JC, Deutchman DE, et al.., (editors) The Concept and Measurement of Quality of Life in the Frail Elderly. Academic Press, Inc.

Baldassarre FG, Arthur HM, DiCenso A, Guyatt G. (2002) Effect of coronary artery bypass graft surgery on older women's health-related quality of life. *Heart and Lung;* Nov-Dec. 31(6): 421-31.

Ball A, Russell E, Seymour D, Primrose W, Garratt A. (2001) Problems in using health survey questionnaires in older patients with physical disabilities. *Gerontology;* 47: 334-340.

Bath P, Philp I, Boydell L, McCormick W, Bray J, Roberts H. (2000) Standardized health check data from community-dwelling elderly people: the potential for comparing populations and estimating need. *Health and Social Care in the Community;* 8: 17-21.

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

Bergner M, Bobbitt RA, Kressel S, et al. (1976) The Sickness Impact Profile: conceptual formulation and methodology for the development of a health status measure. *International Journal of Health Service;* 6: 393-415.

Bergner M, Bobbitt RA, Carter WB et al. (1981) The Sickness Impact Profile: development and final revision of a health status measure. *Medical Care;* 19:787-805.

Berkman B, Chauncey S, Holmes W, Daniels A, Bonander E, Sampson S, Robinson M. (1999) Standardized screening of elderly patients' needs for social work assessment of primary care: use of the SF-36. *Health and Social Work;* 24: 9-16.

Beusterien KM, Steinwald B, Ware J. (1996) Usefulness of the SF-36 Health Survey in measuring health outcomes in the depressed elderly. *Journal of Geriatric Psychiatry and Neurology;* 9: 13-21.

Bindman AB, Keane D, Lurie N. (1990) Measuring health changes among severely ill patients. *Medical Care;* 28(12): 1142-1152.

Bland JM, Altman DG. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet;* I: 307-310.

Bombardier C, Melfi CA, Paul J, Green R, Hawker G, Wright J, Coyte P. (1995) Comparison of a generic and a disease-specific measure of pain and physical function after knee replacement surgery. *Medical Care;* 33: AS131-AS144.

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

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

Bowling A, Windsor J. (1997) Discriminative power of the health status questionnaire 12 in relation to age, sex, and long-standing illness: findings from a survey of households in Great Britain. *Journal of Epidemiology and Community Health;* 51: 564-573.

Bowling A. (1998) Measuring health-related quality of life among older people. Editorial. *Aging and Mental Health;* 2(1): 5-6.

Brach JS, Van Swearingen JM, Newman AB, Kriska AM. (2002) Identifying early decline of physical function in community-dwelling older women: performance-based and self-report measures. *Physical Therapy;* Apr. 82: 320-328.

Brazier JE, Jones N, Kind P. (1993) Testing the validity of the EuroQol and comparing it with the SF-36 Health Survey questionnaire. *Quality of Life Research;* 2: 169-180.

Brazier JE, Walters SJ, Nicholl JP, Kohler B. (1996) Using the SF-36 and EuroQol on an elderly population. *Quality of Life Research;* 5: 195-204.

Breithaupt K, McDowell I. (2001) Considerations for measuring functioning of the elderly: IRM dimensionality and scaling analysis. *Health Services and Outcomes Research Methodology;* 2: 37-50.

Buckwalter K, Piven M. (1999) Depression. In: Stone J, Wyman J, Salisbury S. Clinical Gerontological Nursing. (2nd Edition). Philadelphia: W.B Saunders Co.

Burstrom K, Johannesson M, Diderichsen F. (2001) Health-related quality of life by disease and socio-economic group in the general population in Sweden. *Health Policy;* Jan. 55: 51-69.

Cairl RE, Pfeiffer E, Keller DM, Burke H, Samis HV. (1983) An evaluation of the reliability and validity of the Functional Assessment Inventory. *Journal of the American Geriatrics Society;* 31: 607-612.

Carver DJ, Chapman CA, Thomas VS, Stadnyk KJ, Rockwood K. (1999) Validity and reliability of the Medical Outcomes Study Short Form-20 questionnaire as a measure of quality of life in elderly people living at home. *Age and Ageing;* 28: 169-174.

Cleary PD, Jette AM. (2000) Reliability and validity of the functional status questionnaire. *Quality of Life Research;* 9(6 A): 747-753.

Coast J, Peters TJ, Richards SH, Gunnell DJ. (1998) Use of the EuroQol among elderly acute care patients. *Quality of Life Research;* 7: 1-10.

Cochrane T, Davey RC, Munro J, Nicholl J. (1998) Exercise, physical function and health perceptions of older people. *Physiotherapy;* 84: 598-602.

Condello C, Padoani W, Uguzzoni U, Caon F, De Leo D. (2003) Personality disorders and self-perceived quality of life in an elderly psychiatric outpatient population. *Psychopathology;* 36:78-83.

Cousins SO. (1997) Validity and reliability of self-reported health of persons aged 70 and older. *Health Care for Women International;* 18: 165-174.

Crockett AJ, Cranston JM, Moss JR, Alpers JH. (1996) The MOS SF-36 Health Survey questionnaire in severe chronic airflow limitation: comparison with the Nottingham Health Profile. *Quality of Life Research;* 5: 330-338.

Crumbaugh J, Maholick L. (1964) An experimental study in existentialism: the psychometric approach to Frankl's concept of noogenic neurosis. *Journal of Clinical Psychology;* 20: 200-207.

DeBon M, Pace P, Kozin F, Kaplan RM. (1995) Validation of self-reported dysfunction in older adults. *Journal of Clinical Geropsychology;* 1: 283-292.

De Bruin AF, De Witte LP, Stvene F, Diederiks JPM. (1992) Sickness Impact Profile: the state of art of a generic functional status measure. *Social Science and Medicine;* 35: 1003-14.

Degl'Innocenti A, Elmfeldt D, Hansson L, Breteler M, James O, Lithell H, Olofsson B, Skoog I, Trenkwalder P, Zanchetti A, Wiklund I. (2002) Cognitive function and health-related quality of life in elderly patients with hypertension-baseline data from the study on cognition and prognosis in the elderly (SCOPE). *Blood Pressure;* 11(3): 157-65.

De Leo D, Diekstra RFW, Lonnqvist J, et al. (1998) LEIPAD; an internationally applicable instrument to assess quality of life in the elderly. *Behavioral Medicine;* 24: 17-27.

Dexter PR, Stump TE, Tierney WM, Wolinsky FD. (1996) The psychometric properties of the SF-36 health survey among older adults in a clinical setting. *Journal of Clinical Geropsychology;* 2: 225-237.

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.

Doetch TM, Alger BH, Glasser M, Levenstein J. (1994) Detecting depression in elderly outpatients: findings from depression symptom scales and the Dartmouth COOP charts. *Family Medicine;* 26: 519-523.

Doraiswamy PM, Khan ZM, Donahue RM, Richard NE. (2002) The spectrum of quality of life impairments in recurrent geriatric depression. *Journal of Gerontology: Medical Sciences;* Feb; 57(2): M134-7.

EASY-Care Information Sheet (2003) EASY-Care - An international standard for assessing the health and social care needs of older people. English version. January. Centre for Health Ageing, The University of Sheffield.

EASY-Care Training Pack (2003) The Single Assessment Process - information and procedural guidance for trainers and practitioners using EASY-Care as a contact and overview assessment tool. August. Centre for Health Ageing, The University of Sheffield.

Eiser C, Moore R. (2001) Quality of life measures in chronic disease of childhood. *Health Technology Assessment.* 5(4).

Ekman I, Fagerberg B, Lundman B. (2002) Health-related quality of life and sense of coherence among elderly patients with severe chronic heart failure in comparison with healthy controls. *Heart and Lung;* Mar-Apr. 31(2): 94-101.

Ferrans CE, Ferrell BR. (1990) Development of a quality of life index for patients with cancer. *Oncology Nursing Forum;* 17: 15-19(supplement).

Ferrans CE, Powers MJ. (1985) Quality of Life Index: development and psyshometric properties. *Advances in Nursing Science;* 8: 15-24.

Fernandez Buergo AM, Esnaola S, Esquisabel R, Ricarte F, Goicoetxea E, Amaya Diez AM, Alejandre P. (2002) Validity and reliability of a questionnaire for screening older people for comprehensive geriatric assessment. *European Journal of General Practice;* 8(2): 50-5.

Fillenbaum GG. (1978) Validity and reliability of the Multidimensional Functional Assessment Questionnaire. In: Multidimensional Functional Assessment: the OARS Methodology. Durham, NC, Duke University Centre for the Study of Aging and Human Development.

Fillenbaum GG, Smyer MA. (1981) The development, validity and reliability of the OARS Multidimensional Functional Assessment Questionnaire. *Journal of Gerontology;* 36: 428-434.

Fillenbaum GG. (1985) Screening the elderly. A brief instrumental activities of daily living measure. *Journal of the American Geriatrics Society;* 33: 698-706.

Fillenbaum GG. (1988) Multidimensional functional assessment of older adults: the Duke Older Americans Resources and Services procedures. Hillsdale, New Jersey and Hove/London: Lawrence Erlbaum Associates.

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

Fletcher A, Dickinson E, Philp I. (1992) Review - audit measures: quality of life instruments for everyday use with elderly patients. *Age and Ageing;* 21: 142-150.

Fowler RW, Congdon P, Hamilton S. (2000) Assessing health status and outcomes in a geriatric day hospital. *Public Health;* 114: 440-445.

Froom J. (1988) Preface: WONCA committee on international classification. Statement on functional status assessment. Calgery, October. In: Lipkin M (editor).

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. (2002a) Quality of life measurement: bibliographic study of patient assessed health outcome measures. *British Medical Journal;* 324 (7351): 1417-1421.

Garratt AM, Schmidt L, Fitzpatrick R. (2002b) Patient-assessed health outcome measures for diabetes: a structured review. *Diabetic Medicine;* 19(1): 1-11.

Garratt AM; in collaboration with the UK Back Pain Exercise and Manipulation Trial. (2003) Rasch analysis of the Roland Disability Questionnaire. *Spine.* 28(1): 79-84.

George LK, Fillenbaum GG. (1985) OARS methodology: A decade of experience in geriatric assessment. *Journal of the American Geriatrics Society;* 33: 607-615.

Girotto JA, Schreiber J, Nahabedian MY. (2003) Breast reconstruction in the elderly: preserving excellent quality of life. *Annals of Plastic Surgery;* Jun. 50(6): 572-8.

Golden R, Teresi J, Gurland B. (1984) Development of indicator scales for the Comprehensive Assessment and Referral Evaluation (CARE) interview schedule. *Journal of Gerontology;* 39: 138-146.

Goldsmith G, Brodwick M. (1989) Assessing the functional status of older patients with chronic illness. *Family Medicine;* 21: 38-41.

Groessl EJ, Kaplan RM, Cronan TA. (2003) Quality of well-being in older people with osteoarthritis. *Arthritis and Rheumatism (Arthritis Care and Research);* 49(1): 23-28.

Guralnik JM, Simonsick EM, Ferrucci L, et al. (1994) A short physical performance battery assessing lower extremity function: association with self-reported disability and prediction of mortality and nursing home admission. *Journal of Gerontology: Medical Sciences;* 49: M85-M94.

Gurland B, Kuriansky J, Sharpe L, Simon R, Stiller P, Birkett P. (1977) The comprehensive assessment and referral evaluation (CARE) - rationale, development and reliability. *International Journal of Aging and Human Development;* 8: 9-42.

Gurland B, Golden R, Teresi J, Challop J. (1984) The SHORT-CARE: an efficient instrument for the assessment of depression, dementia and disability. *Journal of Gerontology;* 39: 166-169.

Gurland B, Wilder D. (1984) The CARE interview revisited: development of an efficient, systematic clinical assessment. *Journal of Gerontology;* 39: 129-137.

Guyatt GH, Feeny DH, Patrick DL. (1993a) Measuring health-related quality of life. *Annals of Internal Medicine;* 118: 622-629.

Guyatt GH, Eagle DJ, Sackett B, Willan A, Griffith L, McIlroy W, Patterson CJ, Turpie I. (1993b) Measuring quality of life in the frail elderly. *Journal of Clinical Epidemiology;* 46: 1433-1444.

Hage C, Mattsson E, Stahle A. (2003) Long-term effects of exercise training on physical activity level and quality of life in elderly coronary patients - a three to six year follow-up. *Physiotherapy Research International;* 8(1): 13-22.

Hamilton S, Congdon P, Fowler R. (1996) Measuring outcomes of care in a day hospital setting. *Elderly Care;* 8: 14-17.

Harada ND, Chiu V, King AC, Stewart AL. (2001) An evaluation of three self-report physical activity instruments for older adults. *Medicine and Science in Sports and Exercise;* Jun. 33(6): 962-70.

Harel Z, Deimling GT. (1984) Social resources and mental health: an empirical refinement. *Journal of Gerontology;* 39: 747-752.

Hawthorne G, Richardson J, Osborne R, McNeil H. (1997) The Australian Quality of Life (AQoL) Instrument; initial validation. Working Paper 66. Centre for Health Program Evaluation. Monash University and the University of Melborne, Australia.

Hayes V, Morris J, Wolfe C, Morgan M. (1995) The SF-36 health survey questionnaire: is it suitable for use with older adults? *Age and Ageing;* 24: 120-125.

Hayes JA, Black NA, Jenkinson C, Young JD, Rowan KM, Daly K, Ridley S. (2000) Outcome measures for adult critical care: a systematic review. *Health Technology Assessment;* 4(34).

Health Outcome Institute (1996) Twelve-item health status questionnaire (HSQ-12) version 2.0 user guide. Bloomington, MN: Health Outcomes Institute.

Heikkinen E, Waters WE, Brzezinski ZJ (editors) (1983) The elderly in eleven countries: a socio-economic survey. World Health Organisation Regional Office for Europe.

Heslin JM, Soveri PJ, Winoy JB, Lyons RA, Buttanshaw AC, Kovacic L, Daley JA, Gonzalo E. (2001) Health status and service utilisation of older people in different European countries. *Scandinavian Journal of Primary Health Care;* Dec. 19(4): 218-22.

Heyrman J, Van Hoeck K. (1996) Chapter 12: Health outcome measures for older people. In: Hutchinson A, McColl, Christie M, Rccalton C. (Ed) *Health Outcome Measures in Primary and Out-Patient Care.* Harwood Academic Publishers, United Kingdom. pp.167-175.

Higginson IJ, Carr A. (2001) Measuring quality of life: using quality of life measures in the clinical setting. Education and Debate. *British Medical Journal;* 322(26May): 1297-1300.

Hill S, Harries U. (1994) Assessing the outcome of health-care for the older person in community settings: should we use the SF-36? *Outcomes Briefing;* 4: 26-27.

Hill S, Harries U, Popay J. (1996) Is the short form 36 (SF-36) suitable for routine health outcomes assessment in health-care for older people? Evidence from preliminary

work in community-based health services in England. *Journal of Epidemiology and Community Health;* 50: 94-98.

Ho SF, O'Mahony MS, Steward JA, Breay P, Buchalter M, Burr ML. (2001) Dyspnoea and quality of life in older people at home. *Age and Ageing;* Mar. 30(2): 155-9.

Hobson JP, Meara RJ. (1997) Is the SF-36 health survey questionnaire suitable as a self-report measure of the health status of older adults with Parkinson's disease? *Quality of Life Research;* 6: 213-216.

Hofman A, Rocca A, Brayne C. et al. (1991) The prevalence of dementia in Europe: a collaborative study of 1980-1990 findings. *International Journal of Epidemiology;* 20:736-49.

Hunt SM, McKenna SP, McEwen J, Backett EM, Williams J, Papp E. (1980) A quantitative approach to perceived health status: a validation study. *Journal of Epidemiology and Community Health;* 34: 281-286.

Husted JA, Cook RJ, Farewell VT, Gladman DD. (2000) Methods for assessing responsiveness: a critical review and recommendations. *Journal of Clinical Epidemiology;* 53: 459-468.

Hutler Asberg K. (1988) The common language of Katz's Index of ADL in six studies of aged and disabled patients. *Scandinavian Journal of Caring Sciences;* 2:171-178.

Iglesias CP, Birks YF, Torgerson DJ. (2001) Improving the measurement of quality of life in older people: the York SF-12. *Quarterly Journal of Medicine;* 94: 695-698.

Inaba K, Goecke M, Sharkey P, Brenneman F. (2003) Long-term outcomes after injury in the elderly. *Journal of Trauma;* Mar. 54(3): 486-91.

Ingram SS, Seo PH, Martell RE et al. (2002) Comprehensive assessment of the elderly cancer patient: the feasibility of self-report methodology. *Journal of Clinical Oncology;* (Feb 1) 20: 770-775.

Irvine D, O'Brien-Pallas LL, Murray M, Cockerill R, Sidani S, Laurie-Shaw B, Lochhaas-Gerlach J. (2000) The reliability and validity of two health status measures for evaluating outcomes of home-care nursing. *Research in Nursing and Health;* 23: 43-54.

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

Jaglal S, Lakhani Z, Schatzker J. (2000) Reliability, validity and responsiveness of the Lower Extremity Measure for patients with a hip fracture. *The Journal of Bone and Joint Surgery;* 82-A (7: July): 955-962.

Jannink-Nijlant JMM, Diederiks JPM, Brouwers MAH, Metsemakers JFM. (1999) Screening for mobility disorders by the Mobility Control subscale of the short version of the Sickness Impact Profile. *Clinical Rehabilitation;* 13: 492-497.

Jenkinson C, Peto V, Fitzpatrick R, Greenhall R, Hyman N. (1995) Self-reported functioning and well-being in patients with Parkinson's Disease: comparison of the Short-form health survey (SF-36) and the Parkinson's disease questionnaire (PDQ-39). *Age and Ageing;* 24: 505-509.

Jenkinson C, Jenkinson D, Shepperd S, Layte R, Petersen S. (1997) Evaluation of treatment for congestive heart failure in patents aged 60 years and older using generic measures of health status (SF-36 and COOP charts). *Age and Ageing;* 26: 7-13.

Jenkinson C, McGee H. (1998). Health Status Measurement, a Brief but Critical Introduction. Radcliffe Medical Press Ltd, Oxford.

Jette AM, Davies AR, Cleary PD, Calkins DR, Rubenstein LV, Fink A, Kosecoff J, Young RT, Brook RH, Delbanco TL. (1986) The Functional Status Questionnaire: reliability and validity when used in primary care. *Journal of General Internal Medicine;* 1(May/June): 143-149.

Jette AM, Cleary PD. (1987) Functional Disability Assessment. *Physical Therapy;* 67: 1854-59.

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.

Kane RA, Kane RL. (1984) Assessing the elderly; a practical guide to measurement. D.C. Heath and Company, Lexington.

Kaplan RM, Bush JW, Berry CC. (1976) Health status: Types of validity and the index of well-being. *Health Service Research;* 11:478.

Kaplan RM, Atkins CJ, Timms R. (1984) Validity of a Quality of Well-Being Scale as an outcome measure in chronic obstructive pulmonary disease. *Journal of Chronic Disease;* 37: 85-95.

Kaplan RM, Anderson JP. (1988) A General Health Policy Model: update and applications. *Health Service Research;* 23: 203-235.

Kaplan RM, Anderson JP, Ganiats TJ. (1993) The Quality of Well-Being Scale: rationale for a single quality of life index. In: Walker SR, Rosser RM (Eds). Quality of Life Assessment: key issues in the 1990s. Dordrecht, Netherlands: Kluwer Academic Publishers: 65-94.

Katz S, Ford AB, Moskowitz R, et al. (1963) Studies of illness in the aged: The index of ADL – a standardised measure of biological and psychosocial function. *Journal of the American Medical Association;* 185: 914-919.

Katz JN, Larson MG, Phillips CB, Fossel AH, Liang MH. (1992) Comparative measurement sensitivity of short and longer health status instruments. *Medical Care;* October. 30(10): 917-925.

Kempen GI, Van Sonderen E, Sanderman R. (1997) Measuring health status with the Dartmouth COOP charts in low-functioning elderly. Do the illustrations affect the outcomes? *Quality of Life Research;* 6: 323-328.

Kind P, Dolan P, Gudex C, Williams A. (1998) Variations in population health status: results froma United Kingdom national questionnaire survey. *British Medical Journal;* 316: 736-72.

Kirby L, Lehmann P, Majeed A. (1998) Dementia in people aged 65 years and older: a growing problem. *Population Trends;* 92: 23-8.

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

Kleinpell RM, Ferrans CE. (2002) Quality of life of elderly patients after treatment in the ICU. *Research in Nursing and Health;* Jun. 25(3): 212-21.

Kliempt P, Ruta D, McMurdo M. (2000) Measuring the outcomes of care in older people: a non-critical review of patient-based measures. I. General health status and quality of life instruments. *Reviews in Clinical Gerontology;* 10: 33-42.

Kosinski M, Keller SD, Hatoum HT, Kong SX, Ware JE Jr. (1999) The SF-36 Health Survey as a generic outcome measure in clinical trials of patients with osteoarthritis and rheumatoid arthritis: tests of data quality, scaling assumptions and score reliability. *Medical Care;* May. 37(5 Suppl): MS10-22.

Landgraf JM, Nelson EC. (1992) Dartmouth COOP primary care network. Summary of the WONCA/COOP international health assessment field trial. *Australian Family Physician;* 21: 255-269.

Larson JL, Kapella MC, Wirtz S, Covey MK, Berry J. (1998) Reliability and validity of the functional performance inventory in patients with moderate to severe chronic obstructive pulmonary disease. *Journal of Nursing Measurement;* 6(1): 55-73.

Last JM. (1995) Ed. A Dictionary of Epidemiology. Third Edition. Oxford University Press. Oxford.

Lawton MP, Moss M, Fulcomer M, Kleban MH. (1982) A research and service oriented multilevel assessment instrument. *Journal of Gerontology;* 37: 91-99.

Lawton MP. (1991) A multidimensional view of quality of life in frail elders. In: The Concept and Measurement of Quality of Life in the Frail Elderly. Academic Press, London.

Liang J, Levin JS, Krause NM. (1989) Dimensions of the OARS mental health measures. *Journal of Gerontology;* 44: 127-138.

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

Liang MH. (2000) Longitudinal construct validity: establishment of clinical meaning in patient evaluative instruments. *Medical Care;* 38(9 Suppl): II84-90.

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.

Liddle J, March L, Carfrae B, Finnegan T, Druce J, Schwarz J, Brooks P. (1996) Can occupational therapy intervention play a part in maintaining independence and quality of life in older people? *Australian and New Zealand Journal of Public Health;* 20: 574-578.

Lim LL-Y, Fisher JD. (1999) Use of the 12-item Short-Form (SF-12) Health Survey in an Australian heart and stroke population. *Quality of Life Research;* 8: 1-8.

Linn MW, Linn BS. (1984) Self-evaluation of life function (self) scale: a short, comprehensive self-report of health for elderly adults. *Journals of Gerontology;* 39: 603-612.

Lisse J, Espinoza L, Zhao SZ, Dedhiya SD, Osterhaus JT. (2001) Functional status and health-related quality of life of elderly osteoarthritic patients treated with Celecoxib. *Journal of Gerontolgy: Medical Sciences;* Mar.56A(3): M167-75.

Livingston G, Watkin V, Manela M, Rosser R, Katona C. (1998) Quality of life in older people. *Aging and MentalHealth;* 2: 20-23.

Lyons RA, Perry HM, Littlepage BNC. (1994) Evidence for the validity of the Short-Form 36 questionnaire (SF-36) in an elderly population. *Age and Ageing;* 23: 182-184.

Lyons R, Crome P, Monaghan S, Killalea D, Daley J. (1997) Health status and disability among elderly people in three UK districts. *Age and Ageing;* 26: 203-209.

Mahoney FI, Barthel DW. (1965) Functional evaluation: the Barthel Index. *Maryland St. Medical Journal;* 14: 61-5.

Mallinson S. (1998) The Short-Form 36 and older people: some problems encountered when using postal administration. *Journal of Epidemiology and Community Health;* 52: 324-328.

Mangione CM, Marcantonio ER, Goldman L, et al. (1993) Influence of age on measurement of health status in patients undergoing elective surgery. *The Journal of the American Geriatrics Society;* 41: 377-383.

Manz BD, Mosier R, Nusser-Gerlach MA, Bergstrom N, Agrawal S. (2000) Pain assessment in the cognitively impaired and unimpaired elderly. *Pain Management Nursing;* 1: 106-115.

McColl E, Jacoby A, Thomas L. et al. (2001) Design and use of questionnaires: a review of best practice applicable to surveys of health service staff and patients. *Health Technology Assessment;* 5(31).

McCusker J, Bellavance F, Cardin S, Belzile E. (1999) Validity of an activities of daily living questionnaire among older patients in the emergency department. *Journal of Clinical Epidemiology;* 52: 1023-1030.

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, Teno J, Lu JFR, Sherbourne CD, Ware JE. (1990) The use of standardised measures of functional status and well-being among cognitively impaired and intact elders: Results from the Medical Outcomes Study. Paper presented at the 43rd Annual Scientific Meeting of the Gerontological Society of America, Boston, MA.

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, Ware JE, Lu JFR, Sherbourne CD. (1994a) The MOS 36-item Short-Form Health Survey (SF-36): III. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Medical Care;* 32(1): 40-66.

McHorney CA, Kosinski M, Ware JE. (1994b) 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: 551-567.

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.

Mitchell SL, Stott DJ, Martin BJ, Grant SJ. (2001) Randomised controlled trial of quadriceps training after proximal femoral fracture. *Clinical Rehabilitation;* 15: 282-290.

Morgan A, Hickson L, Worrall L. (2002) The impact of hearing impairment on quality of life of older people. *Asia Pacific Journal of Speech Language and Hearing;* 7(1): 39-53.

Morishita L, Boult C, Ebbitt B, Rambel M, Fallstrom K, Gooden T. (1995) Concurrent validity of administering the Geriatric Depression Scale and the physical functioning dimension of the SIP by telephone. *Journal of the American Geriatrics Society;* 43: 680-683.

Morris WW, Buckwalter KC. (1988). Functional assessment of the elderly: the Iowa Self-assessment Inventory. In: Measurement of Nursing Outcomes. Volume One. Measuring Client Outcomes (Waltz CF, Strickland OL, editors), pp. 328-351. New York, Springer.

Morris WW, Buckwalter KC, Cleary A, Gilmer JS, Hatz DL, Studer M. (1989) Issues related to the validation of the Iowa Self-assessment Inventory. *Journal of Educational and Psychological Measurement;* 49: 853-861.

Morris WW, Buckwalter KC, Cleary TA, Gilmer JS, Hatz DL, Studer M. (1990) Refinement of the Iowa Self-Assessment Inventory. *Gerontologist;* 30: 243-247.

Murray M, Lefort S, Ribeiro V. (1998) The SF-36: reliable and valid for the institutionalised elderly? *Aging and Mental Health;* 2: 24-27.

National Health Service (NHS) Research and Development (R&D) Strategic Review (1999). Ageing and age-associated disease and disability. Report of Topic Working Group, June 1999.

National Service Framework for Older People (2001). Modern Standards and Service Models. Department of Health. March 2001.

Nelson EC, Wasson J, Kirk J, et al. (1987) Assessment of function in routine clinical practice: description of the COOP Chart method and preliminary findings. *Journal of Chronic Disease;* 40(suppl 1): 55S-63S.

Nelson E, Landgraf J, Hays R, Wasson J, Kirk J. (1990) The functional status of patients. How can it be measured in physicians' offices? *Medical Care;* 28: 1111-1126.

Neumann PJ, Araki SS, Gutterman EM. (2000) The use of proxy respondents in studies of older adults: lessons, challenges, and opportunities. *Journal of the American Geriatrics Society;* 48: 1646-1654.

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.

Nygren C, Iwarsson S, Isacsson A, Dehlin O. (2001) Quality of care in geriatric rehabilitation: clients' perceptions, ADL dependence and subjective well-being in a one-year perspective. *Scandinavian Journal of Occupational Therapy;* 8(3): 148-56.

O'Mahony PG, Rodgers H, Thomson RG, Dobson R, James OFW. (1998) Is the SF-36 suitable for assessing health status of older stroke patients. *Age and Ageing;* 27: 19-22.

Osborne RH, Hawthorne G, Lew EA, Gray LC. (2003) Quality of life assessment in the community-dwelling elderly: validation of the Assessment of Quality of Life (AQoL) instrument and comparison with the SF-36. *Journal of Clinical Epidemiology;* 138-147.

Overcash J, Extermann M, Parr J, Perry J, Balducci L. (2001) Validity and reliability of the FACT-G Scale for Use in the Older Person with Cancer. *American Journal of Clinical Oncology (CCT);* 24(6): 591-596.

Page SA, Verhoef MJ, Emes CG. (1995) Quality of life, bypass surgery and the elderly. *Canadian Journal of Cardiology;* 11(9): 777-782.

Parker SG, Peet SM, Jagger C, Farhan M, Castleden CM. (1998) Measuring health status in older patients. The SF-36 in practice. *Age and Ageing;* 27: 13-18.

Patrick DL, Erickson PE. (1993) Chapter 2: Assessing health-related quality of life doe clinical decision-making. In: Walker SR, Rosser RM (editors) Quality of Life Assessment - Key issues in the 1990's. Kluwer Academic Publishers, London.

Pettit T, Livingston G, Manela M, Kitchen G, Katona C, Bowling A. (2001) Validation and normative data of health status measures in older people: the Islington study. *International Journal of Geriatric Psychiatry;* 16: 1061-1070.

Pfeiffer E. (1975) (editor) Multidimensional functional assessment: the OARS methodology. A manual. Durham, North Carolina: Duke University Centre for the Study of Aging and Human Development.

Pfeiffer E, Johnson TM, Chiofolo RC. (1981) Functional assessment of elderly subjects in four service settings. *Journal of the American Geriatrics Society;* 29: 433-437.

Pfeiffer BA, McClelland T, Lawson J. (1989) Use of the Functional Assessment Inventory to distinguish among the rural elderly in five service settings. *Journal of American Geriatrics Society;* 37: 243-248.

Philp I. (1997) Can a medical and social assessment be combined? *Journal of the Royal Society of Medicine;* 90 (suppl.32): 11-13.

Philp I. (2000) Measuring quality of care. *Age and Ageing;* Mar.29(2): 95-6.

Philp I, Newton P, McKee KJ, Dixon S, Rowse G, Bath PA. (2001) Geriatric assessment in primary care: formulating best practice. *British Journal of Community Nursing;* 6: 290-295.

Philp I, Lowles R, Armstrong G, Whitehead C. (2002) Repeatability of standardized tests of functional impairment and well-being in older people in a rehabilitation setting. *Disability and Rehabilitation;* 24: 243-249.

Pierre U, Wood Dauphinee SL, Korner Bitensky N, Gayton D, Hanley J. (1998) Proxy use of the Canadian SF-36 in rating health status of the disabled elderly. *Journal of Clinical Epidemiology;* 51: 983-990.

Rai GS, Kelland P, Rai SGS, Wientjes HJFM. (1995) Quality of life cards - a novel way to measure quality of life in the elderly. *Archives of Gerontology and Geriatrics;* 21: 285-289.

Radosevich D, Pruitt M. (1995) Twelve-item health status questionnaire. HSQ-12 Version 2.0. Bloomington, MN: Health Outcomes Institute.

Raphael D, Brown I, Renwick R, Cava M, Weir N, Heathcote K. (1995a). The quality of life of seniors living in the community: a conceptualization with implications for public health practice. *Canadian Journal of Public Health;* 86: 228-233.

Raphael D, Smith TF, Brown I, Renwick R. (1995b) Development and properties of the short and brief versions of the Quality of Life Profile - Seniors Version. *International Journal of Health Sciences;* 6: 161-168.

Raphael D, Brown I, Renwick R, Cava M, Weir N, Heathcote K. (1997) Measuring the quality of life of older persons: a model with implications for community and public health nursing. *International Journal of Nursing Studies;* 34: 231-239.

Rebello P, Ortega F, Baltar JM, Alvarez Ude F, Alvarez Navascues R, Alvarez Grande J. (2001) Is the loss of health-related quality of life during renal replacement therapy lower in elderly patients than in younger patients? *Nephrology Dialysis Transplantation;* Aug.16(8): 1675-80.

Reker GT, Wong PT. (1984) Psychological and physical well-being in the elderly: the Perceived Well-Being Scale (PWB). *Canadian Journal on Aging;* 3: 23-32.

Reker GT. (1995) Reliability and validity of the Perceived Well-being Scale - revised (PWB-R). *Unpublished Manuscript.* Department of Psychology, Trent University, Peterborough. Ontario. Canada.

Repetto L, Ausili Cefaro G, Gallo C, Rossi A, Manzione L. (2001) Quality of life in elderly cancer patients. *Annals of Oncology;* 12: S49-S52.

Resnick B, Nahm ES. (2001) Reliability and validity testing of the revised 12-item Short-Form Health Survey in older adults. *Journal of Nursing Measurement;* Fall. 9(2): 151-61.

Resnick B, Parker R. (2001) Simplified scoring and psychometrics of the revised 12-item Short-Form Health Survey. *Outcomes Management for Nursing Practice;* Oct-Dec. 5(4): 161-6.

Reuben DB, Rubenstein LV, Hirsch SH, Hays RD. (1992) Value of functional status as a predictor of mortality: results of a prospective study. *American Journal of Medicine;* 93: 663-669.

Reuben DB, Valle LA, Hays RD, Siu AL. (1995) Measuring physical function in community-dwelling older persons: a comparison of self-administered, interviewer-administered, and performance-based measures. *Journal of the American Geriatrics Society;* 43: 17-23.

Reza M, Taylor CD, Towse K, Ward JD, Hendra TJ. (2002) Insulin improves well-being for selected elderly type 2 diabetic subjects. *Diabetes Research and Clinical Practice;* Mar.55(3): 201-7.

Richardson J. (2001) The EASY-Care assessment system and its appropriateness for older people. *Nursing Older People;* 13: 17-19.

Robinson BE, Lund CA, Keller D, Cuervo CA. (1986) Validation of the Functional Assessment Inventory against a multidisciplinary home-care team. *Journal of the American Geriatrics Society;* 34: 851-854.

Rockwood K. (1995) Integration of research methods and outcome measures: comprehensive care for the frail elderly. *Canadian Journal on Aging;* 14: 151-164.

Rosenberg M. (1965) Society and the Adolescent Self-Image. Princetown University Press. Princetown.

Rosser RM, Watts VC. (1978) The measurement of illness. *Journal of the Operational Research Society*; 29: 529-40.

Rosser R, Cottee M, Rabin R, Selai C. (1992) Chapter 7: Index of health-related quality of life. In: Hopkins A (Ed) Measures of the quality of life and the uses to which such measures may be put. London: Royal College of Physicians.

Rosser RM, Allison R, Butler C, et al. (1993) Chapter 8: The Index of Health-Related Quality of Life (IHQL): a new tool for audit and cost-per QALY analysis. In: Walker RS, Rosser RM. (Eds) Quality of Life Assessment: key issues in the 1990s. Dordrecht: Kluwer Academic.

Rothman ML, Hedrick S, Inui T. (1989) The Sickness Impact Profile as a measure of the health status of non-cognitively impaired nursing-home residents. *Medical Care;* 27: S157-S167.

Rubenstein LV, Calkins DR, Greenfield S, et al. (1989) Health status assessment for elderly patients. Report of the Society of General Internal Medicine Task Force on Health Assessment. *Journal of the American Geriatrics Society;* 37: 562-569.

Sarvimaki A, Stenbock-Hult HB. (2000) Quality of life in old age described as a sense of well-being, meaning and value. *Journal of Advanced Nursing;* 32: 1025-1033.

Schofield M, Mishra G. (1998) Validity of the SF-12 compared with the SF-36 health survey in pilot studies of the Australian longitudinal study on women's health. *Journal of Health Psychology;* 3: 259-271.

Seki E, Watanabe Y, Sunayama S et al., (2003) Effects of phase III cardiac rehabilitation programs on health-related quality of life in elderly patients with coronary artery disease: Juntendo Cardiac Rehabilitation Program (J-CARP). *Circulation Journal;* Jan. 67(1): 73-7.

Seymour DG, Ball AE, Russell EM, Primrose WR, Garratt AM, Crawford JR. (2001) Problems in using health survey questionnaires in older patients with physical disabilities. The reliability and validity of the SF-36 and the effect of cognitive impairment. *Journal of Evaluation in Clinical Practice;* 7: 411-418.

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.

Sherman S, Reuben D. (1998) Measures of functional status in community-dwelling elders. *Journal of General Internal Medicine;* 13: 817-823.

Simpson P. (2002) Clinical outcomes in transition program for older adults with hip fracture. *Outcomes Management;* 6(2): 86-92.

Single Assessment Process (1999) Assessment Tools and Scales. Department of Health. September 2002.

Siu AL, Reuben DB, Ouslander JG, Osterweil D. (1993a) Using multidimensional health measures in older persons to identify risk of hospitalisation and skilled nursing placement. *Quality of Life Research;* 2: 253-261.

Siu AL, Ouslander JG, Osterweil D, Reuben DB, Hays RD. (1993b) Change in self-reported functioning in older persons entering a residential care facility. *Journal of Clinical Epidemiology;* 46: 1093-1101.

Slivinske LR, Fitch VL, Morawski DP. (1996) The Wellness Index: developing an instrument to assess elders' well-being. *Journal of Gerontological Social Work;* 25: 185-204.

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.

Spilker B (1996) Quality of Life and Pharmacoeconomics in Clinical Trials. Lipincott-Raven: Philadelphia 1996.

Spitzer WO, Dobson AJ, Hall J, et al. (1981) Measuring the quality of life of cancer patients: a consise QL-Index for use by physicians. *Journal of Chronic Disease;* 34: 585-597.

Stadnyk K, Calder J, Rockwood K. (1998) Testing the measurement properties of the Short Form-36 Health Survey in a frail elderly population. *Journal of Clinical Epidemiology;* 51: 827-835.

Stewart AL, Hays RD, Ware JE. (1988) The MOS Short-Form General Health Survey: reliability and validity in a patient population. *Medical Care;* 26: 724-735.

Stolee P, Stadnyk K, Myers AM, Rockwood K. (1996) An individualised approach to outcome measurement in geriatric rehabilitation. *Gerontologist;* 36: 97 (abstract).

Stones MJ, Kozma A. (1989) Multidimensional assessment of the elderly via a microcomputer: the SENOTS program and battery. *Psychology and Aging;* 4: 113-118.

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

Stuck AE, Siu AL, Wieland GD, Adams J, Rubenstein LZ. (1993) Comprehensive geriatric assessment: a meta-analysis of controlled trials. *Lancet;* 342: 1032-1036.

Suzuki M, Ohyama N, Yamada K, Kanamori M. (2002) The relationship between fear of falling, activities of daily living and quality of life among elderly individuals. *Nursing and Health Sciences;* Dec.4(4): 155-61.

Tamim H, McCusker J, Dendukuri N. (2002) Proxy reporting of quality of life using the EQ-5D. *Medical Care;* Dec.40(12): 1186-95.

Tedesco C, Manning S, Lindsay R, Alexander C, Owen R, Smucker ML. (1990) Functional assessment of elderly patients after percutaneous aortic balloon valvuloplasty: New York Heart Association classification versus functional status questionnaire. *Heart and Lung: Journal of Critical Care;* 19: 118-125.

Teresi JA, Golden RR, Gurland BJ. (1984a) Concurrent and predictive validity of indicator scales developed for the Comprehensive Assessment and Referral Evaluation interview schedule. *Journals of Gerontology;* 39: 158-165.

Teresi JA, Golden RR, Gurland BJ, Wilder DE, Bennett RG. (1984b) Construct validity of indicator-scales developed from the comprehensive assessment and referral evaluation interview schedule. *Journals of Gerontology;* 39: 147-157.

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 EuroQol Group. (1990) EuroQol: a new facility for the measurement of health-related quality of life. *Health Policy;* 16:199-208.

The Expert Patient: a new approach to chronic disease management for the 21st Century (2001). Department of Health. September 2001.

The Single Assessment Process and EASY-Care as the Contact and Overview Assessment tool (2003). Training Pack - Appendix 2. August. Centre for Health Ageing, The University of Sheffield.

Theiler R, Bischoff HA, Good M, Uebelhart D. (2002) Rofecoxib improves quality of life in patients with hip or knee osteoarthritis. *Swiss Medicine Weekly;* Nov 2.132(39-40): 566-73.

Thorsen H, McKenna SP, Gottschalck L. (1995) Perceived health in three groups of elderly people. A validity study of the Danish version of the Nottingham Health Profile. *Danish Medical Bulletin;* 42: 105-108.

Tibblin G, Tibblin B, Peciva S, Kullman S, Svardsudd K. (1990) The Goteborg Quality of Life Instrument – an assessment of well-being and symptoms among men born 1913 and 1923. *Scandinavian Journal of Primary Health Care;* 8 (Supplement 1): 33-38.

Tidermark J, Zethraeus N, Svensson O, Tornkvist H, Ponzer S. (2002a) Femoral neck fractures in the eldery: functional outcome and quality of life according to the EuroQol. *Quality of Life Research;* 11: 473-481.

Tidermark J, Zethraeus N, Svensson O, Tornkvist H, Ponzer S. (2002b) Quality of life related to fracture displacement among elderly patients with femoral neck fractures treated with internal fixation. *J Orthop Trauma;* 16(1): 34-8.

Tidermark J, Bergstrom G, Svensson O, Tornkvist H, Ponzer S. (2003a) Responsiveness of the EuroQol (EQ-5D) and SF-36 in elderly patients with displaced femoral neck fractures. *Quality of Life Research;* Dec; 12(8): 1069-79.

Tidermark J, Blomfeldt R, Ponzer S, Soderqvist A, Tornkvist H. (2003b) Primary total hip arthroplasty with a Burch-Schneider antiprotrusion cage and autologous bone grafting for acetabular fractures in elderly patients. *Journal of Orthopaedic Trauma;* Mar. 17(3): 193-7.

Van Balen R, Steyerberg EW, Polder JJ, Ribbers TLM, Habbema JDF, Cools HJM. (2001) Hip fracture in elderly patients: outcomes for function, quality of life, and type of residence. *Clinical Orthopaedics and Related Research;* Sep. 390: 232-43.

Van Balen R, Essink-Bot ML, Steyerberg E, Cools H, Habbema DF. (2003) Quality of life after hip fracture: a comparison of four health status measures in 208 patients. *Disability and Rehabilitation;* May 20. 25(10): 507-19.

Walters SJ, Munro JF, Brazier JE. (2001) Using the SF-36 with older adults: a cross-sectional community-based survey. *Age and Ageing;* 30: 337-343.

Ware JE, Sherbourne CD, Davies AR. (1992) Developing and testing the MOS 20-Item Short-Form Health Survey: A general population application. In: Stewart AL, Ware JE. (editors). Measuring functioning and well-being: the Medical Outcomes Study approach (pp.277-290). Durham, NC: Duke University Press.

Ware JE, Sherbourne CD. (1992) The MOS 36-Item Short-Form Health Survey (SF-36). I. Conceptual framework and item selection. *Medical Care;* 30: 473-483.

Ware J, Kosinski M, Keller SD. (1994) SF-36 Physical and Mental Health Summary Scales: A User's Manual. Boston, MA: The Health Institute, New England Medical Centre.

Ware JE, Kosinski M, Keller SD. (1995) SF-12: How to Score the SF-12 Physical and Mental Health Summary Scales. The Health Institute, New England Medical Center. Boston, MA. Second Edition.

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

Ware JE, Gandek B. (1998) Methods for testing data quality, scaling assumptions and reliability: the IQOLA project approach. *Journal of Clinical Epidemiology;* 51(11): 945-952.

Ware JE, Kosinski M, Dewey JE. (2000) How to Score Version Two of the SF-36 Health Survey. Lincoln, RI. Quality Metric, Inc.

Ware JE, Kosinski M, Turner-Bowker DM, Gandek B. (2002) How to Score Version Two of the SF-36 Health Survey (with a supplement documenting Version 1) Lincoln, RI. Quality Metric, Inc.

Weinberger M, Samsa GP, Hanlon JT, Schmader K, Doyle ME, Cowper PA, Uttech KM, Cohen HJ, Feussner JR. (1991) An evaluation of a brief health status measure in elderly veterans. *Journal of the American Geriatrics Society;* 39: 691-694.

Weinberger M, Nagle B, Hanlon JT, Samsa GP, Schmader K, Landsman PB, Uttech KM, Cowper PA, Cohen HJ, Feussner JR. (1994) Assessing health-related quality of life in elderly outpatients: telephone versus face-to-face administration. *Journal of the American Geriatrics Society;* 42: 1295-1299.

Wells G, Beaton D, Shea B, Boers M, Simon L, Strand V, Brooks P, Tugwell P. (2001) Minimal clinically important differences: review of methods. *The Journal of Rheumatology;* 28(2): 406-12.

Wiebe S, Guyatt G, Weaver B, Matijevic S, Sidwell C. (2003) Comparative responsiveness of generic and specific quality-of-life instruments. *Journal of Clinical Epidemiology;* 56:52-50.

Wildner M, Sangha O, Clark DE, Doring A, Manstetten A. (2002) Independent living after fractures in the Elderly. *Osteoporosis International;* 13:579-585.

Wissing U, Unosson M. (2001) A follow-up study of ulcer healing, nutirtion, and life-situation in elderly patients with leg ulcers. *The Journal of Nutrition, Health and Aging;* 5(10): 37-42.

Wissing U, Unosson M. (2002) Life situation and function in elderly people with and without leg ulcers. *Scandinavian Journal of Caring Sciences;* 16:59-65.

Wolinsky FD, Stump TE. (1996) A measurement model of the Medical Outcomes Study 36-Item Short-Form Health Survey in a clinical sample of disadvantaged, older, black, and white men and women. *Medical Care;* 34: 537-548.

Wolinsky FD, Wan GJ, Tierney WM. (1998) Changes in the SF-36 in 12 months in a clinical sample of disadvantaged older adults. *Medical Care;* 36: 1589-1598.

Wood Dauphinee S, Gauthier L, Gandek B, Magnan L, Pierre U. (1997) Readying a US measure of health status, the SF-36, for use in Canada. *Clinical and Investigative Medicine;* 20: 224-238.

World Health Organisation (1947) Constitution of the World Health Organisation (WHO). WHO Chroncile 1-29. Geneva.

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.

Yarnold PR, Bryant FB, Repasy AB, Martin GJ. (1991) The factor structure and cross-sectional distributional properties of the Beth Israel/UCLA Functional Status Questionnaire. *Journal of Behavioral Medicine;* 14 (Apr): 141-153.

Yarnold PR, Stille FC, Martin GJ. (1995) Cross-sectional psychometric assessment of the functional status questionnaire: use with geriatric versus non-geriatric ambulatory medical patients. *International Journal of Psychiatry in Medicine;* 25: 305-317.

Yip JY, Wilber KH, Myrtle RC, Grazman DN. (2001) Comparison of older adult subject and proxy responses on the SF-36 health-related quality of life instrument. *Aging and Mental Health;* 5: 136-142.